Article

Early prediction of preeclampsia in pregnancy with cell-free RNA

Mira N. Moufarrej[1], Sevahn K. Vorperian[2], Ronald J. Wong[3], Ana A. Campos[3], Cecele C. Quaintance[3], Rene V. Sit[4], Michelle Tan[4], Angela M. Detweiler[4], Honey Mekonen[4], Norma F. Neff[4], Courtney Baruch-Gravett[5], James A. Litch[5], Maurice L. Druzin[6], Virginia D. Winn[6], Gary M. Shaw[3], David K. Stevenson[3] & Stephen R. Quake[1,4,7 ✉]

Liquid biopsies that measure circulating cell-free RNA (cfRNA) offer an opportunity to study the development of pregnancy-related complications in a non-invasive manner and to bridge gaps in clinical care[1–4]. Here we used 404 blood samples from 199 pregnant mothers to identify and validate cfRNA transcriptomic changes that are associated with preeclampsia, a multi-organ syndrome that is the second largest cause of maternal death globally[5]. We find that changes in cfRNA gene expression between normotensive and preeclamptic mothers are marked and stable early in gestation, well before the onset of symptoms. These changes are enriched for genes specific to neuromuscular, endothelial and immune cell types and tissues that reflect key aspects of preeclampsia physiology[6–9], suggest new hypotheses for disease progression and correlate with maternal organ health. This enabled the identification and independent validation of a panel of 18 genes that when measured between 5 and 16 weeks of gestation can form the basis of a liquid biopsy test that would identify mothers at risk of preeclampsia long before clinical symptoms manifest themselves. Tests based on these observations could help predict and manage who is at risk for preeclampsia—an important objective for obstetric care[10,11].

Advances in obstetrics and neonatology have substantially mitigated many of the adverse pregnancy outcomes related to preterm birth and preeclampsia[3]. Nonetheless, the standard of care implemented today focuses on how to treat a mother and child once a complication has been diagnosed, which proves both insufficient and costly[1,2,4,12]: Preeclampsia and related hypertensive disorders cause 14% of maternal deaths each year globally, second only to haemorrhage[5], and cost nearly US$2 billion in care in the first year following delivery[2]. Worse, three out of five maternal deaths in the USA are preventable and often associated with a missed or delayed diagnosis[13]. Such outcomes highlight the need for tools that would aid in identifying which mothers are at risk for preeclampsia before clinical presentation[10,11].

Formally defined as new-onset hypertension with proteinuria or other organ damage (for example, renal, liver or brain) occurring after 20 weeks of gestation[14], preeclampsia can clinically manifest anytime thereafter, including into the post-partum period[15]. So far, no recommended test exists that can predict the future onset of preeclampsia early in pregnancy[10]. Liquid biopsies that measure plasma cfRNA suggest a means to achieve this[16]; there have been promising results both in the confirmation of preeclampsia at clinical diagnosis[17,18] and earlier in pregnancy[19]. The prediction of preeclampsia early in gestation, before symptoms present, could guide the prophylactic use of potential therapeutic agents[10] such as low-dose aspirin[11].

Preeclampsia is specific to humans[6] and a few non-human primates[7], and consequently, elucidating its pathogenesis has proven challenging. Broadly, it is accepted that preeclampsia occurs in two stages—abnormal placentation early in pregnancy followed by systemic endothelial dysfunction[6,8,9]. Preeclampsia can present with a diversity of symptoms and efforts to subclassify the disease on the basis of the timing of onset[20] have had mixed success[6,9,21,22]. Separate efforts have focused on subtyping preeclampsia molecularly using placental gene expression and histology[23,24]. As cfRNA is derived from many tissues in the body[25,26], liquid biopsies present a potential means to indirectly observe pathogenesis in real time and to identify physiological changes associated with preeclampsia for proposed subtypes.

Here we report that cfRNA transcriptomic changes can distinguish between normotensive and preeclampsia pregnancies throughout the course of pregnancy, irrespective of preeclampsia subtype. The majority of these cfRNA changes are most marked early in pregnancy—well before the onset of symptoms. Neuromuscular, endothelial and immune cell types and tissues contribute to these cfRNA changes, consistent with important aspects of the pathogenesis of preeclampsia and also suggesting new approaches to stratify the disease. These observations enabled us to identify and independently validate a panel of 18 genes that when measured between 5 and 16 weeks of gestation form a predictive signature of preeclampsia risk. cfRNA measurements also reflect the multifactorial nature of preeclampsia and provide a means

[1]Department of Bioengineering, Stanford University, Stanford, CA, USA. [2]ChEM-H and Department of Chemical Engineering, Stanford University, Stanford, CA, USA. [3]Department of Pediatrics, Stanford University School of Medicine, Stanford, CA, USA. [4]Chan Zuckerberg Biohub, San Francisco, CA, USA. [5]Global Alliance to Prevent Prematurity and Stillbirth (GAPPS), Lynnwood, WA, USA. [6]Department of Obstetrics and Gynecology, Stanford University School of Medicine, Stanford, CA, USA. [7]Department of Applied Physics, Stanford University, Stanford, CA, USA. ✉e-mail: steve@quake-lab.org
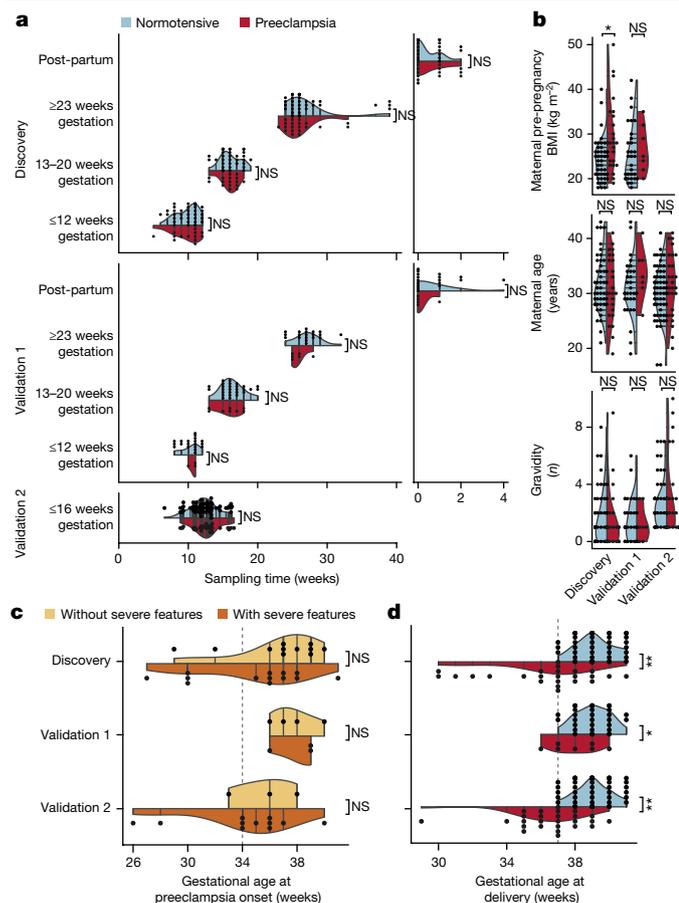
Fig. 1 | Comparing sample, maternal and pregnancy characteristics for normotensive and preeclampsia groups across cohorts. a, Matched sample collection time both across gestation (left) and after delivery (right). **b**, Maternal characteristics ($P = 0.02$ comparing BMI in the discovery cohort). **c**, Matched gestational age at preeclampsia onset regardless of preeclampsia symptom severity. **d**, Gestational age at delivery ($P = 6 \times 10^{-7}$, 0.04, 8×10^{-9}) for the discovery ($n = 49$ normotensive [37, 36, 39, 30]; $n = 24$ with preeclampsia [13, 17, 20, 17]), validation 1 ($n = 32$ normotensive [19, 27, 19, 19]; $n = 7$ with preeclampsia [3, 8, 6, 5]) and validation 2 ($n = 61$ normotensive [61]; 26 with preeclampsia [28]) cohorts. Square brackets indicate the sample number per collection group. For **a**, statistics were calculated by sample group. For **b–d**, statistics were calculated by cohort group (NS = not significant, *$P < 0.05$, **$P \le 10^{-7}$; two-sided (**a–c**) and one-sided (**d**) Mann–Whitney rank test).

to monitor maternal organ health in a non-invasive manner. Together, these results show that cfRNA measurements can form the basis for clinically relevant tests that would predict preeclampsia months before presentation, manage who is at risk for specific organ damage and help to characterize the pathogenesis of preeclampsia in real time.

Clinical study design

To identify changes associated with preeclampsia well before traditional diagnosis, we designed a prospective study and recruited pregnant mothers at their first clinical visit to Stanford's Lucile Packard Children's Hospital. For each participant, we analysed cfRNA for samples collected before or at 12 weeks, between 13 and 20 weeks, at or after 23 weeks of gestation and post-partum. We then split this larger group into discovery ($n = 73$, (49 normotensive, 24 with preeclampsia)) and validation 1 ($n = 39$, (32 normotensive, 7 with preeclampsia)) cohorts. We also obtained samples from an independent cohort of 87 mothers (validation 2); these samples were collected at several separate institutions before 16 weeks of gestation (61 normotensive, 26 with preeclampsia) (Fig. 1a).

All cohorts included individuals of diverse racial and ethnic backgrounds in approximately matched proportions across normotensive and preeclampsia groups (Extended Data Table 1). A pregnancy was considered to be normotensive if it was both uncomplicated and went to full-term (37 or more weeks), or as preeclampsia with or without severe features on the basis of current guidelines (see Methods). For mothers who developed preeclampsia, all antenatal blood samples were collected before diagnosis. Our final analysis included a subset of those samples that passed predefined quality metrics (Extended Data Fig. 1, Supplementary Note 1, Supplementary Table 1, Methods).

Across gestational time points in all cohorts, we found no significant difference in sampling time between preeclampsia and normotensive groups ($P \ge 0.26$, 0.11, 0.46) (values are reported as discovery, validation 1, validation 2; two-sided Mann–Whitney rank test unless otherwise specified). Known risk factors for preeclampsia, such as pre-pregnancy maternal body mass index (BMI), maternal age and gravidity followed expected trends. BMI was significantly different between preeclampsia and normotensive groups in the discovery cohort alone ($P = 0.02$, 0.45, not available), whereas maternal age and gravidity were not ($P \ge 0.29$, 0.16, 0.2) (Fig. 1b, Extended Data Table 1). In validation 2, history of preterm birth and mode of delivery were significantly different between normotensive and preeclampsia groups. Other demographic factors such as race, ethnicity and nulliparity differed across cohorts but not between case groups within each cohort ($P \le 0.05$, two-sided chi-squared test for categorical or ANOVA for continuous variables with Bonferroni correction; Extended Data Table 1, Supplementary Table 2).

In mothers who later developed preeclampsia, we observed no significant difference ($P = 0.14$, 1.0, 0.4) in gestational age at onset between those who did not experience severe symptoms ($n = 11, 4, 3^*$) and those who did experience severe symptoms ($n = 13, 3, 13^*$) (*denotes incomplete data for the specified cohort) (Fig. 1c). Furthermore, 21 mothers who developed preeclampsia also delivered preterm ($n = 9, 1, 11$) as compared with no mothers in the normotensive group; this was reflected by significantly different gestational ages at delivery ($P = 10^{-7}$, 0.04, 10^{-9}; one-sided Mann-Whitney rank test) (Fig. 1d) and lower fetal weight at delivery (Extended Data Table 1), which is consistent with epidemiological evidence that preeclampsia increases the risk of spontaneous or indicated preterm delivery[2,27].

Identifying preeclampsia-related cfRNA changes

A total of 544 differentially expressed genes (DEGs) were altered across gestation and post-partum between mothers who later developed preeclampsia with or without severe features and normotensive mothers who did not experience complications ($P \le 0.05$; see Methods). Most DEGs were annotated as protein-coding and a small fraction (43; 8%) were other types, including 11 mitochondrial transfer RNAs, 6 long non-coding RNAs, 8 pseudogenes and 1 small nucleolar RNA (snoRNA). These changes in gene expression occurred most notably before 20 weeks of gestation, as indicated by a clear bimodal distribution with two peaks centred around a $\log_2$-transformed fold change ($\log_2(FC)$) of $+0.8$ and -0.6 (Fig. 2a). Changes in gene expression were also most stable before 20 weeks of gestation, at which point over 50% of genes had a coefficient of variation (CV) < 1 as compared to 31% of genes at or after 23 weeks of gestation and 36% at post-partum (Extended Data Fig. 2a).

We then asked whether a subset of genes approximately proportional in number to the total sample number ($n = 49, 49, 57, 46$ for 12 weeks or less, 13–20 weeks, 23 weeks or more of gestation, and post-partum, respectively) was sufficient to segregate preeclampsia ($n = 13, 16, 20, 17$) and normotensive ($n = 36, 33, 37, 29$) samples across gestation. We found that 24–32 genes were sufficient to separate preeclampsia and normotensive samples across gestation and at post-partum with good specificity (86% [75–93%], 79% [66–88%], 97% [90–100%] and 90% [78–96%]) and sensitivity (85% [64–95%], 88% [69–96%], 65% [47–80%] and 71% [51–86%]) (values in square brackets are 90% confidence intervals (CI);
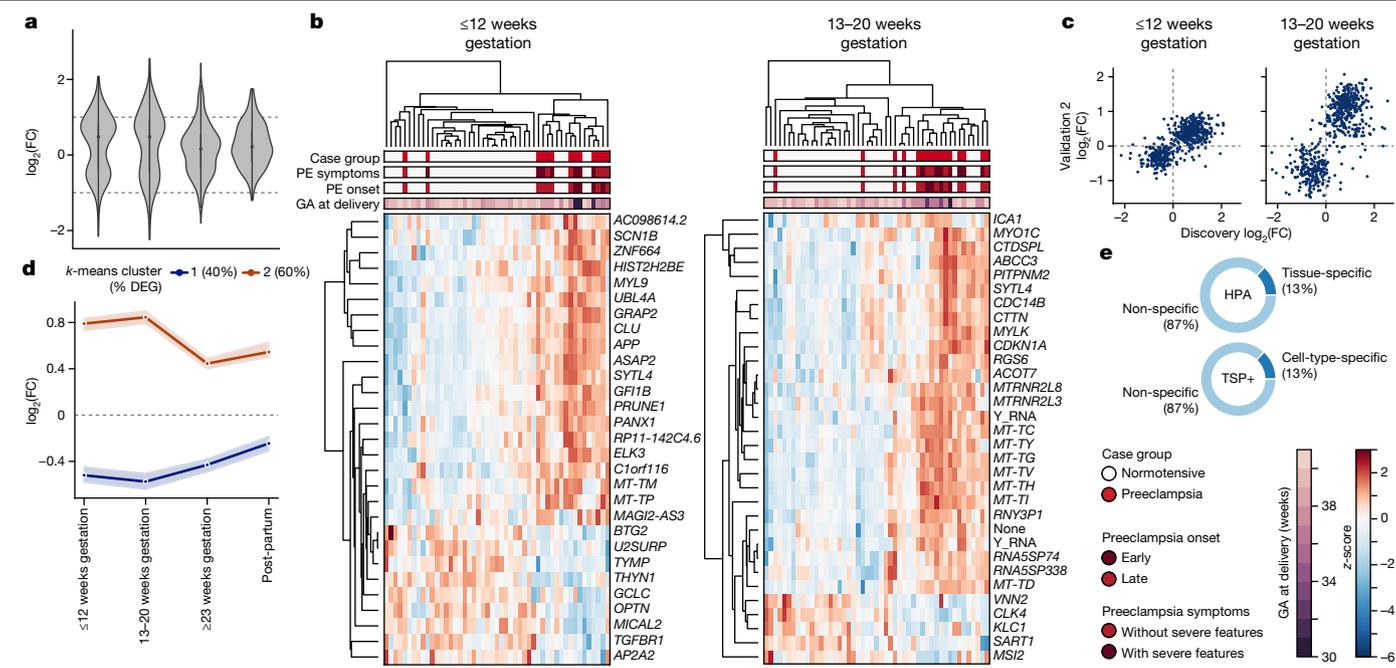
Fig. 2 | Before 20 weeks of gestation, cfRNA measurements segregate preeclampsia and normotensive samples and are enriched for neuromuscular, endothelial and immune cell types and tissues.
a, Distribution of $\log_2$(FC) for DEGs ($n = 544$) with dashed lines at $\log_2$(FC) = ± 1. **b**, Before 20 weeks of gestation, a subset of DEGs can separate preeclampsia (PE) and normotensive samples despite differences in symptom severity, preeclampsia onset subtype and gestational age (GA) at delivery. *HIST2H2BE* is also known as *H2BC21*. See Supplementary Table 3 for more information on

genes included in heatmaps. **c**, Comparison of $\log_2$(FC) for DEGs between the discovery and the validation 2 cohorts reveals strong agreement. **d**, DEGs for preeclampsia as compared to normotensive samples can be described as either increased (orange) or decreased (dark blue) in preeclampsia over gestation. Points indicate median per trend and shaded region indicates 95% CI. **e**, Approximately 13% of DEGs are tissue- or cell-type-specific when compared with the Human Protein Atlas (HPA) and an augmented Tabula Sapiens (TSP+) atlas.

Fig. 2b, Extended Data Fig. 2b). See also Extended Data Fig. 3, Supplementary Table 3.

Nearly all 544 DEG changes showed strong agreement in both validation cohorts as compared to the discovery cohort across gestation but not post-partum. Specifically, more than 82% and 92% of genes across gestation had the same $\log_2$(FC) sign, with a Spearman correlation of at least 0.67 and 0.71 for validation 1 and validation 2 respectively ($P < 10^{-15}$; two-sided *t*-test) as compared with 60% and 0.35 post-partum (Fig. 2c, Extended Data Fig. 2c). Finally, we asked whether symptom severity correlated with $\log_2$(FC) magnitude for these 544 DEGs common to both preeclampsia subtypes. We found that on average, symptom severity did not influence $\log_2$(FC) magnitude as reflected by a slope of nearly one across gestation (Extended Data Fig. 2e).

cfRNA changes reflect preeclampsia pathophysiology
The 544 identified DEGs could be well categorized into two longitudinal trends (Fig. 2d, Extended Data Fig. 4a, c). Resembling a V shape, the first trend (group 1) described the longitudinal behaviour of 216 genes (40%), for which measured levels were reduced in preeclampsia samples ($-1.3\times$ to $-1.5\times$) across gestation with a minimum between 13 and 20 weeks. Peaking in early gestation before 20 weeks ($1.75\times$), the second trend (group 2) described the behaviour of 328 genes (60%) that had significantly increased levels in preeclampsia samples before 20 weeks and to a lesser extent after 23 weeks of gestation ($1.3\times$). For group 1 but not group 2, gene changes were far less evident post-partum and trended towards no difference between preeclampsia and normotensive, which may reflect a placental contribution. DEGs were also enriched for genes previously implicated in preeclampsia[28] broadly (30 gene overlap, $P = 0.006$; one-sided hypergeometric test) and more specifically through placental biopsies[23,24] (two of nine previously identified genes overlap, *PIK3CB*, *TAP1*) (Extended Data Table 2).

Approximately 13% of DEGs were tissue- or cell-type-specific (Fig. 2e). Genes that were decreased in preeclampsia across gestation (group 1) were broadly enriched for the immune system, whereas those genes increased in preeclampsia across gestation (group 2) were enriched for nervous, muscular, endothelial and immune contributions as reflected by cell type and pathway enrichment ($P \leq 0.05$; one-sided hypergeometric test with multiple hypothesis correction, see Methods) (Extended Data Fig. 2e, Extended Data Table 2, Supplementary Table 4). Consistent with the known pathogenesis of preeclampsia, we identified a strong endothelial-linked signal underscored by contributions from capillary aerocytes ($P = 0.03$), an endothelial cell type specific to the lungs[29], platelets ($P = 10^{-33}$) and several platelet-related pathways like platelet degranulation ($P = 10^{-9}$) and platelet activation, signalling and aggregation ($P = 10^{-8}$) among others. We also identified a small, borderline-significant placental contribution ($P = 0.16$) from two genes (*IGF2*, *TGM2*) in group 1 with established roles in trophoblast development[30,31].

We found increased nervous and muscular contributions for preeclampsia as emphasized by contributions from excitatory neurons ($P = 0.02$), oligodendrocytes ($P = 0.005$) and smooth muscle ($P = 0.0003$), and terms like muscle contraction ($P = 0.02$) and dilated cardiomyopathy ($P = 0.01$). The immune system also contributes to both increased (for example, mesenchymal stem cells, total peripheral blood mononuclear cells) and decreased (granulocytes, T cells) changes across gestation. Genes in both groups were enriched for signalling pathways (that is, secretion by cell, integrin-mediated signalling pathway, regulation of IκB kinase, NF-κB signalling). Group 2 was also enriched for cellular compartments such as the cell periphery, cell junctions and extracellular space, consistent with reports that preeclampsia may be associated with signalling from the feto-placental complex[32].
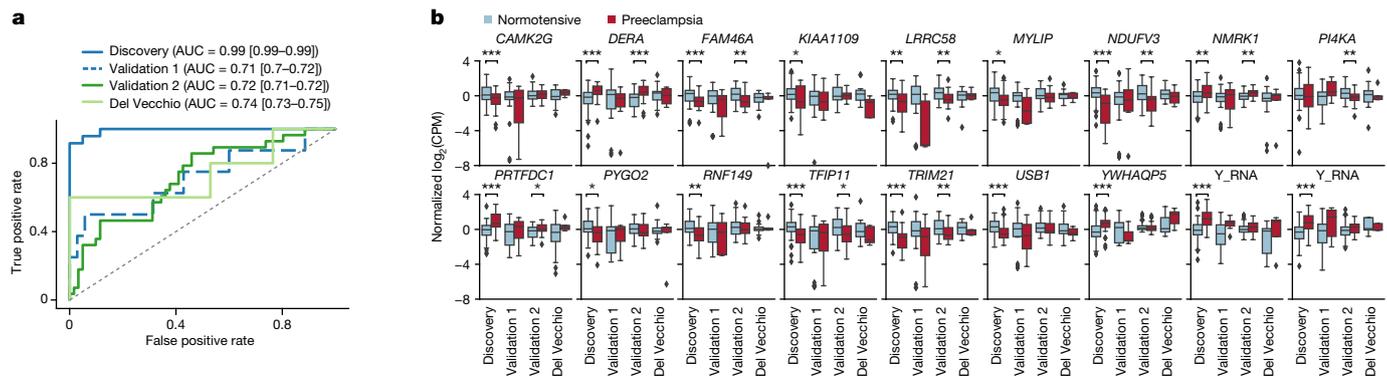
Fig. 3 | A subset of cfRNA changes can predict risk of preeclampsia early in gestation. a, Classifier performance as quantified by receiver operator characteristic curve (ROC) for samples collected in early gestation between 5 and 16 weeks, with AUROC and corresponding 90% CI noted per cohort. **b**, Prediction of preeclampsia incorporates cfRNA levels for 18 genes for which normalized centred $\log_2(FC)$ trends hold across the discovery ($n = 61$ normotensive, 24 preeclampsia), validation 1 ($n = 35$ normotensive, 8 preeclampsia), validation 2 ($n = 61$ normotensive, 28 preeclampsia) and Del

Vecchio ($n = 17$ normotensive or other complication, 5 preeclampsia) cohorts as confirmed using univariate analysis (*$P \le 0.05$, **$P \le 0.01$, ***$P \le 0.005$; one-sided Mann–Whitney rank test with Benjamini–Hochberg correction). See Supplementary Table 5 for exact P values. For box plots, centre line, box limits, whiskers and outliers represent the median, upper and lower quartiles, 1.5× interquartile range and any outliers outside that distribution, respectively. Plot limits are −8 to 4 to better visualize the main distribution. $\log_2(CPM)$, $\log_2$-transformed counts per million reads.

Risk prediction early in gestation

As gene expression changes associated with preeclampsia pathogenesis across gestation were readily detected irrespective of symptom severity, we sought to build a classifier that could identify mothers at risk of preeclampsia at or before 16 weeks of gestation (Extended Data Fig. 2e, Supplementary Note 2). We trained a logistic regression model on the discovery cohort ($n = 61$ normotensive, 24 with preeclampsia). After training, the final model performed well, with a near-perfect area under the receiver operating characteristic curve (AUROC) (0.99 [0.99–0.99]), good specificity (85% [77–91%]) and perfect sensitivity (100% [92–100%]) (Fig. 3a, Extended Data Table 3). We then tested this model on validation 1 ($n = 35$ normotensive, 8 preeclampsia) and two other independent cohorts, which were collected at separate institutions: validation 2 ($n = 61$ normotensive, 28 preeclampsia) and the cohort used by Del Vecchio and colleagues[19] ($n = 8$ normotensive, 5 with preeclampsia, 7 with gestational diabetes, 2 with chronic hypertension). Across these cohorts, the final model once again performed well, with consistent AUROC (0.71 [0.70–0.72], 0.72 [0.71–0.72], 0.74 [0.73–0.75]), sensitivity (75% [46–92%], 56% [42–72%], 60% [26–87%]) and specificity (56% [43–70%], 69% [59–78%], 100% [89–100%]) (all reported as validation 1, validation 2, Del Vecchio, value [90% CI]) (Fig. 3a, Extended Data Table 3).

Misclassified individuals did not predominantly belong to one racial or ethnic group; rather, the fraction of misclassified individuals for each race and ethnicity matched the cohort distribution as a whole. Notably, this holds true even for a group not included in the discovery cohort (American Indian or Alaskan Native) which made up 23% of both case and control for the validation 2 cohort. For false negatives in validation 2 and Del Vecchio, we find a shift to later gestational ages at collection (13.5 ± 2, 12.5 ± 2 weeks) as compared to preeclampsia samples that were correctly classified (12 ± 2, 12 ± 0 weeks; mean ± s.d. for validation 2, Del Vecchio) (Extended Data Fig. 5a). This suggests that in practice, there may an optimal collection window to reduce false negatives. Indeed, if we only consider samples before 14 weeks of gestation, we observe a 9% and a 15% increase in sensitivity, with corresponding AUROC values of 0.73 and 0.90 for validation 2 and Del Vecchio, respectively. There were no false positives from the Del Vecchio cohort, suggesting that the model can distinguish between preeclampsia and other risks like chronic hypertension or gestational diabetes. The model also proved well-calibrated, estimating a slightly increased probability of preeclampsia for gestational diabetes (0.15 ± 0.08) and chronic

hypertension (0.18 ± 0.13)—known preeclampsia risk factors[33,34]—as compared to the estimate for normotensive samples (0.1 ± 0.08) (mean ± s.d., Extended Data Fig. 5b). These increased probabilities for other risk factors affected the AUROC of the test (0.74 [0.73–0.75] as compared to 0.8 [0.79–0.81] for only preeclampsia versus normotensive samples, Extended Data Table 3) (all reported as value [90% CI]).

Finally, we inspected the 18 genes (Fig. 3b, Extended Data Table 4) used by the model to yield probability estimates. Eight genes were annotated in the Human Protein Atlas (HPA, v.19)[35] as enhanced or enriched in the placenta (*TENT5A* (also known as *FAM46A*) and *MYLIP*), neuromuscular (*CAMK2G*, *NDUFV3*, *PI4KA* and *PRTFDC1*) and immune system (*RNF149* and *TRIM21*). Univariate analysis further confirmed that nine of the gene trends (that is, decreased or increased gene levels in preeclampsia) observed in the discovery dataset are upheld in validation 2 ($P \le 0.05$; one-sided Mann–Whitney rank test with Benjamini–Hochberg correction) (Fig. 3b, Supplementary Table 5). We also found that most models trained using a subset of the 18 initial genes can predict future preeclampsia onset with varying performance. Notably, performance improved across all metrics (sensitivity, specificity and AUROC) as we increased the number of genes included for model training (Supplementary Table 6, Extended Data Fig. 5c).

Preeclampsia as a multifactorial disease

By comparing mothers who later developed preeclampsia with or without severe symptoms, we identified 503 DEGs ($P \le 0.05$). As there were no significant differences in symptom severity as related to the timing of preeclampsia onset (Fig. 1c), we believe that our observations contrasting preeclampsia with and without severe symptoms are not obscured by differences in preeclampsia-onset type. DEGs could be well categorized into four longitudinal trends (Fig. 4a, Extended Data Fig. 4b, d). Two groups (groups 1 and 3) described the temporal behaviour of 217 genes (44%), for which measured levels were either consistently increased (group 1) or reduced (group 3) in preeclampsia with as compared to without severe symptoms (±1.8×) across gestation and trended towards no change post-partum. By contrast, groups 2 and 4 (286 genes, 56%) changed signs in mid-gestation, beginning as slightly increased (group 2, 1.2×) or decreased (group 4, −1.2×) in severe preeclampsia and then moving to decreased (group 2, −1.4×) or unchanged (Group 4, 1×) at 23 weeks or more of gestation.

Analysis of the enriched cell types and tissues of origin for each of these groups revealed that increased gene differences in severe preeclampsia were driven by contributions from endothelial cells and
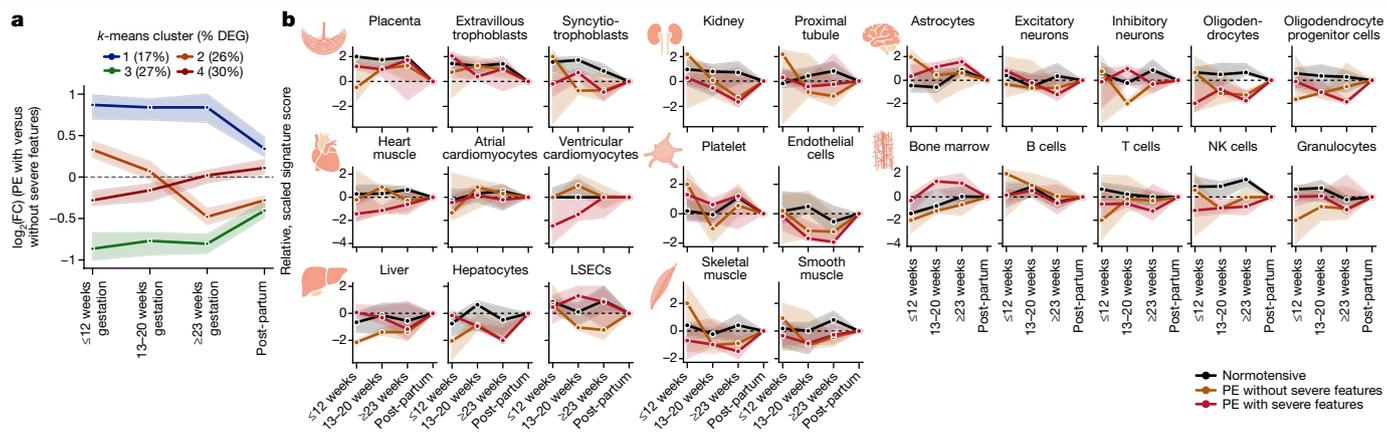
Fig. 4 | cfRNA measurements reflect the multifactorial nature and pathogenesis of preeclampsia over pregnancy before diagnosis. a, DEGs for preeclampsia with as compared to without severe features (*n* = 503) can be described by four longitudinal trends. **b**, Comparison of organ and cell-type changes over gestation for eight organ systems reflects the multifactorial nature of preeclampsia and provides a possible means to comprehensively monitor maternal organ health. Points indicate median per sample group and shaded region indicates 95% CI in **a** and 75% CI in **b**. LSECs, liver sinusoidal endothelial cells; NK cells, natural killer cells.

the adaptive immune system (bone marrow). By contrast, genes that changed signs over gestation were enriched for innate immune cell types (for example, granulocytes and neutrophils for group 2, thymus for group 4) (Extended Data Table 5). Quantifying the total cfRNA signal confirmed an increased bone marrow signal for only severe preeclampsia across gestation and a decreased granulocyte signal for only preeclampsia without severe features at 12 weeks or less of gestation (Fig. 4b). Functional enrichment analysis further revealed pathways specific to genes that were only decreased for severe preeclampsia in early gestation (group 4); for example, axon guidance, nervous development and metabolism of RNA (*P* < 0.05; one-sided hypergeometric test with multiple hypothesis correction; see Methods).

cfRNA reflects maternal organ health

We then investigated the possibility of monitoring organ health in a non-invasive manner. We focused on eight organ systems (Fig. 4b) relevant to preeclampsia presentation with consequences such as proteinuria, impaired liver function, renal insufficiency and epilepsy. We found substantial shifts in total contributions for all systems. We observed an increased astrocyte signal before 20 weeks of gestation and decreased oligodendrocytes and excitatory neurons at 23 or more weeks of gestation for all preeclampsia relative to normotensive (Fig. 4b). Although placental contributions increased over pregnancy with a peak in late gestation as expected, placental tissue and syncytiotrophoblast contributions were reduced for preeclampsia pregnancies before 20 weeks of gestation. Finally, we observed a decreased signal in hepatocyte, kidney, endothelial cell and smooth muscle signatures across gestation and an increased platelet signal before 12 weeks of gestation for preeclampsia. These tissue- and cell-type-specific changes are consistent both with common preeclampsia pathogenesis and with the specific, prominent diagnoses in our cohort (for example, thrombocytopenia, proteinuria, impaired liver function and renal insufficiency).

Discussion

Our findings provide molecular evidence that supports the present physiological understanding of preeclampsia pathogenesis: early abnormal placentation and systemic endothelial dysfunction[6]. Early in gestation, we observe a reduced placental signal for preeclampsia, regardless of onset type or symptom severity. Concurrently, platelets and endothelial cells drive changes in cfRNA in preeclampsia samples regardless of symptom severity as compared to normotensive samples and between individuals with preeclampsia with and without severe symptoms, especially before 20 weeks of gestation. Increases in cell-type-specific cfRNA may occur in part through signalling and secretion by cells, as underscored by functional enrichment analysis. The innate and adaptive immune system also heavily contribute to cfRNA changes in preeclampsia, with marked shifts related to bone marrow, T cells, B cells, granulocytes and neutrophils—consistent with previous studies on the maternal–placental interface and preeclampsia[6,36–38,39].

Given the diversity of clinical presentations in preeclampsia, we propose a non-invasive means of monitoring a mother's risk of specific organ damage. The cfRNA changes that we characterized here reflect dysfunction in at least five organ systems (brain, liver, kidney, muscle and bone marrow), and can in some cases further distinguish between preeclampsia with and without severe symptoms. As a molecular lens into maternal health, liquid biopsies present an opportunity both as a research and clinical tool to learn about the pathogenesis of a human disease in humans and as a predictor of maternal health. Here we have shown proof of principle that cfRNA measurements can form the basis for a robust liquid biopsy test, which predicts preeclampsia very early in gestation. If validated in controlled clinical studies of suitably large, racially and ethnically diverse populations, such tests could help to discover and manage individuals who are at risk for preeclampsia, complementing recent efforts based on clinical and laboratory data[40]. We have also shown here that cfRNA measurements reflect who is at risk for specific organ damage. Together, these results form the basis for a series of clinical tests that can be used to help to characterize and stratify the pathogenesis of preeclampsia in real time, and thereby to meet key objectives for obstetric care.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-022-04410-z.

1. Blencowe, H. et al. Born too soon: the global epidemiology of 15 million preterm births. *Reprod. Health* **10**, S2 (2013).
2. Stevens, W. et al. Short-term costs of preeclampsia to the United States health care system. *Am. J. Obstet. Gynecol.* **217**, 237–248 (2017).
3. Basso, O. et al. Trends in fetal and infant survival following preeclampsia. *JAMA* **296**, 1357–1362 (2006).

Article

4. Beam, A. L. et al. Estimates of healthcare spending for preterm and low-birthweight infants in a commercially insured population: 2008–2016. *J. Perinatol.* **40**, 1091–1099 (2020).

5. Say, L. et al. Global causes of maternal death: a WHO systematic analysis. *Lancet Glob. Health* **2**, e323–e333 (2014).

6. Phipps, E. A., Thadhani, R., Benzing, T. & Karumanchi, S. A. Pre-eclampsia: pathogenesis, novel diagnostics and therapies. *Nat. Rev. Nephrol.* **15**, 275–289 (2019).

7. 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).

8. Steegers, E. A. P., von Dadelszen, P., Duvekot, J. J. & Pijnenborg, R. Pre-eclampsia. *Lancet* **376**, 631–644 (2010).

9. Burton, G. J., Redman, C. W., Roberts, J. M. & Moffett, A. Pre-eclampsia: pathophysiology and clinical implications. *BMJ* **366**, l2381 (2019).

10. McCarthy, F. P., Ryan, R. M. & Chappell, L. C. Prospective biomarkers in preterm preeclampsia: a review. *Pregnancy Hypertens.* **14**, 72–78 (2018).

11. Henderson, J. T., Vesco, K. K., Senger, C. A., Thomas, R. G. & Redmond, N. Aspirin use to prevent preeclampsia and related morbidity and mortality: updated evidence report and systematic review for the US preventive services task force. *JAMA* **326**, 1192–1206 (2021).

12. Behrman, R. E. & Butler, A. S. in *Preterm Birth: Causes, Consequences, and Prevention* (eds Berhman, R. E. & Butler, A. S.) Ch. 12 (National Academic Press, 2007).

13. Petersen, E. E. et al. Vital signs: pregnancy-related deaths, United States, 2011–2015, and strategies for prevention, 13 states, 2013–2017. *MMWR Morb. Mortal. Wkly Rep.* **68**, 423–429 (2019).

14. The American College of Obstetricians and Gynecologists. Hypertension in pregnancy. Report of the American College of Obstetricians and Gynecologists' Task Force on Hypertension in Pregnancy. *Obstet. Gynecol.* **122**, 1122–1131 (2013).

15. Goel, A. et al. Epidemiology and mechanisms of de novo and persistent hypertension in the postpartum period. *Circulation* **132**, 1726–1733 (2015).

16. Whitehead, C. L., Walker, S. P. & Tong, S. Measuring circulating placental RNAs to non-invasively assess the placental transcriptome and to predict pregnancy complications. *Prenat. Diagn.* **36**, 997–1008 (2016).

17. Munchel, S. et al. Circulating transcripts in maternal blood reflect a molecular signature of early-onset preeclampsia. *Sci. Transl. Med.* **12**, eaaz0131 (2020).

18. Tsang, J. C. H. et al. Integrative single-cell and cell-free plasma RNA transcriptomics elucidates placental cellular dynamics. *Proc. Natl Acad. Sci. USA* **114**, E7786–E7795 (2017).

19. Del Vecchio, G. et al. Cell-free DNA methylation and transcriptomic signature prediction of pregnancies with adverse outcomes. *Epigenetics* **16**, 642–661 (2021).

20. von Dadelszen, P., Magee, L. A. & Roberts, J. M. Subclassification of preeclampsia. *Hypertens. Pregnancy* **22**, 143–148 (2003).

21. Raymond, D. & Peterson, E. A critical review of early-onset and late-onset preeclampsia. *Obstet. Gynecol. Surv.* **66**, 497–506 (2011).

22. 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. *J. Perinat. Med.* **41**, 485–504 (2013).

23. Leavey, K. et al. Unsupervised placental gene expression profiling identifies clinically relevant subclasses of human preeclampsia. *Hypertension* **68**, 137–147 (2016).

24. Benton, S. J., Leavey, K., Grynspan, D., Cox, B. J. & Bainbridge, S. A. The clinical heterogeneity of preeclampsia is related to both placental gene expression and placental histopathology. *Am. J. Obstet. Gynecol.* **219**, 604.e1–604.e25 (2018).

25. Koh, W. et al. Noninvasive in vivo monitoring of tissue-specific global gene expression in humans. *Proc. Natl Acad. Sci. USA* **111**, 7361–7366 (2014).

26. Vorperian, S. K., Moufarrej, M. N., Tabula Sapiens Consortium & Quake, S. R. Cell types of origin in the cell free transcriptome in human health and disease. Preprint at *bioRxiv* https://doi.org/10.1101/2021.05.05.441859 (2021).

27. The American College of Obstetricians and Gynecologists. Gestational hypertension and preeclampsia: ACOG practice bulletin, number 222. *Obstet. Gynecol.* **135**, e237–e260 (2020).

28. Uzun, A., Triche, E. W., Schuster, J., Dewan, A. T. & Padbury, J. F. dbPEC: a comprehensive literature-based database for preeclampsia related genes and phenotypes. *Database* **2016**, baw006 (2016).

29. Gillich, A. et al. Capillary cell-type specialization in the alveolus. *Nature* **586**, 785–789 (2020).

30. Harris, L. K., Crocker, I. P., Baker, P. N., Aplin, J. D. & Westwood, M. IGF2 actions on trophoblast in human placenta are regulated by the insulin-like growth factor 2 receptor, which can function as both a signaling and clearance receptor. *Biol. Reprod.* **84**, 440–446 (2011).

31. Robinson, N. J., Baker, P. N., Jones, C. J. P. & Aplin, J. D. A role for tissue transglutaminase in stabilization of membrane-cytoskeletal particles shed from the human placenta. *Biol. Reprod.* **77**, 648–657 (2007).

32. Arck, P. C. & Hecher, K. Fetomaternal immune cross-talk and its consequences for maternal and offspring's health. *Nat. Med.* **19**, 548–556 (2013).

33. Weissgerber, T. L. & Mudd, L. M. Preeclampsia and diabetes. *Curr. Diab. Rep.* **15**, 9 (2015).

34. Sibai, B. M. et al. The impact of prior preeclampsia on the risk of superimposed preeclampsia and other adverse pregnancy outcomes in patients with chronic hypertension. *Am. J. Obstet. Gynecol.* **204**, 345.e1–345.e6 (2011).

35. Uhlen, M. et al. A genome-wide transcriptomic analysis of protein-coding genes in human blood cells. *Science* **366**, eaax9198 (2019).

36. Han, X. et al. Differential dynamics of the maternal immune system in healthy pregnancy and preeclampsia. *Front. Immunol.* **10**, 1305 (2019).

37. Ander, S. E., Diamond, M. S. & Coyne, C. B. Immune responses at the maternal–fetal interface. *Sci. Immunol.* **4**, eaat6114 (2019).

38. Szarka, A., Rigó, J., Lázár, L., Beko, G. & Molvarec, A. Circulating cytokines, chemokines and adhesion molecules in normal pregnancy and preeclampsia determined by multiplex suspension array. *BMC Immunol.* **11**, 59 (2010).

39. Ibarra, A. et al. Non-invasive characterization of human bone marrow stimulation and reconstitution by cell-free messenger RNA sequencing. *Nat. Commun.* **11**, 400 (2020).

40. Marić, I. et al. Early prediction of preeclampsia via machine learning. *Am. J. Obstet. Gynecol MFM* **2**, 100100 (2020).

Methods

Clinical study design

The discovery and validation 1 cohorts were collected as part of a longitudinal, prospective study. We enrolled pregnant mothers (aged 18 years or older) receiving routine antenatal care on or before 12 weeks of gestation at Lucile Packard Children's Hospital at Stanford University, following study review and approval by the Institutional Review Board (IRB) at Stanford University (21956). All individuals signed informed consent before enrolment. Whole-blood samples for plasma isolation were then collected at three distinct time points during their pregnancy course and once (or twice for two individuals) post-partum. To split the larger Stanford cohort into Discovery and Validation 1, we first allocated samples using sequencing batches of which there were three. We allocated the sequencing batch with the most preeclampsia samples to Discovery to ensure sufficient statistical power and the second most preeclampsia samples to Validation 1. Sequencing batches themselves contained randomly allocated samples based on the individual such that all samples from the same individual were in the same batch. For the final sequencing batch, we randomly allocated individuals to either Discovery or Validation 1 such that all samples from 1 individual were part of the same group (either Discovery or Validation 1) and we maintained at least a 1–2 case to control ratio in both groups.

The validation 2 cohort was collected as part of the Global Alliance to Prevent Prematurity and Stillbirth (GAPPS) Pregnancy Biorepository at Yakima Valley Memorial Hospital, Swedish Medical Center and the University of Washington Medical Center under review and approval by Advarra IRB (CR00195799). Samples were processed and sequenced at Stanford under the same IRB as above (21956). All individuals signed informed consent before enrolment. Whole-blood samples for plasma isolation were collected at a single time point (or two time points in the case of two individuals with preeclampsia) before or at 16 weeks of gestation.

For all three cohorts, we chose a case to control ratio of approximately 1–2 to increase statistical power. We also ensured that case and control groups were matched for race and ethnicity, and that all included individuals did not have chronic hypertension or gestational diabetes. No other matching or exclusion criterion were used; we performed no further sample selection prior to sample processing. Mothers were defined as having preeclampsia on the basis of current American College of Obstetrics and Gynecology (ACOG) guidelines (see below). Mothers were defined as controls if they had uncomplicated term pregnancies and either normal spontaneous vaginal or caesarean deliveries. For mothers who developed preeclampsia, all antenatal samples included in this study were collected before clinical diagnosis.

We processed samples from 88 individuals (60 normotensive, 28 with preeclampsia) in the discovery cohort, 43 individuals (34 normotensive, 9 with preeclampsia) in the validation 1 cohort and 87 individuals (61 normotensive, 26 with preeclampsia) in the validation 2 cohort. For some cohorts, only a subset of these individuals was included in the final analysis after filtering samples on the basis of pre-defined quality metrics (see 'Sample quality filtering', Supplementary Note 1, Supplementary Table 1)

We tested for within-cohort (normotensive versus preeclampsia) and across-cohort differences in demographic variables using a two-sided chi-squared test and ANOVA for categorical and continuous variables. respectively. We then applied Bonferroni correction and reported any differences as significant if adjusted $P \leq 0.05$. Investigators were blinded during data collection as pregnancy outcomes were not known yet. Investigators were not blinded during data analysis as analysis methods required knowledge of outcome (that is, supervised learning).

Definition of preeclampsia

Preeclampsia was defined per the ACOG guidelines[27] on the basis of two diagnostic criteria: (1) new-onset hypertension developing on or after 20 weeks of gestation; and (2) new-onset proteinuria or in its absence, thrombocytopenia, impaired liver function, renal insufficiency, pulmonary oedema or cerebral or visual disturbances.

New-onset hypertension was defined when the systolic and/or diastolic blood were at least 140 or 90 mmHg, respectively, on at least two separate occasions between four hours and one week apart. Proteinuria was defined when either 300 mg protein was present within a 24-h urine collection or an individual urine sample contained a protein/creatinine ratio of 0.3 mg dl^{-1}, or if these were not available, a random urine specimen had more than 1 mg protein as measured by dipstick. Thrombocytopenia, impaired liver function and renal insufficiency were defined as a platelet count of less than 100,000 per μl, liver transaminase levels two or more times higher than normal and serum creatinine level of higher than 1.1 mg dl^{-1}, respectively.

Symptoms were defined as severe per the ACOG guidelines. Specifically, preeclampsia was defined as severe if any of the following symptoms were present and diagnosed as described above: new-onset hypertension with systolic and/or diastolic blood pressure of at least 160 and/or 110 mmHg, respectively, thrombocytopenia, impaired liver function, renal insufficiency, pulmonary oedema, new-onset headache unresponsive to medication and unaccounted for otherwise or visual disturbances.

Finally, a pregnant mother was considered to have early-onset preeclampsia if onset occurred before 34 weeks of gestation and late onset thereafter.

Sample preparation

Plasma processing. At Lucile Packard Children's Hospital, blood samples were collected in either EDTA-coated (368661, Becton-Dickinson) or Streck cfRNA BCT (218976, Streck) tubes at ≤12, 13–20 and ≥23 weeks of gestation, and post-partum for each participant. Within 30 min, tubes were then centrifuged at 1,600 g for 30 min at room temperature. Plasma was transferred to 2-ml microfuge tubes and centrifuged at 13,000 g for another 10 min in a microfuge. One-millilitre aliquots were then transferred to 2-ml Sarstedt screw cap microtubes (50809242, Thermo Fisher Scientific) and stored at −80 °C until analysis.

At GAPPS, blood samples were collected in EDTA-coated tubes at ≤16 weeks of gestation from a network of collection sites. Per standard operating procedure, tubes were then centrifuged within 2 h of collection at 2,500 rpm for 10 min at room temperature in a swinging bucket rotor. Plasma was transferred to 2-ml cryovials in at most 1-ml aliquots and stored at −80 °C until analysis. Sample volume was also recorded.

cfRNA isolation. In 96-sample batches, cfRNA from 1-ml plasma samples was extracted in a semi-automated manner using the Opentrons 1.0 system and Norgen Plasma/Serum Circulating and Exosomal RNA Purification 96-Well Kit (Slurry Format) (29500, Norgen). Samples were subsequently treated with Baseline-ZERO DNAse (DB0715K, Lucigen) for 20 min at 37 °C. DNAse-treated cfRNA was then cleaned and concentrated into 12 μl using Zymo RNA Clean and Concentrator-96 kits (R1080).

After cfRNA extraction from plasma samples, isolated RNA concentrations were estimated for a randomly selected 11 samples per batch using the Bioanalyzer RNA 6000 Pico Kit (5067-1513, Agilent) per the manufacturer's instructions.

Sequencing library preparation. cfRNA sequencing libraries were prepared with the SMARTer Stranded Total RNAseq Kit v2 - Pico Input Mammalian Components (634419, Takara) from 4 μl of eluted cfRNA according to the manufacturer's instructions. Samples were barcoded using the SMARTer RNA Unique Dual Index Kit – 96U Set A (634452, Takara), and then pooled in an equimolar manner and sequenced on Illumina's NovaSeq platform (2 × 75 bp) to a mean depth of 54, 33 and 38 million reads per sample for discovery, validation 1, and validation 2 cohorts, respectively. Some samples (12, 61 and 0 for discovery,

validation 1 and validation 2 cohorts) were not sequenced owing to failed library preparation.

Bioinformatic processing

For each sample, raw sequencing reads were trimmed using Trimmomatic (v.0.36) and then mapped to the human reference genome (hg38) with STAR (v.2.7.3a). Duplicate reads were then removed by GATK's (v.4.1.1) MarkDuplicates tool. Finally, mapped reads were sorted and quantified using htseq-count (v.0.11.1) generating a counts table (genes × samples). Read statistics were estimated using FastQC (v.0.11.8).

Across samples, the bioinformatic pipeline was managed using Snakemake (v.5.8.1). Read and tool performance statistics were aggregated across samples and steps using MultiQC (v.1.7). After sample quality and gene filtering, all gene counts were adjusted to $\log_2$-transformed counts per million reads (CPM) with trimmed mean of M values (TMM) normalization[41].

Sample quality filtering

For every sequenced sample, we estimated three quality parameters as previously described[42,43]. To estimate RNA degradation in each sample, we first counted the number of reads per exon and then annotated each exon with its corresponding gene ID and exon number using htseq-count. Using these annotations, we measured the frequency of genes for which all reads mapped exclusively to the 3′-most exon as compared to the total number of genes detected. RNA degradation for a given sample can then be approximated as the fraction of genes in which all reads mapped to the 3′-most exon. To estimate the number of reads that mapped to genes, we summed counts for all genes per sample using the counts table generated from bioinformatic processing above. To estimate DNA contamination, we quantified the ratio of reads that mapped to intronic as compared to exonic regions of the genome.

After measuring these metrics across nearly 700 samples, we empirically estimated RNA degradation and DNA contamination's 95th percentile bound. We considered any given sample an outlier, low-quality sample if its value for at least one of these metrics was greater than or equal to the 95th percentile bound or if no reads were assigned to genes.

Once values for each metric were estimated across the entire dataset, we visualized: (1) whether low-quality samples clustered separately using hierarchical clustering (average linkage, Euclidean distance metric); and (2) whether sample quality drove variance in gene measurements using principal component analysis (PCA). These analyses were performed in Python (v.3.6) using Scikit-learn for PCA (v.0.23.2), Scipy for hierarchical clustering (v.1.5.1) and nheatmap for heat map and clustering visualization (v.0.1.4).

After confirming sample quality, 404 samples from 199 mothers (142 normotensive, 57 with preeclampsia) were included in the final analysis (Supplementary Table 1). Specifically, 209, 106 and 89 samples from 73, 39 and 87 participants (49, 32 and 61 normotensive; 24, 7 and 26 with preeclampsia) were included in discovery, validation 1 and validation 2, respectively.

Gene filtering

We performed filtering to identify well-detected genes across the entire cohort. Specifically, we used a basic cut-off that required a given gene be detected at a level of at least 0.5 CPM in at least 75% of discovery samples after removing outlier samples. Following this step, we retain 7,160 genes for differential expression analysis.

Differential expression analysis

Differential expression analysis was performed in R using Limma (v.3.38.3). To identify gene changes associated with preeclampsia across gestation and post-partum, we used a mixed-effects model. We performed differential expression analysis using two design matrices: (1) examine the interaction between time to preeclampsia onset or delivery for normotensive and preeclampsia symptoms (that is,

preeclampsia with or without severe symptoms); and (2) examine the interaction between time to preeclampsia onset or delivery for normotensive and preeclampsia broadly. In both design matrices, we included time to preeclampsia onset or delivery for normotensive (continuous variable), whether a sample was collected post-partum (binary variable), the interaction between time and preeclampsia symptoms for (1) or preeclampsia for (2), the interaction between whether a sample is post-partum and preeclampsia symptoms for (1) and preeclampsia for (2), and 7–8 confounding factors.

In (1), we defined preeclampsia symptoms categorically using three levels–normotensive, preeclampsia without severe symptoms, preeclampsia with severe symptoms. In (2), we defined whether a sample was preeclampsia using a binary, indicator variable (0 = normotensive, 1 = preeclampsia). The 7–8 confounding variables included were maternal race (categorical variable), maternal ethnicity (binary variable), fetal sex (binary variable), maternal pre-pregnancy BMI group (categorical variable), maternal age (continuous variable, only included in design 1), and sequencing batch (categorical variable). We defined time to preeclampsia onset or delivery as the difference between gestational age at onset or delivery and gestational age at sample collection. We defined BMI group as follows: obese (BMI ≥ 30), overweight (25 ≤ BMI < 30), healthy (18.5 ≤ BMI < 25), underweight (BMI < 18.5). We chose to model time to preeclampsia onset or delivery as a continuous variable, specifically a natural cubic spline with four degrees of freedom to account for the range across which samples were collected (one to three months per collection period). We also blocked for participant identity (categorical variable), modelling it as a random effect to account for auto-correlation between samples from the same person.

Per the Limma-Voom guide, to account for sample auto-correlation over time, we ran the function voomWithQualityWeights twice. We first ran it without any blocking on participant identity, and used this base estimation to approximate sample auto-correlation on the basis of participant identity using the function duplicateCorrelation. After estimating correlation, voomWithQualityWeights was run again, this time blocking for participant identity and including the estimated auto-correlation level. A linear model was then fit for each gene using lmFit and differential expression statistics were approximated using Empirical Bayes (eBayes) methods. For comparing preeclampsia with versus without severe symptoms, we contrasted the relevant coefficients (makeContrasts) and then applied Empirical Bayes as opposed to directly after lmFit.

DEGs were then identified using the relevant design matrix coefficients with Benjamini–Hochberg multiple hypothesis correction at a significance level of 0.05. For design 1, we identified DEGs related to three comparisons: preeclampsia without severe symptoms versus normotensive (1,759 DEGs); severe preeclampsia versus normotensive (1,198 DEGs); and preeclampsia with versus without severe symptoms (503 DEGs). We find 544 genes in common for preeclampsia without and with severe symptoms versus normotensive. These 544 DEGs are examined in Fig. 2 and the related main text. For design 2, we identified DEGs related to preeclampsia versus normotensive alone (330 DEGs), which we used as the initial gene set for building a logistic regression model (see Supplementary Note 2). Finally, we removed the effect of sequencing batch on estimated logCPM values with TMM normalization for the discovery cohort using the limma-voom function, removeBatchEffect.

$\log_2$-transformed fold change and CV estimation

We define $\log_2$-transformed fold-change ($\log_2(FC)$) as the difference between the median gene level (logCPM; see 'Bioinformatic processing') between preeclampsia and normotensive samples for a given sample collection period (that is, ≤12, 13–20 and ≥23 weeks of gestation, or post-partum). In the case in which a given person had multiple samples included into a specific collection period, we only used the values associated with the first collected sample to avoid artificially reducing within-group (preeclampsia or normotensive) variance due to auto-correlation among samples from the same person.

We then quantified the relative dispersion around the estimated $\log_2(FC)$ for each gene using an approximation for CV. Specifically, we consider CV to be the ratio between an error bound, ∂, and the estimated $\log_2(FC)$. We defined the error bound, ∂, as the one-sided error bound associated with the lower (or upper in the case of negative $\log_2(FC)$ values) 95% CI as estimated by bootstrapping. This definition of ∂ as a one-sided bound that approaches 0 (equivalent to no FC) allowed us to explore how confident we could be in an estimated $\log_2(FC)$. For instance, a CV = 1 would indicate that at the boundary of proposed values, the $\log_2(FC)$ for a given gene becomes effectively 0. Similarly, a CV > 1 would indicate even less confidence in a proposed average $\log_2(FC)$ and indicate that at the boundary, the estimated $\log_2(FC)$ changes signs (that is, a negative $\log_2(FC)$ becomes a positive one or vice versa).

Hierarchical clustering analysis

For each sample collection period, hierarchical clustering was performed using DEGs with an $|\log_2(FC)| \geq 1$ and CV < 0.5 or 0.4 in the case of the 13–20 weeks of gestation time point so that the number of genes used did not exceed the number of samples. For each gene that passed these thresholds, we calculated a z-score across all samples (at most 1 per individual, the earliest collected sample in a given group) in each sample collection period using the function StandardScaler in Scikit-learn (v.0.23.2), Average linkage hierarchical clustering with a Euclidean distance metric was then performed for both rows (gene z-scores) and columns (samples in same collection group) for a given matrix in Python using Scipy (v.1.5.1). Clustering and corresponding heat maps were visualized using nheatmap (v.0.1.4).

Longitudinal dynamics analysis

To group gene changes by longitudinal behaviour, we performed k-means clustering on a matrix in which each row was a DEG and each column was the estimated $\log_2(FC)$ for a given sample collection period (N genes × 4 time points). We measured the sum of squared distances for every 'k' between 1 and 16 (4^2), in which 16 represents the maximum possible k (4 time points with 2 possibilities each, $\log_2(FC) > 0$ or $\log_2(FC) < 0$). We then identified the optimal k clusters by using the well-established elbow method to identify the smallest 'k' that best explained the data, visually described as the elbow (or knee) of a convex plot like that for the sum of squared distances versus k (Extended Data Fig. 4a, c). To do so, we either visually inspected and identified the elbow (Fig. 2d) or used the function KneeLocator as implemented in the package Kneed (v.0.7.0) (Fig. 4a). We used visual inspection for Fig. 2d as we observed that given two k-values (for example, k = 2, 3) with a similar sum of squared distances, KneeLocator would choose the larger value, whereas we preferred a simpler model. Having identified the optimal number of clusters, k, we labelled every DEG with its assigned cluster and visualized average behaviour (median) and the 95% CI (bootstrapped using 1,000 iterations) per cluster using Seaborn line plot (v.0.10.0).

To confirm that the identified patterns were not spurious (that is, an artifact of the k-means clustering algorithm), we permuted the data columns ($\log_2(FC)$ per time point) for each gene thereby scrambling any time-related structure while preserving its overall distribution. We then visualized the result using Seaborn line plot as described above. After permutation, we observed no longitudinal patterns, which were instead replaced by nearly flat, uninformative trends (Extended Data Fig. 4b, d).

Correlation analysis

To verify DEGs identified in the discovery cohort, we compared $\log_2(FC)$ values for the discovery cohort as compared to both validation 1 and validation 2 cohorts. We calculated the percentage of genes for which the $\log_2(FC)$ had the same sign across cohorts (that is, both positive or both negative) and the Spearman correlation using the function scipy.

stats.spearmanr. We did not calculate $\log_2(FC)$ values for DEGs at ≤12 weeks of gestation in validation 1 because of small sample numbers (3 preeclampsia samples before 12 weeks).

We also sought to compare whether symptom severity (without or with severe preeclampsia) correlated with $\log_2(FC)$ magnitude for 544 DEGs identified as common to all preeclampsia in design 1. To do so, we calculated the slope of a best-fit line in which x and y were defined as $\log_2(FC)$ values for preeclampsia without (x) and with (y) severe features versus normotensive. We would expect a slope > 1 and < 1 if $\log_2(FC)$ magnitudes for preeclampsia with as compared to without severe symptoms were larger or smaller on average, respectively. Similarly, a slope = 1 would reflect that symptom severity did not correlate with $\log_2(FC)$ magnitude for the 544 DEGs tested.

Finally, to confirm that the identified correlations were significant, we permuted the data columns ($\log_2(FC)$ per cohort) for each gene, thereby scrambling any structure while preserving its overall distribution. We then calculated the same statistics. After permutation, we observe about 50–55% $\log_2(FC)$ agreement, as expected at random, a Spearman correlation of 0 and a slope of 0.

Defining cell-type- and tissue-specific gene profiles

Cell-type gene profiles were identified as previously described[26] and briefly summarized below. We also briefly describe an adapted, similar method to derive tissue gene profiles.

On the tissue level, for genes and tissues (and some blood and immune cell types) measured in the HPA (v.19)[35], we calculated a Gini coefficient per gene as a measure of tissue specificity. As a measure of inequality, Gini coefficient values closer to 1 represent genes that are tissue-specific. We defined a given gene Y as specific to tissue X included in the HPA if Gini(Y) ≥ 0.6 and max expression for Y is in tissue X. The aforementioned method can identify genes that are expressed in several tissues (for example, group-enriched) as opposed to only one. Specifically, it is possible to have a gene Y where Gini(Y) ≥ 0.6 and the gene is expressed in more than 1 tissue (for example, enrichment in placenta and muscle). To this end, when tracking a single cell type or tissue's trajectory over gestation (for example, Fig. 4b), where the specificity of a given gene profile is especially important, we imposed a further constraint to ensure that any gene signal only reflects the named tissue (for example, any gene named placenta-specific is specific to the placenta alone). Specifically, we required that genes be annotated by HPA as 'Tissue-enriched' or 'Tissue-enhanced' and term this reference 'HPA strict'.

On the cell-type level, we identified cell-type-specific gene profiles using both Tabula Sapiens v.1.0 (TSP) and individual cell atlases. We used individual cell atlases to identify gene profiles for cell types from missing tissues in TSP (for example, placenta, brain) or tissues that are known to be important in preeclampsia with additional annotations in individual single-cell atlases (for example, liver, kidney). First, for genes and cell types measured in TSP, we defined a given gene Y as specific to cell type X included in TSP if Gini(Y) ≥ 0.8 and max mean expression for Y is in cell type X. We combined subtype annotations for neutrophil and endothelial cells into single parent categories called 'neutrophil' and 'endothelial cell' respectively; as subtype annotations were based on manifold clustering, it was unclear whether they were truly distinct enough to be distinguished at a whole-body level for our purposes. Next, for genes and cell types described in individual tissue single-cell atlases, we required that a gene be differentially expressed in the specific single-cell atlases and tissue-specific per the HPA (Gini ≥ 0.6). The following single-cell atlases were used for each organ: (1) placenta[44,45]; (2) liver[46]; (3) kidney[47]; (4) heart[48]; and (5) brain[49].

We then created an augmented reference, which we term TSP+. For TSP+, we took the union of TSP and individual atlas gene annotations. A small number of genes had conflicting double annotations in TSP as compared to at most one individual tissue single-cell atlas. In these rare instances, which most often occurred for genes related

Article

to cell types missing in TSP (for example, placental or brain cell types), we used the individual single-cell atlas label.

Determining cell type and tissue of origin

We determined whether a given cell type or tissue was enriched in preeclampsia by comparing preeclampsia DEGs with cell-type and tissue gene profiles using a one-sided hypergeometric test. For every tissue (HPA) or cell type (TSP+) with at least two DEGs specific to it, we performed the following. First, we defined a hypergeometric distribution (scipy.stats.hypergeom, (v.1.5.1)) with the following parameters, where category refers to tissue when using HPA and cell type when using TSP+: M = number of genes specific to any category; n = number of genes specific to this category; N = number of DEGs in this k-means longitudinal cluster specific to this category. Next, we estimated a P value using the survival function (1 − cumulative distribution function (CDF)) for the specified distribution. Specifically, a P value is defined as the cumulative probability, $P(X > (n\_DEGs\_specific\_to\_this\_category − 1))$, that the distribution takes a value greater than the number of DEGs specific to this category − 1. Finally, once we estimated a P value for every cell type (TSP+) or tissue (HPA) identified in each DEG k-means longitudinal cluster, we adjusted for multiple hypotheses using Benjamini−Hochberg with a significance threshold of 0.05.

Defining relative signature score per cell type or tissue

We define a signature score as the sum of logCPM values over all genes in a given tissue or cell-type gene profile[26]. We required that a cell-type or tissue gene profile have at least five specified genes to be considered for signature scoring in cfRNA. Genes were defined as specific to a given tissue on the basis of the reference HPA strict, and to a given cell type on the basis of the reference TSP+ (see 'Defining cell-type- and tissue-specific gene profiles' for details).

To account for our previous observation that baseline cfRNA levels vary between individuals—the consequence of biological and technical (for example, sample processing) factors, we chose to calculate relative as opposed to absolute signature scores. For each individual whose post-partum sample passed sample quality control (QC) (see 'Sample quality filtering' for details), we estimated a relative signature score, which was defined as the difference between the signature score at a given gestational time point and the post-partum sample. For both the discovery cohort and the validation 1 cohort, 49 normotensive individuals and 24 individuals with preeclampsia had a post-partum sample that survived sample QC. After normalization, all samples at post-partum had a similar baseline (0). We note that one can define a relative signature score based on any sampled time point for a given person. We chose the post-partum sample because we were interested in tracking maternal organ health over gestation.

Finally, we scaled (that is, z-score) the relative signature scores for a given cell type or tissue by dividing by the interquartile range, a robust alternative to standard deviation, using the sklearn.preprocessing class, RobustScaler. This accounted for differing gene profile lengths and gene expression levels, and allowed us to compare both different cell-type and tissue contributions and case groups per cell type or tissue.

Having defined a relative signature score per cell type and tissue, we visualized average behaviour (median) and the 75% CI, a non-parametric estimation of standard deviation, (bootstrapped relative signature score per case group and time point using 1,000 iterations) using Seaborn line plot (v.0.10.0).

Functional enrichment analysis

Functional enrichment analysis was performed using the tool GProfiler (v.1.0.0) for the following data sources: Gene Ontology: biological processes and cellular compartments (GO:BP, GO:CC, released 1 May 2021, 2021-05-01), Reactome (REAC, released 7 May 2021, 2021-05-07) and Kyoto Encyclopedia of Genes and Genomes (KEGG FTP, released 3 May 2021, 2021-05-03). To identify GO terms, we excluded electronic GO annotations (IEA) and used a custom background of only the 7,160 genes that were included in the differential expression analysis. We then performed the recommended multiple hypothesis correction (g:SCS) with an experiment-wide significance threshold of a = 0.05 (ref. [50]).

Logistic regression feature selection and training

To build a robust classifier that can identify mothers at risk of preeclampsia at or before 16 weeks of gestation, we first pre-selected features using the set of 330 DEGs when contrasting preeclampsia versus normotensive (see design 2 in 'Differential expression analysis' and Supplementary Note 2) as a starting point.

We normalized gene measurements using a series of steps. First, to correct for batch effect, in which we define batch as a set of samples processed at the same time by a distinct group (for example, discovery cohort = batch, Del Vecchio cohort = batch), we centred the data by subtracting the median logCPM per gene for a given cohort. Next, we scaled gene values for each cohort using its corresponding interquartile range in the discovery cohort. Finally, to account for sampling differences across samples, we used an approach similar to when analysing quantitative PCR with reverse transcription (RT–qPCR) data, and normalized data using multiple internal control (that is, housekeeping) genes. On a per-sample basis, we subtracted the median, normalized logCPM value (centred and scaled) for all internal control genes, which we define as 66 genes for which the measured value did not change across preeclampsia versus normotensive comparisons (all genes with adjusted $P > 0.99$ for preeclampsia versus normotensive; design 2). When calculating the median value for all internal control genes, we excluded any 0 logCPM values as these are likely to have been the consequence of technical dropout.

Model training then used the discovery cohort alone split into 80% for hyperparameter tuning and 20% for model selection and consisted of two stages—further feature pre-selection based on two metrics followed by the construction of a logistic regression model with an elastic net penalty. Using a split discovery cohort for training mitigated overfitting even though all discovery samples were used for differential expression, which defined the initial feature set.

For feature pre-selection, we calculated logFC values using the 80% discovery split for all 330 genes for preeclampsia versus normotensive. We focused on 2 practical metrics measured across the 80% split of discovery samples collected on or before 16 weeks of gestation: gene change size ($|\log_2(FC)|$) and gene change stability (CV). All model hyperparameters were then tuned using AUROC as the outcome metric and fivefold cross validation. Next, we selected the best model including tuned feature pre-selection cut-offs again using AUROC. Specifically, we calculated an AUROC score for both the 80% and the 20% discovery splits separately, and the selected model achieved the best score on both splits.

Finally, we tuned the probability threshold, P, at which a sample is labelled as at risk of preeclampsia if $P(PE) \geq P$ using the entire discovery cohort. To do so, we constructed a receiver operator characteristic curve (ROC) and calculated the false positive rate (FPR) and true positive rate (TPR) at different thresholds, P_i. We identified the threshold, P_i, at which FPR = 10%, and round to the nearest 5 (for example, 0.37 would become 0.35). This yielded a tuned threshold of P = 0.35. All classifications as negative or positive were then made based on this threshold.

To understand the importance of each gene feature, we trained a separate logistic regression model for a subset of all possible feature subsets (307 combinations out of a total of 262,143 for 1–17 genes). No feature pre-selection was performed for this sub-analysis. All model hyperparameters were tuned as previously described. We defined a gene subset as weakly predictive if the model yielded an AUROC > 0.5 on the test set (validation 2).

In all cases, performance metrics were assessed as described below (see next section) and used Scikit-learn (v.0.23.2).

Performance metric analysis

Model performance was assessed using several statistics including sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) and AUROC. Given a 2 × 2 confusion matrix in which rows 1 and 2 represent true negatives and positives and columns 1 and 2 represent negative and positive predictions, respectively, we can define the values in row 1, column 1 as true negatives (TN), row 1, column 2 as false positives (FP), row 2, column 1 as false negatives (FN), and row 2, column 2 as true positives (TP). We can then define the following proportions: (1) sensitivity = TP/(TP + FN); (2) specificity = TN/(TN + FP); (3) PPV = TP/(TP + FP); (4) NPV = TN/(TN + FN). For each proportion, we calculated 90% CIs using Jeffrey's interval[51] and the function, proportion_confint, from statsmodels.stats.proportion. We also approximated AUC and its corresponding 90% CI using the Scikit-learn function, roc_auc_score, and the binormal approximation, respectively.

Statistical analyses

All *P* values reported herein were calculated using the non-parametric Mann–Whitney rank test unless otherwise stated. One-sided tests were performed where required based on the hypothesis tested.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Raw and processed sequencing data (BioProject PRJNA792450) are available through the Sequence Read Archive (SRA, SRP352519) and the Gene Expression Omnibus (GEO, GSE192902), respectively. Data were mapped using the human reference genome (hg38) and annotated using ENSEMBL v.82. We used publicly available data from the HPA (v.19; https://v19.proteinatlas.org/); TSP (v.1.0; https://tabula-sapiens-portal.ds.czbiohub.org/); Gene Ontology: biological processes and cellular compartments (GO:BP, GO:CC, released 1 May 2021, 2021-05-01); Reactome (REAC, released 7 May 2021, 2021-05-07); Kyoto Encyclopedia of Genes and Genomes (KEGG, released 3 May 2021, 2021-05-03); and several previous publications[44–49].

Code availability

All computational analyses were performed using Python 3.6 or R 3.5, and are available on Github at https://github.com/miramou/pe_cfrna.

41. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, R25 (2010).

42. Pan, W. *Development of Diagnostic Methods Using Cell-free Nucleic Acids* (Stanford University, 2016).

43. Moufarrej, M. N., Wong, R. J., Shaw, G. M., Stevenson, D. K. & Quake, S. R. Investigating pregnancy and its complications using circulating cell-free RNA in women's blood during gestation. *Front. Pediatr.* **8**, 605219 (2020).

44. Vento-Tormo, R. et al. Single-cell reconstruction of the early maternal–fetal interface in humans. *Nature* **563**, 347–353 (2018).

45. Suryawanshi, H. et al. A single-cell survey of the human first-trimester placenta and decidua. *Sci. Adv.* **4**, eaau4788 (2018).

46. Aizarani, N. et al. A human liver cell atlas reveals heterogeneity and epithelial progenitors. *Nature* **572**, 199–204 (2019).

47. Stewart, B. J. et al. Spatiotemporal immune zonation of the human kidney. *Science* **365**, 1461–1466 (2019).

48. Litviňuková, M. et al. Cells of the adult human heart. *Nature* **588**, 466–472 (2020).

49. Mathys, H. et al. Single-cell transcriptomic analysis of Alzheimer's disease. *Nature* **570**, 332–337 (2019).

50. Raudvere, U. et al. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.* **47**, W191–W198 (2019).

51. Brown, L. D., Cai, T. T. & DasGupta, A. Interval estimation for a binomial proportion. *Stat. Sci.* **16**, 101–133 (2001).

Author contributions M.N.M., G.M.S., D.K.S. and S.R.Q. conceptualized and designed the study. M.L.D., V.D.W., G.M.S. and D.K.S. designed the cohort study at Lucile Packard Children's Hospital. R.J.W., A.A.C. and C.C.Q. collected and processed whole-blood samples and corresponding sample and participant metadata. C.B.-G. and J.A.L. oversaw whole-blood sample and corresponding sample and participant metadata collection at GAPPS. M.N.M. developed experimental protocols for cfRNA extraction and data generation and processed all samples from cfRNA extraction up to and including library generation for RNA sequencing. R.V.S., M.T., A.M.D., H.M. and N.F.N. developed and executed experimental protocols for RNA sequencing and computational protocols for raw sequencing data transfer. S.K.V. developed computational methods to define genes as cell-type- and tissue-specific in the context of the whole body. M.N.M. and S.K.V. designed the analyses to characterize tissue and cell-type contributions. M.N.M. conceptualized and developed computational analyses, analysed the data and wrote the initial manuscript draft in collaboration with S.R.Q. All authors contributed to writing and editing the manuscript.

Extended Data Fig. 1 | Samples with outlier values for at least one QC metric cluster separately from most non-outlier samples. a–c, For discovery (**a**), validation 1 (**b**), and validation 2 (**c**), hierarchical clustering (left) and PCA (right) reveals that most outlier samples cluster with negative control (NC) samples and separately from non-outlier samples. **d**, **e**, Visualization of other QC metrics like the amount of cfRNA extracted (**d**) and the percent of reads that align uniquely to the human genome (**e**) (n = 209, 106, 89 samples for discovery, validation 1, and validation 2, respectively). For PCA in **a–c**, sample outliers and poorly detected genes drive PCA and serve as leverage points. The top two principal components are visualized when performed using all samples and all genes (leftmost PCA) or only samples that pass QC metrics (middle PCA) reveals that certain samples can act as leverage points. Once sample outliers and lowly detected genes are removed from the cfRNA gene matrix (rightmost PCA), the top two principal components reflect natural variance in the data and are no longer driven by a few leverage points. For box plots, centre line, box limits, whiskers and outliers represent the median, upper and lower quartiles, 1.5× interquartile range and any outliers outside that distribution, respectively.
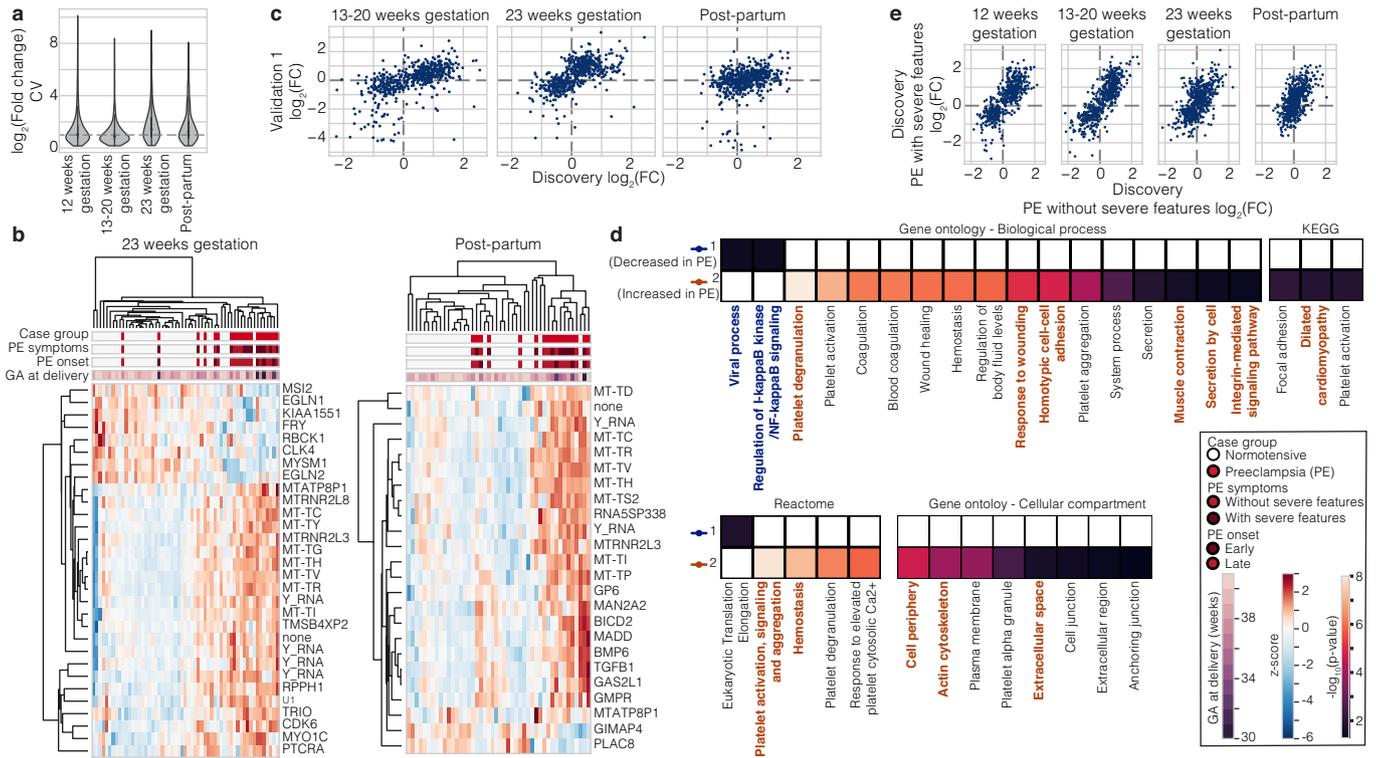
Extended Data Fig. 2 | Across gestation before diagnosis, changes in the cfRNA transcriptome segregate preeclampsia and normotensive samples and reflect known preeclampsia biology. a, Distribution of CVs with dashed line at CV = 1 for all DEGs (n = 544) between preeclampsia as compared to normotensive samples across gestation. **b**, At ≥23 weeks of gestation and post-partum, a subset of DEGs can separate preeclampsia (n = 20, 17) and normotensive (n = 37, 29) samples despite differences in symptom severity, preeclampsia onset subtype, and GA at delivery. **c**, Comparison of log$_2$(FC) for DEGs for preeclampsia as compared to normotensive between discovery and validation 1 reveals good agreement across gestation but not post-partum. **d**, The genes in each longitudinal trend group reflect known preeclampsia

aetiology as highlighted across four databases (GO biological processes, KEGG, the reactome, and GO cellular compartment). Some preeclampsia associated terms are emphasized in bold, coloured text that corresponds to group colour from Fig. 2d (Dark blue and orange indicate decreased and increased in preeclampsia versus normotensive, respectively) (p ≤ 0.05; one-sided hypergeometric test with multiple hypothesis correction, see Methods). **e**, Comparison of log$_2$(FC) for DEGs for preeclampsia without severe features versus normotensive and preeclampsia with severe features versus normotensive in the discovery cohort reveals good agreement along the y=x axis with a slope of 0.93, 1.03, 0.77, and 0.86 at ≤12 weeks, 13–20, ≥23 weeks, and post-partum, respectively.
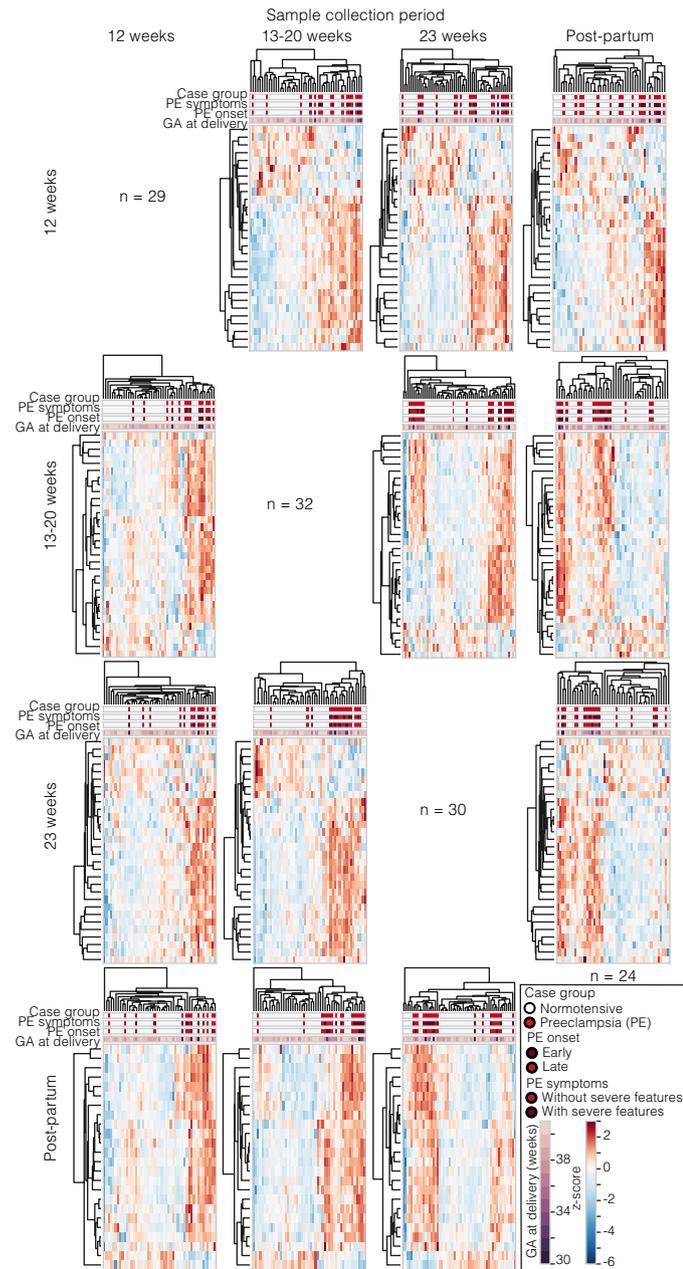
Extended Data Fig. 3 | Across gestation and before diagnosis, changes in the cfRNA transcriptome identified at one time point can moderately segregate preeclampsia and normotensive samples at other time points. DEGs with $|\log_2(FC)| \geq 1$ and CV < 0.5 or 0.4 for the 13–20 week time point were identified at each time point across gestation. Each row visualizes how well a specific DEG subset from a given sample collection period can separate preeclampsia (n = 13, 16, 20, 17) and normotensive (n = 36, 33, 37, 29) samples in all other collection periods (≤12, 13–20, ≥23 weeks of gestation and post-partum respectively). The number of genes identified per sample collection period is highlighted along the main diagonal.
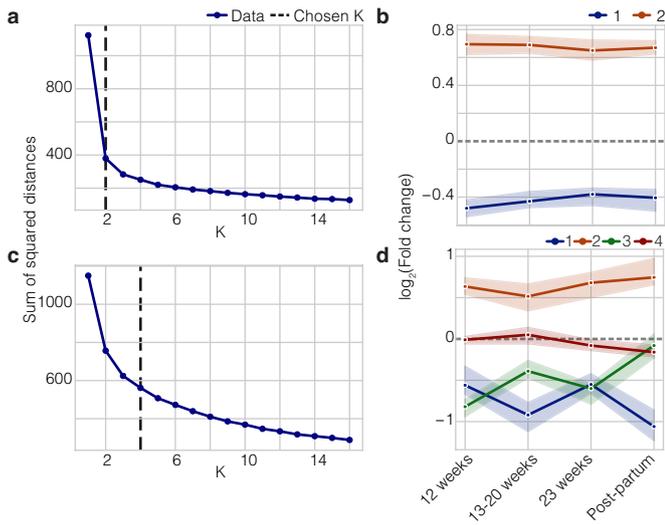
Extended Data Fig. 4 | *k*-means clustering reveals meaningful longitudinal patterns. a, **c**, The chosen *k* clusters (dashed line) comparing a performance metric, the sum of squared distances, and values of *k* for clustering of DEGs for preeclampsia versus normotensive related to Fig. 2d (**a**) and DEGs for preeclampsia with versus without severe features related to Fig. 4a (**c**). **b**, **d**, Following permutation of the data matrix prior to *k*-means clustering, longitudinal changes over gestation are replaced by 2 flat lines for clustering of log$_2$(FC) for preeclampsia versus normotensive (**b**) and 4 uninformative lines for clustering of log$_2$(FC) for preeclampsia with versus without severe features (**d**). For **b**, **d**, points indicate median per trend and shaded region indicates 95% CI.
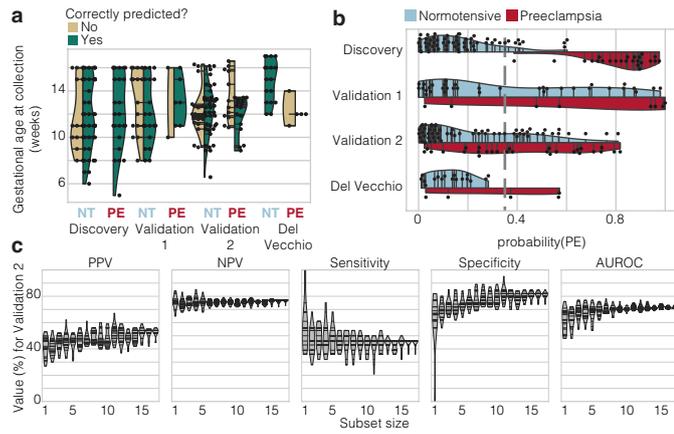
Extended Data Fig. 5 | Examining the logistic regression model used to predict risk of preeclampsia early in gestation. a, Comparison of gestational age at sample collection (weeks) for incorrectly predicted (yellow) or correctly predicted (green) samples across normotensive and preeclampsia groups in the discovery, validation 1, validation 2 and Del Vecchio cohorts shows that incorrectly predicted preeclampsia samples (false negatives) are collected at later gestational ages. **b**, Estimated probability of preeclampsia as outputted by logistic regression for both preeclampsia and normotensive samples shows that the model is well-calibrated across most predictions. Dashed line at 0.35 indicates classifier cut-off where P(PE) ≥ 0.35 constitutes a sample predicted as preeclampsia. **c**, Logistic regression models trained on subsets of 1–18 genes of the initial 18 genes can moderately predict future preeclampsia onset in the validation 2 cohort with improving performance as subset size increases and as characterized by PPV, NPV, sensitivity, specificity and AUROC (left to right).

Extended Data Table 1 | Participant, pregnancy and preeclampsia characteristics across both discovery and validation cohorts

		Maternal pre-pregnancy characteristics					
		Age, years	BMI, kg/m^2	Nulliparous, % (n) *	Smoker, % (n)	History of PTB, % (n) *	History of PE, % (n)
Discovery	NT	32 ± 5	23 ± 3 †, ‡	31 (15)	2 (1)	16 (8)	2 (1)
	PE	31 ± 6	30 ± 8 †	38 (9)	0 (0)	25 (6)	17 (4)
Validation 1	NT	31 ± 5	25 ± 6‡	28 (9)	3 (1)	12 (4)	9 (3)
	PE	34 ± 5	26 ± 4	29 (2)	0 (0)	29 (2)	29 (2)
Validation 2	NT	30 ± 5	NA	0 (0)	5 (3)	0 (0) †	0 (0)
	PE	32 ± 6	NA	0 (0)	0 (0)	50 (13) †	12 (3)

		Maternal ethnicity/race						
		Ethnicity, % (n)* Hispanic	**Race, % (n) *** White	AI/AN	Black	Asian	Other	Unknown
Discovery	NT	22 (11)	73 (36)	0 (0)	4 (2)	20 (10)	2 (1)	0 (0)
	PE	42 (10)	50 (12)	0 (0)	4 (1)	25 (6)	0 (0)	21 (5)
Validation 1	NT	34 (11)	72 (23)	0 (0)	6 (2)	16 (5)	6 (2)	0 (0)
	PE	57 (4)	71 (5)	0 (0)	0 (0)	14 (1)	0 (0)	14 (1)
Validation 2	NT	8 (5)	59 (36)	23 (14)	5 (3)	8 (5)	0 (0)	5 (3)
	PE	4 (1)	62 (16)	23 (6)	4 (1)	4 (1)	0 (0)	4 (1)

		Pregnancy characteristics						
		GA at delivery, weeks *	Mode of delivery % vaginal (n) *	Multi-gestation, % (n)	PTB, % (n) *	Fetal sex, % male (n)	Fetal weight, kg *	SGA, % (n)
Discovery	NT	39 ± 1†	76 (37)	0 (0)	0 †	53 (26)	3.4 ± 0.4 †	0 (0)
	PE	36 ± 3†	46 (11)	12 (3)	38 (9) †	54 (13)	2.8 ± 0.7 †	4 (1)
Validation 1	NT	39 ± 1	56 (18)	3 (1)	0 (0)	53 (17)	3.4 ± 0.5 ‡	0 (0)
	PE	38 ± 1	57 (4)	0 (0)	14 (1)	71 (5)	3.4 ± 0.3	0 (0)
Validation 2	NT	39 ± 1†	75 (46) †	0 (0)	0 (0) †	46 (28)	3.4 ± 0.4 †	0 (0)
	PE	37 ± 2†	35 (9) †	8 (2)	42 (11) †	54 (14)	2.9 ± 0.7 †	8 (2)

		PE characteristics		
		GA at onset, weeks	Onset type, % early (n)	Symptom severity, % severe (n)
Discovery	PE	36 ± 4	25 (6)	54 (13)
Validation 1	PE	38 ± 2	0 (0)	43 (3)
Validation 2	PE	38 ± 2 ‡	12 (3) ‡	50 (13) ‡

Maternal age and BMI, gestational age (GA) at delivery, fetal weight and GA at preeclampsia onset are reported as mean±s.d. All other values are reported as percentages with the corresponding count in parentheses. Small for GA (SGA) was defined as an infant with a birthweight below the 10th centile for their GA at delivery. Pre-pregnancy BMI was not available for individuals in the validation 2 cohort. See Supplementary Table 2 for exact P values. AI/AN, American Indians and Alaska Natives.

*$P \le 0.05$, two-sided chi-squared (categorical) or ANOVA (continuous) test comparing all cohorts with Bonferroni correction.

†$P \le 0.05$, two-sided chi-squared (categorical) or ANOVA (continuous) test comparing preeclampsia and normotensive within each cohort with Bonferroni correction.

‡ denotes that missing values were omitted from reported values for a given feature.

Article

Extended Data Table 2 | Tissue, cell types, and genes previously implicated in preeclampsia enriched in 544 DEGs identified when comparing preeclampsia with or without severe features and normotensive pregnancies

K-means cluster	Specific to	Reference	Adjusted p -value	DEGs specific to cell-type or tissue
1, Decreased in PE	granulocytes	HPA	0.00026	*ICAM3, PTPN6, SNX20, VNN2, LIMD2*
	lymph node	HPA	0.013	*CYTIP, CD48*
	T-cells	HPA	0.02	*TBC1D10C, TRAF3IP3*
2, Increased in PE	platelet	TSP+	2.5x10^{-33}	*GP9, SEPT5, AQP10, TREML1, CMTM5, GRAP2, RP11 - 142C4.6, BMP6, TUBB1, PTGS1, GP1BA, C6orf25, ITGA2B, PTCRA, ITGB3, RP3 -370M22.8, BEND2, SEC14L5, ALOX12, F13A1,GP6, SLA2*
	smooth muscle	HPA	0.00026	*MYL9, PTGS1, MYLK, SYNM, FLNA*
	total PBMC	HPA	0.00077	*GP9, TREML1, TUBB1*
	oligodendrocytes	TSP+	0.005	*PCSK6, SH3TC2*
	granulocytes	HPA	0.02	*BIN2, GFI1B, ITGA2B, MTRNR2L12, NFE2*
	excitatory neurons	TSP+	0.02	*MTURN, NRGN*
	mesenchymal stem cell	TSP+	0.03	*COL6A3, NID1*
	capillary aerocyte	TSP+	0.03	*CLDN5, ESAM*
	PE	dbPEC	0.006	*CLU, ICAM3, TGFB1, IGF2, ADIPOR2, CDKN1A, GPX1, TGFBR1, RXRA, IFNGR2, JUN, PTGS1, F2R, NFKB1, GAPDH, SELP, NFKB2, ILK, GP9, ITGB3, CLDN5, ITGB1, ATP2A3, PDGFA, EGLN1, EGLN2, GP1BA, AC TB, VDR, CASP8*

For every significantly enriched tissue or cell type (*P* ≤ 0.05; one-sided hypergeometric test with Benjamini–Hochberg correction), assigned *k*-means cluster (Fig. 2d), reference and adjusted *P* values are reported from left to right. The rightmost column lists the gene names for all DEGs that were labelled as specific to a given cell type or tissue. The last row lists gene set enrichment with dbPEC, a preeclampsia-specific database of genes.

Extended Data Table 3 | Preeclampsia prediction performance metrics for samples collected early in gestation (5–16 weeks)

	Control (n)	Case (n)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	AUROC
Discovery	61(9)	24 (0)	100 [92-100]	85 [77-91]	73 [59-84]	100 [96-100]	0.99 [0.99-0.99]
Validation 1	35 (15)	8 (2)	75 [46-92]	56 [43-70]	28 [15-46]	91 [77-97]	0.71 [0.7-0.72]
Validation 2	61 (19)	28 (12)	56 [42-72]	69 [59-78]	46 [32-59]	78 [68-86]	0.72 [0.71-0.72]
Del Vecchio[1]	17 (0)	5 (2)	60 [26-87]	100 [89-100]	100 [56-100]	89 [74-97]	0.74 [0.73-0.75]
Del Vecchio[2]	8 (0)	5 (2)	60 [26-87]	100 [79-100]	100 [56-100]	80 [55-94]	0.80 [0.79-0.81]

Control and case sample numbers are reported as the total sample number and, in parentheses, the number of samples misclassified. All other statistics including sensitivity specificity, PPV, NPV and AUROC are reported as the estimated percentage followed by the 90% CI in square brackets. In Del Vecchio[1], the control group is defined as samples from any pregnant mother who did not develop preeclampsia, including those with other underlying or pregnancy-related complications such as chronic hypertension or gestational diabetes, respectively. In Del Vecchio[2], the control group is defined as samples strictly from normotensive pregnant mothers who did not experience complications.

Article

Extended Data Table 4 | Preeclampsia prediction relies on the cfRNA levels of 18 genes

Gene	ENSEMBL	Name	Odds ratio	Biological process [GO]	Molecular function [GO]
CAMK2G	ENSG00000148660	calcium/calmodulin dependent protein kinase II gamma	0.65	nervous system development, protein phosphorylation, regulation of neuron projection development, regulation of skeletal muscle adaptation	calcium-dependent protein serine/threonine phosphatase activity, identical protein binding, protein homodimerization activity
DERA	ENSG00000023697	deoxyribose-phosphate aldolase	1.07	pentose-phosphate shunt, deoxyribose phosphate catabolic process	deoxyribose-phosphate aldolase activity, protein binding
FAM46A	ENSG00000112773	terminal nucleotidyltransferase 5A	0.91	mRNA stabilization, response to bacterium	RNA adenylyltransferase activity, protein binding, RNA binding
KIAA1109	ENSG00000138688	KIAA1109	0.89	synaptic vesicle endocytosis, regulation of epithelial cell differentiation, endosomal transport	protein binding
LRRC58	ENSG00000163428	leucine rich repeat containing 58	0.79		
MYLIP	ENSG00000007944	myosin regulatory light chain interacting protein	0.94	nervous system development, regulation of low-density lipoprotein particle receptor catabolic process	ubiquitin protein ligase activity, cytoskeletal protein binding, protein binding, metal ion binding
NDUFV3	ENSG00000160194	NADH:ubiquinone oxidoreductase subunit V3	0.66	mitochondrial electron transport, NADH to ubiquinone	NADH dehydrogenase (ubiquinone) activity, protein binding, RNA binding
NMRK1	ENSG00000106733	nicotinamide riboside kinase 1	1.42	NAD metabolic process, phosphorylation, NAD biosynthetic process	ribosylnicotinamide kinase activity
PI4KA	ENSG00000241973	phosphatidylinositol 4 kinase alpha	0.9	multi-organism membrane organization, phosphorylation, viral replication complex formation and maintenance	kinase activity, phosphatidylinositol kinase activity, protein binding
PRTFDC1	ENSG00000099256	phosphoribosyl transferase domain containing 1	1.32	guanine salvage, GMP catabolic process, purine ribonucleoside salvage	protein homodimerization activity, hypoxanthine phosphoribosyltransferase activity
PYGO2	ENSG00000163348	pygopus family PHD finger 2	0.95	positive regulation of chromatin binding, developmental growth, regulation of histone H3-K4 methylation, mammary gland development	chromatin binding, protein binding, metal ion binding, histone acetyltransferase regulator activity, histone binding
RNF149	ENSG00000163162	ring finger protein 149	0.82	ubiquitin-dependent protein catabolic process, regulation of protein stability, protein ubiquitination	ubiquitin protein ligase activity, metal ion binding
TFIP11	ENSG00000100109	tuftelin interacting protein 11	0.9	biomineral tissue development, negative regulation of protein binding, spliceosomal complex disassembly	protein binding, nucleic acid binding
TRIM21	ENSG00000132109	tripartite motif containing 21	0.88	negative regulation of protein deubiquitination, regulation of protein localization, regulation of gene expression	zinc ion binding, transcription coactivator activity, identical protein binding
USB1	ENSG00000103005	U6 snRNA biogenesis phosphodiesterase 1	0.85	U6 snRNA 3'-end processing, RNA splicing	poly(U)-specific exoribonuclease activity
Y_RNA	ENSG00000201412	Y RNA	1.85		
Y_RNA	ENSG00000238912	Y RNA	1.08		
YWHAQP5	ENSG00000236564	YWHAQ pseudogene 5	1.09		

For every gene, symbol, ENSEMBL ID, full name, odds ratio (OR) based on the logistic regression coefficient and, if available, a subset of GO biological processes and molecular functions (if available) are reported from left to right.

Extended Data Table 5 | Tissue and cell types enriched in 503 DEGs identified when comparing preeclampsia with as compared to without severe features

K-means cluster	Specific to	Reference	Adjusted p-value	DEGs specific to cell-type or tissue
1, Increased across gestation for severe PE	naive b cell	TSP+	0.0002	*TCL1A, TNFRSF13C*
	B-cells	HPA	0.002	*FCMR, TCL1A*
	neutrophil	TSP+	0.002	*PADI4, MEFV, DOK3*
	bone marrow	HPA	0.01	*PADI4, DOK3*
	endothelial cell	TSP+	0.04	*PDE2A, VWF*
2, Increased and then decreased in early and late gestation respectively for severe PE	granulocytes	HPA	0.0007	*SELL, HK3, MXD1, ALPL*
	neutrophil	TSP+	0.02	*FCN1, HK3*
4, Decreased across gestation for severe PE	thymus	HPA	0.01	*CARMIL2, ARHGAP15*

For every significantly enriched tissue or cell type ($P \le 0.05$; one-sided hypergeometric test with Benjamini–Hochberg correction), assigned *k*-means cluster (Fig. 4a), reference and adjusted *P* values are reported from left to right. The rightmost column lists the gene names for all DEGs that were labelled as specific to a given cell type or tissue.

nature portfolio

Corresponding author(s): Stephen R. Quake

Last updated by author(s): Jan 23, 2022

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used.
Data analysis	Bioinformatic processing: For each sample, raw sequencing reads were trimmed using Trimmomatic (v 0.36) and then mapped to the human reference genome (hg38) with STAR (v 2.7.3a). Duplicate reads were then removed by GATK's (v 4.1.1) MarkDuplicates tool. Finally, mapped reads were sorted and quantified using htseq-count (v 0.11.1) generating a counts table (genes x samples). Read statistics were estimated using FastQC (v 0.11.8). Across samples, the bioinformatic pipeline was managed using Snakemake (v 5.8.1). Read and tool performance statistics were aggregated across samples and steps using MultiQC (v 1.7). Data analysis: Analyses were performed in Python (v 3.6) or R (v 3.5). In python, we used Scikit-learn (v 0.23.2), Scipy (v 1.5.1), nheatmap (v 0.1.4), Kneed (v 0.7.0), Seaborn (v 0.10.0), and GProfiler (v1.0.0). In R, we used Limma (v 3.38.3). All code are available on Github at https://github.com/miramou/pe_cfrna.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Raw and processed sequencing data (BioProject PRJNA792450) are available via the SRA (SRP352519) and GEO (GSE192902), respectively. Data were mapped using the human reference genome (hg38) and annotated using ENSEMBL version 82. We used publicly available data from the HPA (v19, https://v19.proteinatlas.org/), TSP (v1.0, https://tabula-sapiens-portal.ds.czbiohub.org/), Gene ontology: biological processes and cellular compartments (GO:BP, GO:CC, released 2021-05-01), Reactome (REAC, released 2021-05-07), and Kyoto Encyclopedia of Genes and Genomes (KEGG, released 2021-05-03), and publications 43-48.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences   ☐ Behavioural & social sciences   ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	To identify changes associated with PE well before traditional diagnosis, we designed a prospective study and recruited pregnant mothers at their first clinical visit to Stanford's Lucile Packard Children's Hospital, between 5–12 weeks of gestation, of which 131 were included in this study (94 NT, 37 with PE) prior to sample QC. For each participant, we analyzed cfRNA for samples collected at multiple time points. We also obtained samples from an independent cohort collected at several independent institutions (Validation 2), which consisted of 89 samples collected prior to 16 weeks of gestation from 87 mothers (61 NT, 26 PE). For Discovery and Validation 1, we included all enrolled mothers who developed PE and had at least 1 mL banked blood plasma sample remaining. For Validation 2, we included all enrolled mothers who developed PE and had at least 1 mL banked blood plasma sample collected at ≤16 weeks of gestation remaining. In all cases, we chose a case:control ratio of approximately 1:2 to increase statistical power. We also ensured that case and control groups were matched for race and ethnicity. No other matching or exclusion criterion were used, and we performed no further sample selection prior to sample processing.
Data exclusions	Our final analysis included a subset of the available samples that passed pre-defined quality metrics as discussed in the Supplement (Supp. Note 1, Methods, Extended Data Fig. 1). Quality control was performed on a per sample as opposed to per subject basis and so subjects had a variable number of samples that passed QC (Supplementary Table 1).
Replication	To verify our results, we split our larger Stanford cohort into two groups (Discovery, Validation 1) and further replicated any reported findings using 2 independent cohorts (Validation 2, Del Vecchio). The larger Stanford cohort was split into Discovery (n = 88, [60 NT, 28 with PE]) and Validation 1 (n = 43, [34 NT, 9 with PE]) cohorts. We also further replicated our results using 2 independent cohorts: Validation 2 (n = 87, [61 NT, 26 with PE]) and Del Vecchio (n = 22, [8 NT, 5 with PE, 7 with gestational diabetes, 2 with chronic hypertension]), a previously published cohort processed by a completely separate team (Del Vecchio et al. 2020). In all cases, we successfully replicated our initial findings.
Randomization	To allocate samples to the control group (normotensive) or case group (preeclamptic), we relied on current clinical guidelines. Specifically, we defined a pregnancy as normotensive if it was both uncomplicated and went to full-term (≥ 37 weeks) and as preeclamptic based on current clinical guidelines (see Methods). We also explicitly tested for covariate differences between both groups and describe this both in the main text (see section titled Clinical study design, Fig. 1) and extended data (see Methods, Extended Data Table 1). To split our larger Stanford cohort into Discovery and Validation 1, we first allocated samples using sequencing batch of which there were 3. We allocated the sequencing batch with the most PE samples to Discovery to ensure sufficient statistical power and the second most PE samples to Validation 1. Sequencing batches themselves contained randomly allocated samples based on subject ID such that all samples from the same subject were in the same batch. For the final sequencing batch, we randomly allocated subjects to either Discovery or Validation 1 such that all samples from 1 subject were part of the same group (either Discovery or Validation 1) and we maintained at least a 1:2 case to control ratio in both groups.
Blinding	Investigators were blinded during data collection as pregnancy outcomes were not known yet. Investigators were not blinded during data analysis as analysis methods required knowledge of outcome (i.e., supervised learning).

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	Antibodies
☒	Eukaryotic cell lines
☒	Palaeontology and archaeology
☒	Animals and other organisms
☐ ☒	Human research participants
☒	Clinical data
☒	Dual use research of concern

Methods

n/a	Involved in the study
☒	ChIP-seq
☒	Flow cytometry
☒	MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	The cohorts described herein are composed of pregnant mothers (aged 18 years or older) of varied ethnicity, race, age, and BMI among other pre-pregnancy and pregnancy characteristics (Results "Clinical study design", Fig. 1, Extended Data Table 1).
Recruitment	Discovery and Validation 1 were collected as part of a longitudinal, prospective study at Stanford University. We enrolled pregnant mothers (aged 18 years or older) receiving routine antenatal care on or prior to 12 weeks of gestation at Lucile Packard Children's Hospital at Stanford University, following study review and approval by the Institutional Review Board (IRB) at Stanford University (21956). All signed informed consent prior to enrollment. Mothers who agreed to participate in the study after enrollment at their first clinical visit were asked to complete a questionnaire available in multiple languages. Samples were then collected at routine care visits.
	Validation 2 was collected as part of the Global Alliance to Prevent Preterm and Stillbirth (GAPPS) Pregnancy Biorepository at Yakima Valley Memorial Hospital, Swedish Medical Center, and the University of Washington Medical Center under review and approval by Advarra IRB (CR00195799). Mothers were enrolled during routine clinical care visit and then asked to fill out three questionnaires. Samples were drawn during routine care visits, and stored appropriately. Samples were processed and sequenced at Stanford under the same IRB as above (21956). All signed informed consent prior to enrollment.
	In all cohorts, we recruited mothers who came in for routine care. Consequently, participant characteristics in the study reflect the socioeconomic, racial, and ethnic makeup of the larger area around the recruitment site. This presents a possible bias, which we focused on mitigating by reproducing our results using blood samples collected at different institutions (Validation 1, Validation 2, and Del Vecchio) with correspondingly distinct patient populations.
Ethics oversight	Institutional Review Board at Advarra (CR00195799) and Stanford University (21956)

Note that full information on the approval of the study protocol must also be provided in the manuscript.