Spring 2022 – Systems Biology of Reproduction
Discussion Outline – Fetal Development & Birth Systems
Michael K. Skinner – Biol 475/575
CUE 418, 10:35-11:50 am, Tuesday & Thursday
April 21, 2022
Week 15

Fetal Development & Birth Systems

Primary Papers:

1. Glotov, et al. (2015) BMC Systems Biology 9(Suppl 2):S4
2. Pique-Regi, et al. (2019) eLife 8:e52004
3. Winchester, et al. (2022) Scientific Reports 12:3361

Discussion

Student 9: Reference 1 above

- What major diseases are compared with preeclampsia and why?
- What networks were identified and impact?
- What risk factors were identified?

Student 10: Reference 2 above

- What technical approach was used and types of correlations?
- What transcriptome and cellular correlations were made?
- What major insights were provided for preterm birth?

Student 11: Reference 3 above

- What technical approach was used for the study?
- What observations were made for mother, father, child?
- What potential impact to reduce preterm birth?

Glotov et al. BMC Systems Biology 2015, 9(Suppl 2):S4
http://www.biomedcentral.com/1752-0509/9/S2/S4

BMC Systems Biology

RESEARCH Open Access

Molecular association of pathogenetic contributors to pre-eclampsia (pre-eclampsia associome)

Andrey S Glotov[1,2], Evgeny S Tiys[3,4], Elena S Vashukova[1], Vladimir S Pakin[1], Pavel S Demenkov[3,4], Olga V Saik[3,4], Timofey V Ivanisenko[3,4], Olga N Arzhanova[1], Elena V Mozgovaya[1], Marina S Zainulina[1], Nikolay A Kolchanov[3,4], Vladislav S Baranov[1,2], Vladimir A Ivanisenko[3,4*]

From IX International Conference on the Bioinformatics of Genome Regulation and Structure\Systems Biology (BGRS\SB-2014)
Novosibirsk, Russia. 23-28 June 2014

Abstract

Background: Pre-eclampsia is the most common complication occurring during pregnancy. In the majority of cases, it is concurrent with other pathologies in a comorbid manner (frequent co-occurrences in patients), such as diabetes mellitus, gestational diabetes and obesity. Providing bronchial asthma, pulmonary tuberculosis, certain neurodegenerative diseases and cancers as examples, we have shown previously that pairs of inversely comorbid pathologies (rare co-occurrences in patients) are more closely related to each other at the molecular genetic level compared with randomly generated pairs of diseases. Data in the literature concerning the causes of pre-eclampsia are abundant. However, the key mechanisms triggering this disease that are initiated by other pathological processes are thus far unknown. The aim of this work was to analyse the characteristic features of genetic networks that describe interactions between comorbid diseases, using pre-eclampsia as a case in point.

Results: The use of ANDSystem, Pathway Studio and STRING computer tools based on text-mining and database-mining approaches allowed us to reconstruct associative networks, representing molecular genetic interactions between genes, associated concurrently with comorbid disease pairs, including pre-eclampsia, diabetes mellitus, gestational diabetes and obesity. It was found that these associative networks statistically differed in the number of genes and interactions between them from those built for randomly chosen pairs of diseases. The associative network connecting all four diseases was composed of 16 genes (*PLAT, ADIPOQ, ADRB3, LEPR, HP, TGFB1, TNFA, INS, CRP, CSRP1, IGFBP1, MBL2, ACE, ESR1, SHBG, ADA*). Such an analysis allowed us to reveal differential gene risk factors for these diseases, and to propose certain, most probable, theoretical mechanisms of pre-eclampsia development in pregnant women. The mechanisms may include the following pathways: [TGFB1 or TNFA]-[IL1B]-[pre-eclampsia]; [TNFA or INS]-[NOS3]-[pre-eclampsia]; [INS]-[HSPA4 or CLU]-[pre-eclampsia]; [ACE]-[MTHFR]-[pre-eclampsia].

Conclusions: For pre-eclampsia, diabetes mellitus, gestational diabetes and obesity, we showed that the size and connectivity of the associative molecular genetic networks, which describe interactions between comorbid diseases, statistically exceeded the size and connectivity of those built for randomly chosen pairs of diseases. Recently, we have shown a similar result for inversely comorbid diseases. This suggests that comorbid and inversely comorbid diseases have common features concerning structural organization of associative molecular genetic networks.

* Correspondence: salix@bionet.nsc.ru
[3]The Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
Full list of author information is available at the end of the article

BioMed Central

© 2015 Glotov et al.; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

Background

Pre-eclampsia (PE) is the leading cause of maternal and foetal morbidity and mortality. It is a pregnancy complication, predominantly occurring after 20-weeks of gestation, as well as in labour, and it is characterized by multiple organ dysfunction syndromes, including the dysfunction of the kidneys, liver, vascular and nervous systems, and the foetoplacental complex [1,2]. The general clinical symptoms of PE are oedema, proteinuria and hypertension. The clinical outcome of PE may not always be predictable. Either form of PE can be extremely insidious, rapidly progressing, and, even in the absence of one of its general symptoms, may lead to life threatening complications for the mother and foetus [3]. In 70-80% of cases, PE is secondary to an underlying disease [1]. Pre-eclampsia risk factors include cardiovascular diseases (arterial hypertension), kidney, liver and gastrointestinal tract diseases, endocrine disorders (obesity, diabetes mellitus), and autoimmune diseases (antiphospholipid syndrome) [1,3,4]. According to meta-analysis data, women with a history of PE have 1.79 times the risk of venous thromboembolism, 1.81 times the risk of stroke, 2.16 times the risk of ischemic heart disease and 3.7 times the risk of hypertensive disease, compared with women without PE [5]. Thus far, it remains unclear whether the presence of pathological processes before pregnancy predisposes one to PE, or whether defects in multiple organs and systems, induced by PE, are responsible for the development of extragenital diseases in the future. Such joint manifestations of diseases are called comorbidities [6] or syntropies [7]. Likewise, inversely comorbid [8] or dystropic [9] diseases statistically rare co-occur in patients as compared with co-occurrence that can be expected by chance. Previously, for asthma, tuberculosis, certain cancers and neurodegenerative diseases, we have shown that inversely comorbid diseases are more closely related to each other at the molecular level in comparison with randomly chosen pairs of diseases [10].

In recent years, bioinformatics methods have been widely used for modelling different pathological processes, analysing the molecular mechanisms of their development, identifying possible markers, and systematizing available data. Ample evidence regarding the influence of genetics on comorbidities has accumulated in the literature. Computer-based, text-mining methods were developed for efficient extraction of knowledge from the scientific literature. At the present time, COREMINE and MeSHOPs, which analyse the co-occurrence of biomedical terms [11,12], and STRING [13] and the MedScan system, which are based on the parsing of natural language texts [14], are widely used.

We have developed the ANDSystem, which was designed for the automated extraction of knowledge from natural language texts regarding the properties of molecular biological objects and their interactions in living systems [15]. Using this system, we have reconstructed protein-protein networks for proteins that are associated with water-salt metabolism and sodium deposition processes in healthy volunteers [16], as well as protein-protein interaction networks for *Helicobacter pylori*, which are associated with the functional divergence of *H. pylori*, isolated from patients with early gastric cancer [17]. We have also reconstructed associative networks representing molecular genetic interactions between proteins, genes, metabolites and molecular processes associated with myopia and glaucoma [18], and with cardiovascular diseases [19].

In the current study, we used the ANDSystem for the reconstruction of associative networks (the preeclampsia associome) representing molecular genetic interactions between genes associated with PE, diabetes mellitus (DM), gestational diabetes (GD) and obesity (Ob). We conducted an analysis of these networks to reveal differential and common risk factors for these diseases.

Finding pathways common to the indicated multifactorial diseases would contribute to a better understanding of the characteristic features of pre-eclampsia pathogenesis, as well as to the development of new diagnostic, preventative and therapeutic methods.

Results

Pre-eclampsia: its association, via comorbid genes, with diabetes mellitus, obesity and gestational diabetes

The main goal of the current study was to identify comorbid genes whose dysfunction or mutation represent common risk factors for diseases that are concurrent with PE. To this end, we relied on published data [3,4] regarding the four most significant and widespread pathologies concurrent with PE: DM, Ob, GD and pyelonephritis. Furthermore, using the ANDSystem and Pathway Studio software, we reconstructed associative networks (disease-protein/gene-disease) comprising interactions between the above diseases via their associated genes. Subsequently, reduction was achieved by eliminating pyelonephritis, as genes associated with nephritis were not associated with PE and the other analysed disorders. Using the ANDSystem, we identified 1,051 proteins/genes associated with PE, Ob, DM and/or GD. Using Pathway Studio, 1,138 proteins/genes were identified. The results of both programs were in good agreement regarding the number of genes in groups associated with particular diseases (Figure 1). Unfortunately, we were not able to use STRING for the reconstruction of such networks, as this program does not provide information about protein/gene-disease associations.

The number of proteins/genes common to different combinations of the examined diseases is shown in Figure 1. We assumed that comorbid diseases are more

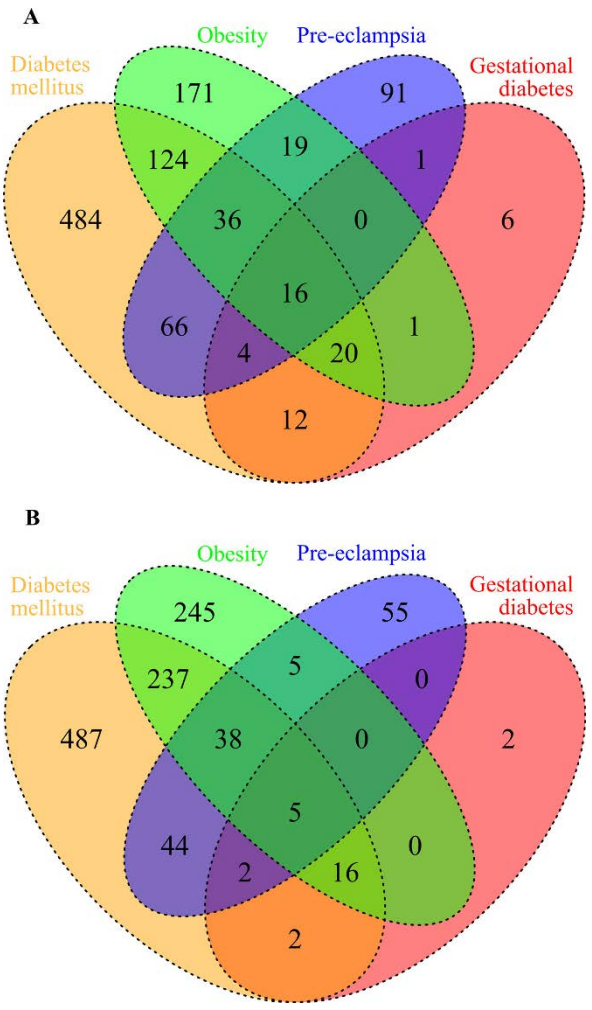


Figure 1 Venn diagram demonstrating the intersections of the lists of genes associated with the analysed diseases (PE, DM, GD, Ob) according to the ANDSystem (A) and Pathway Studio (B).

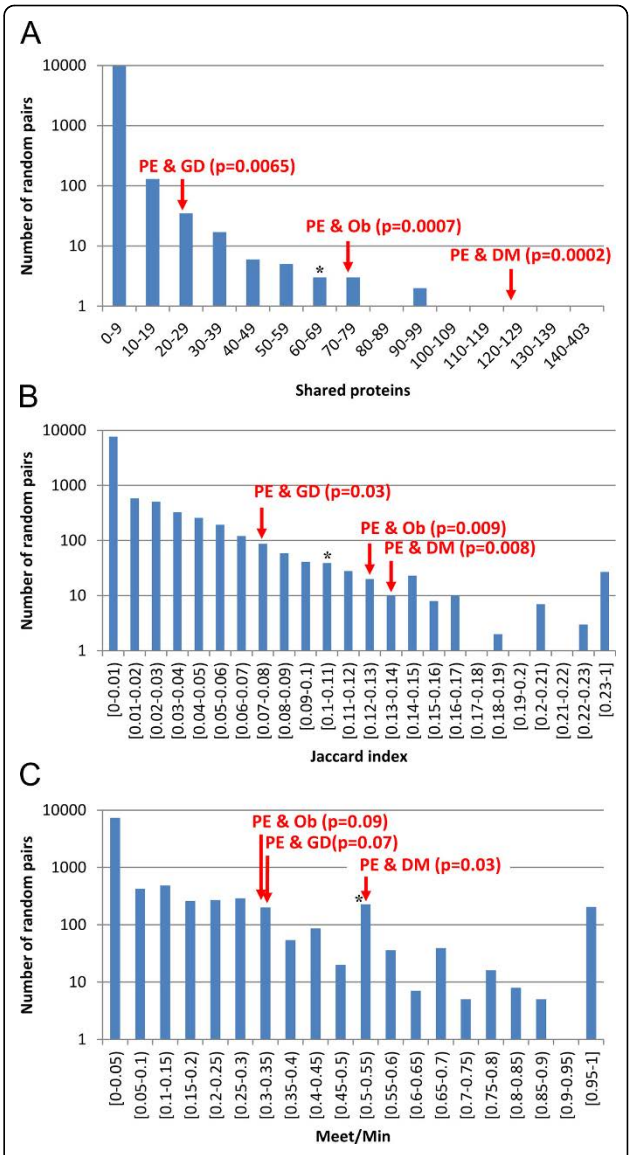


Figure 2 Comparison of analysed and random networks by intersection (A), Jaccard (B) and Meet/Min (C) indices. Bars show the distribution of the value for the features of the associative networks for randomly chosen disease pairs. Arrows indicate PE & GD, PE & DM and PE & Ob comorbid disease pairs. Asterisks indicate the position of inversely comorbid disease pairs (bronchial asthma and pulmonary tuberculosis) [10].

closely interrelated, via the common proteins/genes associated with them, as compared with randomly chosen disease pairs. To test this assumption, we calculated the distribution of three relation indices of random disease-protein/gene-disease networks built for random disease pairs: I_{AB} (number of shared proteins), J_{AB} (Jaccard index) and M_{AB} (Meet/Min). All three disease pairs (PE & DM, PE & GD, PE & Ob) were significantly connected by the I_{AB} and J_{AB} indices at $p < 0.05$ (Figure 2). Only PE & DM pair was significantly different by M_{AB} index ($p < 0.05$) from randomly generated pairs of diseases. Thus, PE and DM were found to be the most significantly associated disease pair for all three relation indices.

Next, we tested the hypothesis whether comorbid proteins/genes common to comorbid disease pairs interact more closely compared to a set of randomly chosen proteins/genes. Comparison of the associative molecular genetic networks with random ones demonstrated that the networks that describe the interactions between the comorbid proteins/genes for all three disease pairs (PE & GD, PE & DM, and PE & Ob) exhibited significantly greater connectivities than those of the random networks ($p < 0.001$).

Of particular interest was an appended analysis of the associative molecular genetic networks built for proteins associated concurrently with four comorbid diseases (PE, DM, Ob and GD). The three programs used to

build this network were the ANDSystem, Pathway Studio and STRING (Figure 3). As Figure 3A shows, the ANDSystem network comprised 32 objects: 16 proteins and 16 genes, as well as 142 interactions. The ANDSystem has an advantageous feature: an object pair can also be associated concurrently with links of several types. For this reason, the number of associated object pairs, 87, was smaller than the number of links. The ANDSystem

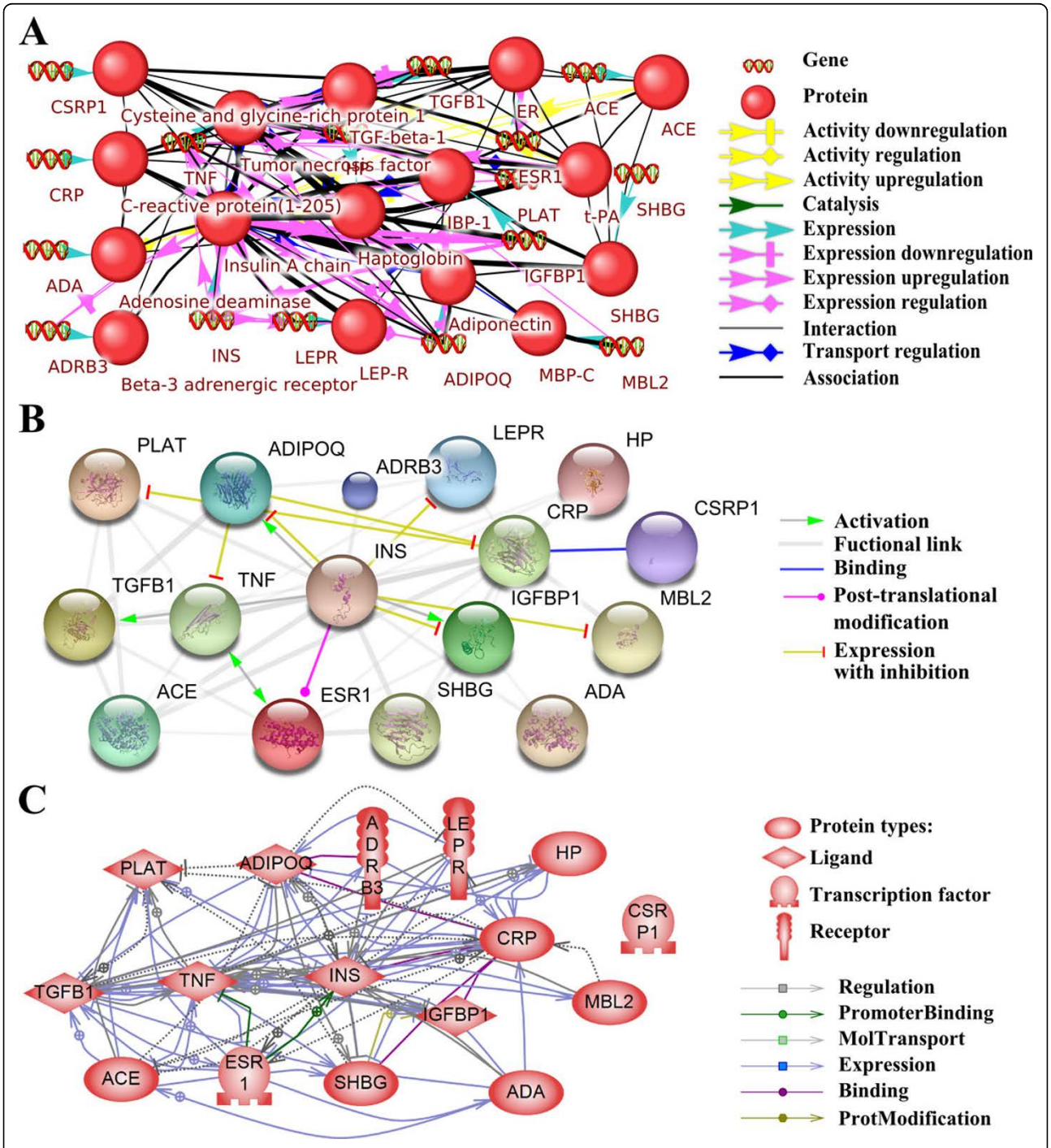


Figure 3 ANDSystem (A), STRING (B) and Pathway Studio (C) networks describing shared proteins/genes. Interactions between proteins/genes concurrently shared in PE, DM, Ob and GD are shown. In the ANDSystem network, proteins are shown as balls, and patterns with double helices designate genes. In the STRING network, proteins/genes are shown as balls (gene names are given). The colour and shape of the arrows indicate the type of association between the objects in all the networks.

represented cases of the regulation of protein activity (six links), including up-regulation (two links) and down-regulation (three links) of protein activity; gene expression regulation (37 links), including up-regulation (seven links) and down-regulation (seven links); protein-protein interactions (two links); protein transport regulation (10 links); catalysis (one link); expression (16 links) and association (70 links). To compare the ANDSystem network with those of the STRING and Pathway Studio, the ANDSystem network was transformed into a protein-protein interaction network, with links from the genes assigned to their respective proteins, while links from genes as separate vertices were deleted from the network. Such a network contained 45 interconnected protein pairs.

The STRING network (Figure 3B) contained 16 proteins/genes, and 45 gene pairs connected by 47 links, including five different types: activation (four links), expression with inhibition (seven links), binding (one link), post-translational modification (one link), and functional links (34 links). The functional links in STRING were determined on the basis of Neighbourhood in the Genome, Gene Fusions, Co-occurrence Across Genomes, Co-Expression, Experimental/Biochemical Data, Association in Curated Databases, and Co-Mentioned in PubMed Abstracts [13].

The network built by Pathway Studio (Figure 3C) contained 16 proteins/genes, and 62 pairs of genes connected by 98 links, including six different types: binding (five links), expression (55 links), molecular transport (19 links), promoter binding (two links), protein modification (one link) and regulation (16 links).

There was a significant difference between the comorbid and random networks ($p < 0.001$), not only for disease pairs, but also for the associative molecular genetic networks that describe the interactions between proteins/genes associated concurrently with all four diseases (PE, DM, GD, Ob) (Figure 3A). These results demonstrated that comorbid proteins/genes are presumably involved in shared biological processes. This can explain the increased number of interactions between proteins/genes, as compared with those for associative molecular genetic networks of randomly chosen proteins/genes. Confirmation of this hypothesis would shed light on the molecular mechanisms underlying the interactions between comorbid diseases.

Analysis of overrepresentation of Gene Ontology (GO) processes

Overrepresentation of GO biological processes was analysed for the group of proteins/genes associated with single diseases (PE, DM, GD and Ob) and pairs of diseases (PE & DM, PE & GD, PE & Ob), as well as concurrently with all four diseases. In each of these cases, more than 1,000 overrepresented processes were found (Additional file 1). Among these were a high number of quite general processes for which thousands of genes have been annotated. The connectivity rate (CR) was calculated for each process listed in Additional file 1 to check how closely the proteins/genes, which caused an overrepresentation of processes, interacted. After ranking the overrepresented biological processes according to the CRs, 313 processes had the highest CR (equal to 1) (see Additional file 1). Just as expected, generalized, nonspecific biological processes had smaller CR values in the majority of cases as compared with specialized processes involving a relatively small number of genes.

Among the overrepresented biological processes with a maximum CR were positive regulation of monooxygenase activity, regulation of fat cell differentiation, regulation of lipid metabolic process, nitric oxide and carbon monoxide metabolism, regulation of protein kinase B signalling cascade, regulation of NF-kappa B transcription factor activity, regulation of glucose metabolism and transport, regulation of cellular response to oxidative stress, regulation of cytokine production, regulation of cell cycle process and others. Thus, the use of the CR index in the GO enrichment analysis revealed the specific GO processes and lower the rank of less informative general processes.

Reconstruction of associative pathways describing potential molecular mechanisms via comorbid genes involved in overrepresented GO biological processes

The next step of the current study was to reconstruct the molecular pathways connecting PE with DM, Ob, and GD, via interactions between the specific and comorbid genes. The Pathway Discovery module of the ANDVisio software was used to trace separate pathways in the network of molecular genetic interactions associated concurrently with all four pathologies. The Pathway Discovery module was used to search for pathways in the network using patterns set by the user.

The patterns were of the following type: <PE> - <any protein/gene specific to PE> - <any comorbid protein/gene> - <any protein/gene specific to Ob or GD, or DM> - <Ob or GD, or DM>. The program chose all the pathways meeting the pattern's criteria: the starting link was PE; the second link of the chain should be one of the proteins/genes associated with PE, exceptions were proteins/genes comorbid for all four diseases (4-comorbid); the third link should be one of the 4-comorbid proteins/genes (PLAT, ADIPOQ, ADRB3, LEPR, HP, TGFB1, TNFA, INS, CRP, CSRP1, IGFBP1, MBL2, ACE, ESR1, SHBG, ADA); the fourth link should be one of the proteins specific to Ob, GD, or DM, with the exception of 4-comorbid proteins/genes. The last link should be one of the diseases (Ob, GD or DM). The total number of identified pathways was more than 50. These were combined into a single pathway network (Figure 4).

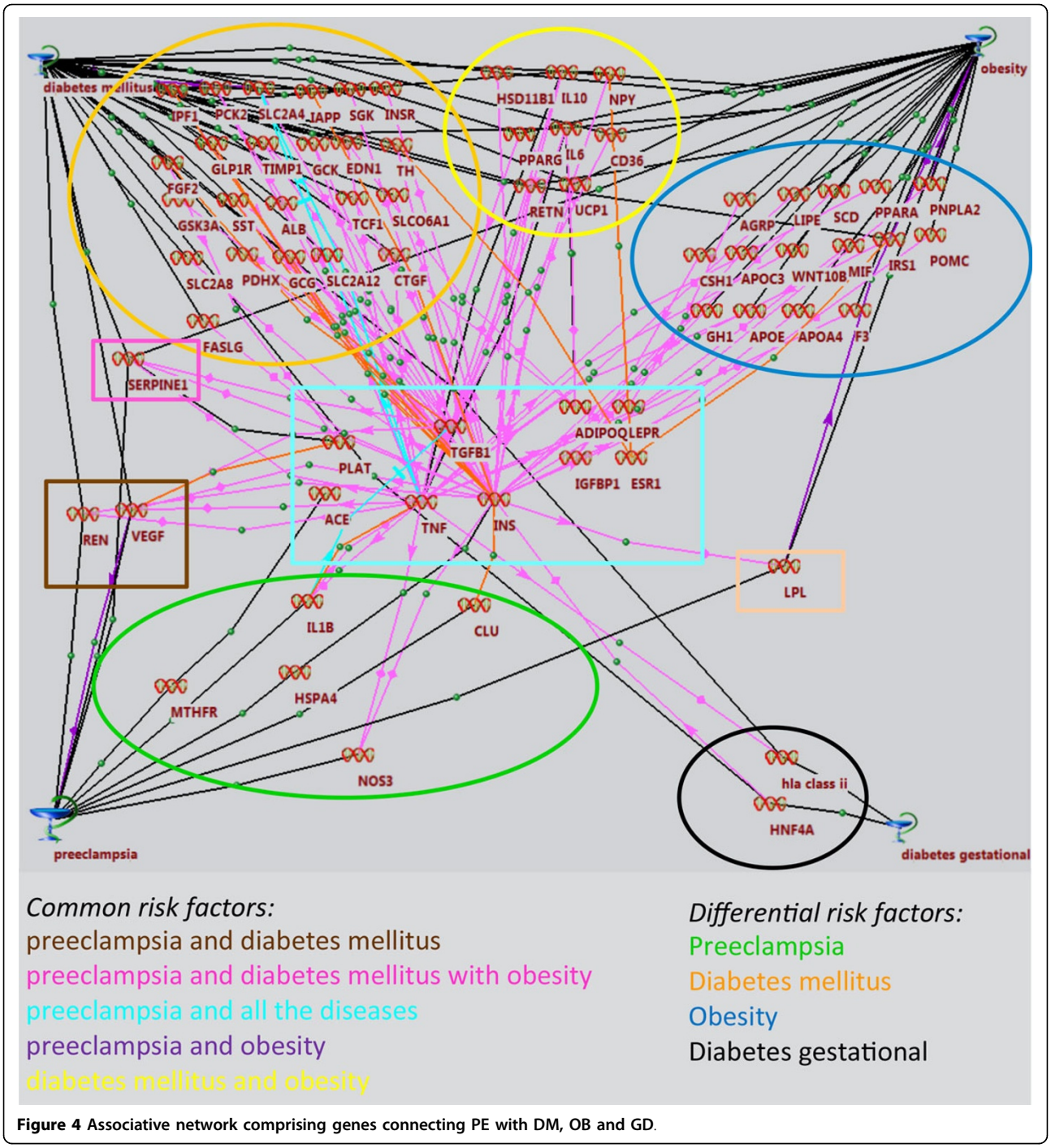


Figure 4 Associative network comprising genes connecting PE with DM, OB and GD.

Common, as well as specific, risk factors were distinguished for the following combinations of diseases: PE and DM; PE and Ob; PE and DM, Ob; PE and DM, Ob, GD; (see Figure 4). The largest number of connections was obtained for the *TNFA, TGFB1* and *INS* genes, which revealed specialized GO processes with maximum CRs, such as: «positive regulation of protein kinase B signalling», «cascade regulation of NF-kappa B transcription factor activity», «regulation of mitosis», «regulation of nuclear division», «regulation of protein secretion MAPK cascade», «positive regulation of protein transport», «regulation of protein complex assembly», «positive regulation of cell migration», «positive regulation of secretion», «positive regulation of cellular component movement», «positive regulation of organelle organization», «regulation of mitotic cell cycle»,

«regulation of immune effector process», «intracellular protein kinase cascade», «regulation of cellular component biogenesis», «regulation of cell cycle process», «regulation of organelle organization», «regulation of cell cycle» (see Additional file 1).

An associative pathway network connecting PE, via the *PLAT, ADIPOQ, LEPR, TGFB1, TNFA, INS, IGFBP1, ACE* and *ESR1* genes, with DM, OB and GD incorporated 66 genes with 167 connections (see Figure 4). Most of these connections (78) corresponded to the "association" type (shown in black). Sixty-nine of them could be referred to "expression regulation" types and 13 as "co-expression" (shown in red); eight comprised "down regulation", "degradation regulation", and "degradation downregulation" (shown in violet).

The differential network of PE risk factors included seven genes (interleukin-1-beta (IL1B), endothelial (NOS3) NO-synthase, heat shock 70 kDa protein 4 (HSPA4), apolipoprotein J (CLU) and 5,10-methylenetetrahydrofolate reductase (MTHFR).

Thus, whereas all the identified PE risk factors might be treated as potential markers of this disease, the most probable molecular mechanism underlying PE, DM, OB and GD includes the pathway starting from the *TGFB1, TNFA, INS* and *ACE* genes, through the *IL1B, NOS3, HSPA4* (HSP74), *CLU* and *MTHFR* genes, and eventually to PE.

Thus, the probable chains of molecular events on the way to combined PE, in this context, are as follows: TGFB1 or TNFA - IL1B - PE; TNFA or INS - NOS3 - PE; INS - HSPA4 or CLU - PE; ACE - MTHFR - PE.

Discussion

The associative networks analysed in this work (see Figures 1, 2, 3, 4) appeared to be significant for the understanding of the nature of PE, thereby supporting the hypothesis that PE represents a stable complex of clinical manifestations [1,3,4]. The key players in the reconstructed networks are comorbid genes which, on the one hand, contribute to the development of PE and its pathogenically related disorders, and, on the other hand, may play the role of "triggers" in the presence of other pre-eclampsia-promoting factors (genes and proteins). Comorbid genes are characteristic of many multifactorial diseases [20]. Moreover, many comorbid diseases may involve various pathophysiological mechanisms [20], and the construction of associative networks makes it possible to understand their molecular interrelations.

An analysis of reconstructed associative networks, which describe interactions between comorbid proteins/genes associated with different pair combinations of PE with DM, Ob, and GD, demonstrated that comorbid diseases differ in a statistically significant manner from random disease pairs. The differences concern both the number of common genes associated with the diseases and the interactions between such genes. The number of vertices in the comorbid networks, as well as the number of interactions between the vertices, exceeded those of random disease pairs. At the same time, the density of connections in the associative molecular genetic network describing the interactions between proteins/genes associated concurrently with all four diseases also differed significantly from those of the random networks formed by random sets of proteins/genes. Interestingly, we also observed the same regularity for inversely comorbid diseases [10]. It has been shown that the associative networks reconstructed for pairs of inversely comorbid diseases, including bronchial asthma and pulmonary tuberculosis, as well as nine pairs formed by neurodegenerative (Parkinson disease, schizophrenia, Alzheimer disease) and cancer diseases (colorectal neoplasms, prostatic neoplasms, lung neoplasms), significantly differed from the networks that describe interactions between random diseases. An example of the mutual arrangement of inversely comorbid (bronchial asthma and pulmonary tuberculosis) and comorbid diseases is shown in Figure 2.

Our current results are in many respects consistent with those of epidemiological studies worldwide. It has been amply demonstrated that the common risk factors of PE were DM, Ob and GD [1,2,21-26]. In most studies, DM is a leading risk factor, as it occurs in more than half of the women with PE [1,2,24]. Furthermore, DM is more strongly associated with a late-onset of the disease, which prevails among all the cases [24,25]. A study of twin gestations supports our reasoning. In this study, an evaluation of associated factors in PE gestations and a comparison of the incidence of pregnancy complications among twins with and without PE demonstrated that a high pregnancy body mass index (BMI) and diabetes were associated with PE [27].

We identified 16 genes encoding shared proteins in the molecular network, built using the literature- and database-mining (ANDSystem, Pathway Studio and STRING), that simultaneously connected with PE, DM, GD and Ob. Most shared genes determined in this study encode proteins controlling energy metabolism, and are associated with the immune response and inflammation.

An analysis of the associations of these genes with PE and DM, GD and Ob obtained in case-control, family-based, and meta-analyses studies, which we conducted using the HuGE Navigator, revealed that 14 of the 16 shared genes were associated with at least one of the diseases (see Table 1). Two genes (*CSRP1* and *PLAT*) had never been shown to be associated with PE and DM, GD and Ob. Four shared genes (*ACE, ADIPOQ,*

Table 1. Statistics of gene-disease associations for PE, DM, GD and Ob obtained with the HuGE Navigator

Gene name	PE	DM	GD	Ob
ACE	39	244	2	77
ADA	-	6	-	-
ADIPOQ	4	156	4	176
ADRB3	1	49	4	145
CRP	2	20	-	28
CSRP1	-	-	-	-
ESR1	7	21	-	36
HP	2	36	-	5
IGFBP1	-	7	-	5
INS	1	88	4	26
LEPR	7	35	1	154
MBL2	4	14	1	1
PLAT	2	3	-	1
SHBG	-	6	1	7
TGFB1	8	33	-	8
TNFA	24	132	5	83

The number of associations determined by case-control, family-based and meta-analysis studies are shown.

MBL2, TNFA) were found to be associated with all the diseases.

We believe that the identification of these genes in the current study is of importance because they encode proteins important for the development of diseases, as confirmed by experimental studies (Table 1).

Angiotensin-converting enzyme (ACE) plays a key role in regulating blood pressure by influencing vascular tone by activating the vasoconstrictor angiotensin II and inactivating the vasodilatory peptide bradykinin. Inter-individual differences in blood ACE levels are at least in part explained by the presence of an insertion/deletion (I/D) polymorphism in intron 16 of the ACE gene, with higher ACE levels observed in D allele carriers. The results of many studies confirmed the association of ACE polymorphism with PE [28]. Other studies have indicated that the ACE gene is a factor that contributes to the manifestation of GD [29], diabetic nephropathy and Ob [30,31].

It has been shown that polymorphisms in the adiponectin gene (*ADIPOQ*) modulate the circulating concentration of adiponectin. Abnormal adiponectin levels, as well as *ADIPOQ* polymorphisms, have been associated with PE [32]. Some variants of this gene are associated with the occurrence of GD [33], while other polymorphisms may contribute to type 2 DM risk [34] and Ob in adults [35].

Mannose-binding lectin (MBL) is involved in the maintenance of an inflammatory environment in the uterus. High MBL levels have been associated with successful pregnancies, whereas low levels are involved in PE development. Association between polymorphisms in the structural and promoter regions of the *MBL2* gene and PE have been evaluated [36]. MBL gene polymorphisms are associated with GD and with type 2 DM [37,38]; in addition, MBL deficiency may confer a risk of Ob and insulin resistance [39].

Tumour necrosis factor-alpha (TNF-α) participates in the immune response and inflammation. Many studies have showed that there is an association between the *TNFA* gene and PE among Europeans [2,40]. The -308 G−>A polymorphism of the *TNFA* promoter gene is involved in the pathophysiology of insulin resistance and GD [41]. The same polymorphism is a genetic risk factor for the development of type 2 DM [42]. Individuals who carry the -308A *TNFA* gene variant have a 23% greater risk of developing obesity compared with controls, and they showed significantly higher systolic arterial blood pressure and plasma insulin levels, supporting the hypothesis that the *TNFA* gene is involved in the pathogenesis of the metabolic syndrome [43].

The PE associome contains more links than each of the individual networks. The identified, shared genes have been classified according to GO. Such a network was needed for a GO overrepresentation analysis. The presence of processes identified by the GO analysis in the pathogenesis of PE is not surprising. The central hypothesis of our understanding of PE is that it results from ischaemia of the placenta, which in turn releases factors into the maternal circulation that are capable of inducing the clinical manifestations of the disease [2]. Multiple pathogenetic mechanisms have been implicated in this disorder, including an imbalance between angiogenic and anti-angiogenic factors, autoantibodies to the type-1 angiotensin II receptor, platelet and thrombin activation, defective deep placentation, intravascular inflammation, endothelial cell activation and/or dysfunction, and oxidative and endoplasmic reticulum stress that promote the differentiation of trophoblasts from a proliferative to an invasive phenotype, regulate cell homeostasis through their involvement in post-translational modifications and protein folding, and induce the release of proinflammatory cytokines and chemokines. Other mechanisms include hypoxia and trophoblast invasion, which down-regulate the expression of transforming growth factor β3 (TGF-β3) and hypoxia-inducible factors (HIF-1α and HIF-2α) [2,44]. These results indicated the contribution of common, non-specific, pathological processes to the development of PE, DG, GD and Ob.

In addition to the identification of common proteins/genes associated with different pathological processes, another goal of the study was to find unique markers for PE. To do so, we reconstructed potential mechanisms of molecular interactions using the ANDSystem software, a program that allows the identification of the largest number of links (see Figure 4). Although the central network

core of these pathways contained only nine common genes (*PLAT, ADIPOQ, LEPR, TGFB1, TNFA, INS, IGFBP1, ACE, ESR1*), it incorporated 68 genes with 174 connections between them, and differential factor risks of PE were identified: the *IL1B, NOS3, HSPA4, CLU* and *MTHFR* genes. The contributions of many of these genes to the pathogenesis of PE has been confirmed by numerous studies [2,45-50]. Here, we showed for the first time that these genes can be specifically involved in the pathogenesis of PE. However, it is not yet clear why these genes have a greater involvement in PE. The possible trigger mechanisms of combined PE are linked to the processes that are carried out by the products of the identified genes, namely, inflammation (*IL1B*), endothelial dysfunction (*NOS3*), heat shock and stress (*HSPA4*), stabilizing cell membranes at diverse fluid-tissue interfaces and protecting the vascular endothelium from an attack by some factors in plasma, such as active complement complexes (*CLU*), and homocysteine metabolism (*MTHFR*).

In addition, the results are of particular importance in regard to the theory of confounding assumptions as false mechanisms of genetic association when the factor is associated with a confound, but not the phenotype, and a confound, in turn, is associated with the phenotype [51,52]. The identified genes can act as such a confound.

Conclusions

The current results broaden our knowledge of the molecular mechanisms of the interactions between comorbid diseases. This reconstruction of associative molecular genetic networks that describe interactions between PE and comorbid diseases (GD, Ob, and DM) differed significantly from partner networks built for random disease pairs. Networks between PE and comorbid diseases had a larger number of genes and links between them. With this in mind, it is of interest that similar features of associative network structure have been observed for inversely comorbid diseases [10]. It can be suggested that comorbid and inversely comorbid relationships between diseases involve larger sets of closely interrelated genes larger than those for random pairs of diseases. In the future, we intend to perform a scale analysis that connects different disease pairs to detect potential comorbid/inversely comorbid diseases for all the possible disease pairs via which these diseases can interact. Reconstruction and analysis of the PE associome is useful for revealing the genetic factors involved in the pathogenesis of the disease and for identifying its differential risk factors, as well as for modelling the theoretical mechanisms of PE development in pregnant women with underlying diseases, such as DB, Ob or GD.

Methods

We used three systems that allowed the automated reconstruction of networks that describe the interactions between proteins/genes and diseases: STRING [13], Pathway Studio [14] and ANDSystem [15].

The ANDSystem was developed for the automated extraction of facts and knowledge regarding the relationships between proteins, genes, metabolites, microRNAs, cellular components, molecular processes, and their associations with diseases from published scientific texts and databases. To extract knowledge from texts in the ANDSystem, the shallow parsing method was applied. Pathway Studio is a software application developed for the navigation and analysis of biological pathways, gene regulation networks and protein interaction maps. The program uses the natural language processing approach to extract knowledge from the texts of scientific publications. STRING is a database and a web resource that contains information about protein-protein interactions (including physical and functional interactions) that is mainly based on the use of text-mining methods.

The associative networks for the considered disease pairs were graphs whose vertices were diseases and human proteins/genes, while the edges were the associations between diseases and proteins.

The following indices of relation between a pair of associative networks were used: (1) the intersection index, $I_{AB} = |A \cap B|$ equal to the intersection size of protein sets A and B composed of proteins concurrently associated with diseases D_A and D_B; (2) the Jaccard index [53] was calculated as the ratio of I_{AB} to the combination of sets A and B involving at least one of the diseases D_A and D_B, $J_{AB} = \frac{I_{AB}}{|A \cup B|}$; (3) Meet/Min [54] was calculated as $M_{AB} = \frac{I_{AB}}{\min(|A|, |B|)}$, where the denominator denotes the size of the minimum of sets A and B.

The statistical significance of the relation indices for the analysed diseases in the associative networks was determined by comparing these networks with the associative ones formed by pairs of randomly chosen diseases. For such an analysis, we used the ANDSystem because this program allows the comparison of reconstructed networks with random ones generated using the ANDCell knowledge base. All the interactions between proteins, genes, metabolites, diseases and other objects described by the ANDSystem are deposited in the ANDCell knowledge base, which is a module of this system [15]. The total number of diseases described in ANDCell was 4,075; of these, 991 were not found to be associated with any human protein. Such diseases were discarded from the analysis. To build the distribution of the relation indices for random disease pairs, 10,000 random disease pairs were generated (see Additional file 2). The P-value for the analysed disease pairs was calculated as the

proportion of 10,000 random networks with the same or larger CR as in the examined pairs of diseases. The associative networks were reconstructed using the ANDSystem and Pathway Studio programs. STRING was not used for this purpose because it gave no information regarding interactions between protein/gene and diseases. The associative networks for the analysed disease pairs included only interactions of the disease-protein/gene type; the interactions between proteins/genes were discounted. As a result, to analyse the interactions between proteins/genes in the associative networks, additional protein/gene-protein/gene associative molecular genetic networks were built using the ANDSystem, Pathway Studio and STRING. The statistical significance of the connectivity of the associative molecular genetics networks built for the analysed disease pairs was also determined by comparing them with random networks. In such a case, for each analysed associative molecular genetic networks, 1,000 random networks were generated using the ANDSystem (only human proteins/genes were considered).

The statistical significance (p-value) of the difference between the connectivity of the analysed network and that of the random networks was also determined, like in the case of the associative networks, as the proportion of random networks with the same or greater number of links between the vertices compared with the number of links in the analysed network. The random molecular genetic networks were built according to the following rules. Proteins/genes considered as vertices in the random networks were taken from the ANDCell knowledge base. To ensure that the proteins/genes in the random networks were represented at a level of study close to that of the proteins/genes from the analysed networks, we considered only those random proteins/genes whose connectivity rate was the same as connectivity rate of proteins/genes from the analysed networks. The set Q_i was formed for each i-th vertex of the analysed network. Q_i was composed of all the proteins/genes from the ANDCell knowledge base having an interaction number in ANDCell equal to the protein/gene interactions in the knowledge base represented by the i-th vertex. The protein/gene for the i-th vertex of the random network was chosen by chance for the set Q_i . The links between the vertices in the random networks were set according to the interactions described in the ANDCell knowledge base.

The results of the automated extraction of information regarding the interactions between proteins/genes and diseases were tested manually. The recognition correctness of the object names in the text, as well as the presence of their interactions, was tested. The lists of shared and specific proteins were reduced by expert evaluation to retain only those participating in the pathogenesis of both diseases for shared proteins, and in the pathogenesis of either disease for specific proteins, as shown previously [10].

The BINGO tool [55] was used to evaluate the overrepresentation of the biological processes for the considered protein/gene set. The enrichment was evaluated using a hypergeometric test with the Benjamini and Hochberg FDR correction using the whole annotation as a reference set. The human Uniprot-GOA Gene Association file (release 2013_05) was used as the custom annotation file. In addition to the statistical significance of the overrepresentation, the overrepresented GO processes were characterized by the CR of the respective proteins/genes in the associative molecular genetics network built for intersection of the four studied diseases. The CR for the protein group of the examined network involved in the overrepresented GO biological process was calculated as the ratio of the number of the protein pairs connected by the network protein pairs of the given group to the number of all possible pairwise combinations of proteins of this group. As is known, the reconstruction quality of the molecular genetic networks is related frequently to the problem of the completeness of information regarding the interactions between proteins. For this reason, to build the network, we took advantage of three independent programs: ANDSystem, Pathway Studio and STRING, with their parameters set by default.

Additional material

Additional file 1: Excel spreadsheet file containing information regarding the characteristics of overrepresented biological processes.

Additional file 2: Excel spreadsheet file containing information regarding the distribution of the relation indices of the disease-protein-disease associative networks.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

Expert analysis of the pathogenetic contributors (diabetes mellitus, gestational diabetes and obesity) was done by ASG, ESV, VSP, ONA, EVM, MSZ and VSB. The development of methods, programs, calculations and analyses of the structural organization of the molecular genetic networks was done by EST, PSD, OVS, TVI, NAK and VAI. All authors read and approved the final manuscript.

Acknowledgements

The work was supported in part by the Russian Science Foundation grant No. 14-24-00123 (development of methods, programs and reconstruction and analysis of the pre-eclampsia associative networks) and Saint-Petersburg State University grant No. 1.38.79.2012 (expert analysis of the pathogenetic contributors: diabetes mellitus, gestational diabetes and obesity).

Declarations

Publication of this article has been funded by the Russian Science Foundation grant No. 14-24-00123.

This article has been published as part of *BMC Systems Biology* Volume 9 Supplement 2, 2015: Selected articles from the IX International Conference on the Bioinformatics of Genome Regulation and Structure\Systems Biology (BGRS \SB-2014): Systems Biology. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcsystbiol/supplements/9/S2.

Authors' details
[1]Federal State Budget scientific Institution "#34;The Research Institute of Obstetrics, Gynecology and Reproductology named after D.O. Ott", St. Petersburg, Russia. [2]Saint-Petersburg State University, St. Petersburg, Russia. [3]The Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia. [4]Novosibirsk State University, Novosibirsk, Russia.

Published: 15 April 2015

References

1. Bilano VL, Ota E, Ganchimeg T, Mori R, Souza JP: **Risk factors of pre-eclampsia/eclampsia and its adverse outcomes in low- and middle-income countries: a WHO secondary analysis.** *PLoS One* 2014, **9**:e91198.
2. Chaiworapongsa T, Chaemsaithong P, Yeo L, Romero R: **Pre-eclampsia part 1: current understanding of its pathophysiology.** *Nat Rev Nephrol* 2014.
3. Young BC, Levine RJ, Karumanchi SA: **Pathogenesis of pre-eclampsia.** *Annu Rev Pathol* 2010, **5**:173-92.
4. Duckitt K, Harrington D: **Risk factors for pre-eclampsia at antenatal booking: systematic review of controlled studies.** *BMJ* 2005, **330**:565.
5. Bellamy L, Casas J-PP, Hingorani AD, Williams DJ: **Pre-eclampsia and risk of cardiovascular disease and cancer in later life: systematic review and meta-analysis.** *BMJ* 2007, **335**:974.
6. Feinstein AR: **The pre-therapeutic classification of co-morbidity in chronic disease.** *Journal of Chronic Diseases* 1970, **23**:455-468.
7. Pfaundler M, Seht L: **Über Syntropie von Krankheitszuständen.** *Zeitschrift für Kinderheilkunde* 1921, **30**:100-120.
8. Ibáñez K, Boullosa C, Tabarés-Seisdedos R, Baudot A, Valencia A: **Molecular evidence for the inverse comorbidity between central nervous system disorders and cancers detected by transcriptomic meta-analyses.** *PLoS genetics* 2014, **10**:e1004173.
9. Freidin MB, Puzyrev VP: *Syntropic genes of allergic diseases* 2010.
10. Bragina EY, Tiys ES, Freidin MB, Koneva LA, Demenkov PS, Ivanisenko VA, Kolchanov NA, Puzyrev VP: **Insights into pathophysiology of dystropy through the analysis of gene networks: an example of bronchial asthma and tuberculosis.** *Immunogenetics* 2014, **66**:457-65.
11. Jenssen TK, Laegreid A, Komorowski J, Hovig E: **A literature network of human genes for high-throughput analysis of gene expression.** *Nat Genet* 2001, **28**:21-8.
12. Cheung WA, Ouellette BFF, Wasserman WW: **Quantitative biomedical annotation using medical subject heading over-representation profiles (MeSHOPs).** *BMC bioinformatics* 2012, **13**:249.
13. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, Lin J, Minguez P, Bork P, Mering C von, Jensen LJ: **STRING v9.1: protein-protein interaction networks, with increased coverage and integration.** *Nucleic Acids Res* 2013, **41(Database)**:D808-15.
14. Nikitin A, Egorov S, Daraselia N, Mazo I: **Pathway studio--the analysis and navigation of molecular networks.** *Bioinformatics* 2003, **19**:2155-7.
15. Demenkov PS, Ivanisenko TV, Kolchanov NA, Ivanisenko VA: **ANDVisio: a new tool for graphic visualization and analysis of literature mined associative gene networks in the ANDSystem.** *In Silico Biol* 2011, **11**:149-61.
16. Larina IM, Kolchanov NA, Dobrokhotov IV, Ivanisenko VA, Demenkov PS, Tiĭs ES, Valeeva OA, Pastushkova LK, Nikolaev EN: **[Reconstruction of associative protein networks connected with processes of sodium exchange' regulation and sodium deposition in healthy volunteers by urine proteome analysis].** *Fiziol Cheloveka* 2012, **38**:107-15.
17. Momynaliev KT, Kashin SV, Chelysheva VV, Selezneva OV, Demina IA, Serebryakova MV, Alexeev D, Ivanisenko VA, Aman E, Govorun VM: **Functional divergence of Helicobacter pylori related to early gastric cancer.** *Journal of proteome research* 2010, **9**:254-67.
18. Podkolodnaya OA, Yarkova EE, Demenkov PS: **Application of the ANDCell computer system to reconstruction and analysis of associative networks describing potential relationships between myopia and glaucoma.** *Russian Journal of Genetics: Applied Research* 2011, **1**:21-28.
19. Sommer B, Tiys ES, Kormeier B, Hippe K, Janowski SJ, Ivanisenko TV, Bragin AO, Arrigo P, Demenkov PS, Kochetov AV, Ivanisenko VA, Kolchanov NA, Hofestädt R: **Visualization and analysis of a cardio vascular disease- and MUPP1-related biological network combining text mining and data warehouse approaches.** *J Integr Bioinform* 2010, **7**:148.
20. Puzyrev VP, Freidin MB: **Genetic view on the phenomenon of combined diseases in man.** *Acta Naturae* 2009, **1**:52-7.
21. Mahaba HM, Ismail NA, El Damaty SI, Kamel HA: **Pre-eclampsia: epidemiology and outcome of 995 cases.** *J Egypt Public Health Assoc* 2001, **76**:357-68.
22. Wendland EM, Duncan BB, Belizán JM, Vigo A, Schmidt MI: **Gestational diabetes and pre-eclampsia: common antecedents?** *Arq Bras Endocrinol Metabol* 2008, **52**:975-84.
23. Schneider S, Freerksen N, Röhrig S, Hoeft B, Maul H: **Gestational diabetes and pre-eclampsia–similar risk factor profiles?** *Early Hum Dev* 2012, **88**:179-84.
24. Ornaghi S, Tyurmorezova A, Algeri P, Giardini V, Ceruti P, Vertemati E, Vergani P: **Influencing factors for late-onset pre-eclampsia.** *J Matern Fetal Neonatal Med* 2013, **26**:1299-302.
25. Lisonkova S, Joseph KS: **Incidence of pre-eclampsia: risk factors and outcomes associated with early-versus late-onset disease.** *Am J Obstet Gynecol* 2013, **209**:544-e1.
26. Dadelszen P von, Magee LA: **Pre-eclampsia: an update.** *Curr Hypertens Rep* 2014, **16**:454.
27. Lučovnik M, Tul N, Verdenik I, Novak Z, Blickstein I: **Risk factors for pre-eclampsia in twin pregnancies: a population-based matched case-control study.** *J Perinat Med* 2012, **40**:379-82.
28. Buurma AJ, Turner RJ, Driessen JH, Mooyaart AL, Schoones JW, Bruijn JA, Bloemenkamp KW, Dekkers OM, Baelde HJ: **Genetic variants in pre-eclampsia: a meta-analysis.** *Hum Reprod Update* 2013, **19**:289-303.
29. Dostálová Z, Bienertová-Vasků AJ, Vasků A, Gerychová R, Unzeitig V: **[Insertion-deletion polymorphism in the gene for angiotensin-converning enzyme (I/D ACE) in pregnant women with gestational diabetes].** *Ceska Gynekol* 2006, **71**:369-73.
30. Yu Z-YY, Chen L-SS, Zhang L-CC, Zhou T-BB: **Meta-analysis of the relationship between ACE I/D gene polymorphism and end-stage renal disease in patients with diabetic nephropathy.** *Nephrology (Carlton)* 2012, **17**:480-7.
31. Mao S, Huang S: **A meta-analysis of the association between angiotensin-converting enzyme insertion/ deletion gene polymorphism and the risk of overweight/obesity.** *J Renin Angiotensin Aldosterone Syst* 2013.
32. Machado JS, Palei AC, Amaral LM, Bueno AC, Antonini SR, Duarte G, Tanus-Santos JE, Sandrim VC, Cavalli RC: **Polymorphisms of the adiponectin gene in gestational hypertension and pre-eclampsia.** *J Hum Hypertens* 2014, **28**:128-32.
33. Low CF, Mohd Tohit ER, Chong PP, Idris F: **Adiponectin SNP45TG is associated with gestational diabetes mellitus.** *Arch Gynecol Obstet* 2011, **283**:1255-60.
34. Chu H, Wang M, Zhong D, Shi D, Ma L, Tong N, Zhang Z: **AdipoQ polymorphisms are associated with type 2 diabetes mellitus: a meta-analysis study.** *Diabetes Metab Res Rev* 2013, **29**:532-45.
35. Wu J, Liu Z, Meng K, Zhang L: **Association of adiponectin gene (ADIPOQ) rs2241766 polymorphism with obesity in adults: a meta-analysis.** *PloS One* 2014, **9**:e95270.
36. Vianna P, Silva GK Da, Santos BP Dos, Bauer ME, Dalmáz CA, Bandinelli E, Chies JA: **Association between mannose-binding lectin gene polymorphisms and pre-eclampsia in Brazilian women.** *Am J Reprod Immunol (New York, NY: 1989)* 2010, **64**:359-74.
37. Megia A, Gallart L, Fernández-Real J-MM, Vendrell J, Simón I, Gutierrez C, Richart C: **Mannose-binding lectin gene polymorphisms are associated with gestational diabetes mellitus.** *J Clin Endocrinol Metab* 2004, **89**:5081-7.
38. Muller YL, Hanson RL, Bian L, Mack J, Shi X, Pakyz R, Shuldiner AR, Knowler WC, Bogardus C, Baier LJ: **Functional variants in MBL2 are associated with type 2 diabetes and pre-diabetes traits in Pima Indians and the old order Amish.** *Diabetes* 2010, **59**:2080-5.
39. Fernández-Real JM, Straczkowski M, Vendrell J, Soriguer F, Pérez Del Pulgar S, Gallart L, López-Bermejo A, Kowalska I, Manco M, Cardona F, García-Gil MM, Mingrone G, Richart C, Ricart W, Zorzano A: **Protection from inflammatory disease in insulin resistance: the role of mannan-binding lectin.** *Diabetologia* 2006, **49**:2402-11.

40. Harmon QE, Engel SM, Wu MC, Moran TM, Luo J, Stuebe AM, Avery CL, Olshan AF: **Polymorphisms in inflammatory genes are associated with term small for gestational age and pre-eclampsia.** *Am J Reprod Immunol* 2014, **71**:472-84.
41. Chang Y, Niu XM, Qi XM, Zhang HY, Li NJ, Luo Y: **[Study on the association between gestational diabetes mellitus and tumor necrosis factor-alpha gene polymorphism].** *Zhonghua Fu Chan Ke Za Zhi* 2005, **40**:676-8.
42. Sefri H, Benrahma H, Charoute H, Lakbakbi El Yaagoubi F, Rouba H, Lyoussi B, Nourlil J, Abidi O, Barakat A: **TNF A -308G>A polymorphism in Moroccan patients with type 2 diabetes mellitus: a case-control study and meta-analysis.** *Mol Biol Rep* 2014.
43. Sookoian SC, González C, Pirola CJ: **Meta-analysis on the G-308A tumor necrosis factor alpha gene variant and phenotypes associated with the metabolic syndrome.** *Obes Res* 2005, **13**:2122-31.
44. Ehsanipoor RM, Fortson W, Fitzmaurice LE, Liao W-XX, Wing DA, Chen D-BB, Chan K: **Nitric oxide and carbon monoxide production and metabolism in pre-eclampsia.** *Reprod Sci* 2013, **20**:542-8.
45. Lachmeijer AM, Nosti-Escanilla MP, Bastiaans EB, Pals G, Sandkuijl LA, Kostense PJ, Aarnoudse JG, Crusius JB, Peña AS, Dekker GA, Arngrímsson R, Kate LP ten: **Linkage and association studies of IL1B and IL1RN gene polymorphisms in pre-eclampsia.** *Hypertens Pregnancy* 2002, **21**:23-38.
46. Serrano NC, Casas JP, Díaz LA, Páez C, Mesa CM, Cifuentes R, Monterrosa A, Bautista A, Hawe E, Hingorani AD, Vallance P, López-Jaramillo P: **Endothelial NO synthase genotype and risk of pre-eclampsia: a multicenter case-control study.** *Hypertension* 2004, **44**:702-7.
47. Chen M, Yuan Z, Shan K: **Association of apolipoprotein J gene 866C–>T polymorphism with pre-eclampsia and essential hypertension.** *Gynecol Obstet Invest* 2005, **60**:133-8.
48. Fekete A, Vér A, Bögi K, Treszl A, Rigó J: **Is pre-eclampsia associated with higher frequency of HSP70 gene polymorphisms?** *Eur J Obstet Gynecol Reprod* 2006, **126**:197-200.
49. Mütze S, Rudnik-Schöneborn S, Zerres K, Rath W: **Genes and the pre-eclampsia syndrome.** *J Perinat Med* 2008, **36**:38-58.
50. Wang XM, Wu HY, Qiu XJ: **Methylenetetrahydrofolate reductase (MTHFR) gene C677T polymorphism and risk of pre-eclampsia: an updated meta-analysis based on 51 studies.** *Arch Med Res* 2013, **44**:159-68.
51. Vanderweele TJ: **Sensitivity analysis: distributional assumptions and confounding assumptions.** *Biometrics* 2008, **64**:645-9.
52. Vanderweele TJ, Mukherjee B, Chen J: **Sensitivity analysis for interactions under unmeasured confounding.** *Stat Med* 2012, **31**:2552-64.
53. Jaccard P: **The distribution of the flora in the alpine zone.** *New Phytol* 1912, **11**:37-50.
54. Goldberg DS, Roth FP: **Assessing experimentally derived interactions in a small world.** *Proc Natl Acad Sci USA* 2003, **100**:4372-4376.
55. Maere S, Heymans K, Kuiper M: **BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks.** *Bioinformatics* 2005, **21**:3448-9.

doi:10.1186/1752-0509-9-S2-S4
Cite this article as: Glotov *et al.*: Molecular association of pathogenetic contributors to pre-eclampsia (pre-eclampsia associome). *BMC Systems Biology* 2015 **9**(Suppl 2):S4.

Submit your next manuscript to BioMed Central and take full advantage of:

- **Convenient online submission**
- **Thorough peer review**
- **No space constraints or color figure charges**
- **Immediate publication on acceptance**
- **Inclusion in PubMed, CAS, Scopus and Google Scholar**
- **Research which is freely available for redistribution**

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central

eLIFE elifesciences.org

RESEARCH ARTICLE

Single cell transcriptional signatures of the human placenta in term and preterm parturition

Roger Pique-Regi[1,2,3]*, Roberto Romero[1,3,4,5,6]*, Adi L Tarca[2,3,7], Edward D Sendler[1], Yi Xu[2,3], Valeria Garcia-Flores[2,3], Yaozhu Leng[2,3], Francesca Luca[1,2], Sonia S Hassan[2,8], Nardhy Gomez-Lopez[2,3,9]*

[1]Center for Molecular Medicine and Genetics, Wayne State University, Detroit, United States; [2]Department of Obstetrics and Gynecology, Wayne State University School of Medicine, Detroit, United States; [3]Perinatology Research Branch, Division of Obstetrics and Maternal-Fetal Medicine, Division of Intramural Research, *Eunice Kennedy Shriver* National Institute of Child Health and Human Development, National Institutes of Health, U.S. Department of Health and Human Services, Detroit, United States; [4]Department of Obstetrics and Gynecology, University of Michigan, Ann Arbor, United States; [5]Department of Epidemiology and Biostatistics, Michigan State University, East Lansing, United States; [6]Detroit Medical Center, Detroit, United States; [7]Department of Computer Science, College of Engineering, Wayne State University, Detroit, United States; [8]Department of Physiology, Wayne State University School of Medicine, Detroit, United States; [9]Department of Immunology, Microbiology, and Biochemistry, Wayne State University School of Medicine, Detroit, United States

***For correspondence:** rpique@wayne.edu (RP-R); prbchiefstaff@med.wayne.edu (RR); ngomezlo@med.wayne.edu (NG-L)

Competing interests: The authors declare that no competing interests exist.

Funding: See page 13

Received: 19 September 2019
Accepted: 12 December 2019
Published: 12 December 2019

Reviewing editor: Stephen Parker, University of Michigan, United States

This is an open-access article, free of all copyright, and may be freely reproduced, distributed, transmitted, modified, built upon, or otherwise used by anyone for any lawful purpose. The work is made available under the Creative Commons CC0 public domain dedication.

Abstract More than 135 million births occur each year; yet, the molecular underpinnings of human parturition in gestational tissues, and in particular the placenta, are still poorly understood. The placenta is a complex heterogeneous organ including cells of both maternal and fetal origin, and insults that disrupt the maternal-fetal dialogue could result in adverse pregnancy outcomes such as preterm birth. There is limited knowledge of the cell type composition and transcriptional activity of the placenta and its compartments during physiologic and pathologic parturition. To fill this knowledge gap, we used scRNA-seq to profile the placental villous tree, basal plate, and chorioamniotic membranes of women with or without labor at term and those with preterm labor. Significant differences in cell type composition and transcriptional profiles were found among placental compartments and across study groups. For the first time, two cell types were identified: 1) lymphatic endothelial decidual cells in the chorioamniotic membranes, and 2) non-proliferative interstitial cytotrophoblasts in the placental villi. Maternal macrophages from the chorioamniotic membranes displayed the largest differences in gene expression (e.g. *NFKB1*) in both processes of labor; yet, specific gene expression changes were also detected in preterm labor. Importantly, several placental scRNA-seq transcriptional signatures were modulated with advancing gestation in the maternal circulation, and specific immune cell type signatures were increased with labor at term (NK-cell and activated T-cell signatures) and with preterm labor (macrophage, monocyte, and activated T-cell signatures). Herein, we provide a catalogue of cell types and transcriptional profiles in the human placenta, shedding light on the molecular underpinnings and non-invasive prediction of the physiologic and pathologic parturition.

Introduction

Parturition is essential for the reproductive success of viviparous species (**Romero et al., 2006a**); yet, the mechanisms responsible for the onset of labor remain to be elucidated (**Norwitz et al., 1999**; **Norwitz et al., 2015**). Understanding human parturition is essential to tackle the challenge of prematurity, which affects 15 million neonates every year (**Muglia and Katz, 2010**; **Blencowe et al., 2012**; **Romero et al., 2014a**). Bulk transcriptomic studies of the cervix (**Hassan et al., 2006**; **Hassan et al., 2007**; **Hassan et al., 2009**; **Bollopragada et al., 2009**; **Dobyns et al., 2015**), myometrium (**Charpigny et al., 2003**; **Romero et al., 2014b**; **Mittal et al., 2010**; **Mittal et al., 2011**; **Chan et al., 2014**; **Stanfield et al., 2019**), and chorioamniotic membranes (**Haddad et al., 2006**; **Mittal et al., 2009**; **Nhan-Chang et al., 2010**) revealed that labor is a state of physiological inflammation; however, finding specific pathways implicated in preterm labor still remains an elusive goal. A possible explanation is that gestational tissues, and especially the placenta, are heterogeneous composites of multiple cell types, and elucidating perturbations in the maternal-fetal dialogue requires dissection of the transcriptional activity at the cell type level, which is not possible using bulk analyses. Recent microfluidic and droplet-based technological advances have enabled characterization of gene expression at single-cell resolution (scRNA-seq) (**Klein et al., 2015**; **Macosko et al., 2015**). Previous work in humans (**Tsang et al., 2017**; **Pavličev et al., 2017**; **Vento-Tormo et al., 2018**) and mice (**Nelson et al., 2016**) demonstrated that scRNA-seq can capture the multiple cell types that constitute the placenta and identify their maternal or fetal origin. Such studies showed that single-cell technology can be used to infer communication networks across the different cell types at the maternal-fetal interface (**Vento-Tormo et al., 2018**). Further, the single-cell-derived placental signatures were detected in the cell-free RNA present in maternal circulation (**Tsang et al., 2017**), suggesting that non-invasive identification of women with early-onset preeclampsia is feasible. However, these studies included a limited number of samples and did not account for the fact that different pathologies can arise from dysfunction in different placental compartments. In addition, the physiologic and pathologic processes of labor have never been studied at single-cell resolution.

Results and discussion

In this study, a total of 25 scRNA-seq libraries were prepared from three placental compartments: basal plate (BP), placental villous (PV), and chorioamniotic membranes (CAM) (**Figure 1A**). These were collected from nine women in the following study groups: term no labor (TNL), term in labor (TIL), and preterm labor (PTL). scRNA-seq libraries were prepared with the 10X Chromium system and were processed using the 10X Cell Ranger software, resulting in 79,906 cells being captured and profiled across all samples (**Supplementary file 1**). We used Seurat (**Butler et al., 2018**) to normalize expression profiles and identified 19 distinct clusters, which were assigned to cell types based on the expression of previously reported marker genes (**Tsang et al., 2017**; **Pavličev et al., 2017**; **Vento-Tormo et al., 2018**) (see Materials and methods, **Figure 1—figure supplement 1** and **Supplementary file 2–3**). The uniform manifold approximation and projection (UMAP **Becht et al., 2019**) was used to display these clusters in two dimensions (**Figure 1B**). With this approach, the local and global topological structure of the clusters is preserved, with subtypes of the major cell lineages (trophoblast, lymphoid, myeloid, stromal, and endothelial sub-clusters) being displayed proximal to each other. The trophoblast lineage reconstruction displayed in **Figure 1—figure supplement 2** shows the progression from cytotrophoblasts to either extravillous trophoblasts or syncytiotrophoblasts, which recapitulates the differentiation structure previously reported (**Tsang et al., 2017**; **Vento-Tormo et al., 2018**).

The cell type composition differed both among placental compartments (**Figure 1C**) and due to the presence of physiologic and pathologic processes of labor (i.e. term in labor and preterm labor) (**Figure 1D**). While extravillous trophoblasts (EVT) were present in all three compartments, cytotrophoblasts (CTB) were especially pervasive in the placental villi, which is explained by the fact that CTBs are abundant in the parenchyma of the placentas. CTBs were also present in the basal plate since this placental compartment is adjacent to the placental villi (**Figure 1A**). The phenotypic similarities between trophoblasts in proximity to the decidua parietalis (layer attached to the chorioamniotic membranes) and those found in the basal plate have been previously documented

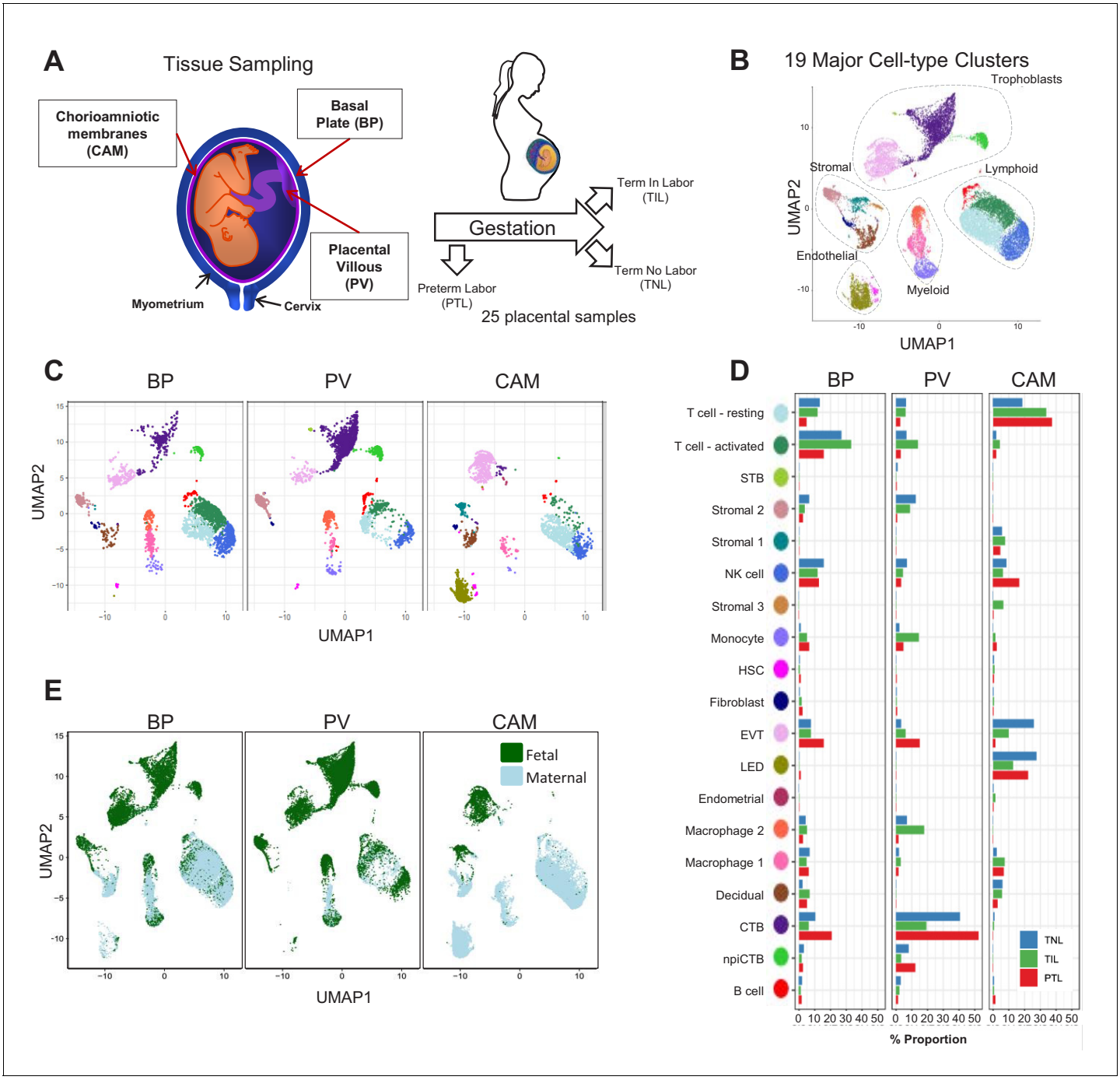


Figure 1. Transcriptional map of the placenta in human parturition. (**A**) Study design illustrating the placental compartments and study groups. (**B**) Uniform Manifold Approximation Plot (UMAP), where dots represent single cells and are colored by cell type. (**C**) Distribution of single-cell clusters by placental compartments. (**D**) Average proportions of cell types by placental compartments and study groups. (**E**) Distribution of single cells by maternal or fetal origin. STB, Syncytiotrophoblast; EVT, Extravillous trophoblast; CTB, cytotrophoblast; HSC, hematopoietic stem cell; npiCTB, non proliferative interstitial cytotrophoblast; LED, lymphoid endothelial decidual cell.

The online version of this article includes the following figure supplement(s) for figure 1:

Figure supplement 1. Heatmap of the top gene expression markers defining each cell-type.

Figure supplement 2. UMAP plot highlighting the trophoblast cell-types and their inferred differentiation path using slingshot R package.

Figure supplement 3. Single marker gene expression UMAP plot for genes differentially expressed between CTB and npiCTB.

Figure supplement 4. Analysis of the fetal/maternal origin of the cell-types based on data from three pregnancies with a male fetus.

Figure supplement 5. Alluvial diagram showing the correspondence between our final curated cluster labels and automated cell-labeling methods.

Figure 1 continued on next page

Figure 1 continued

Figure supplement 6. Heatmap showing the correspondence between our final curated cluster labels and automated cell-labeling methods.
Figure supplement 7. Uniform Manifold Approximation Plot (UMAP), where dots representing single cells and color represents Seurat predicted cell type labels.
Figure supplement 8. Doublet analysis by DoubletFinder.

(***Genbačev et al., 2015***; ***Garrido-Gomez et al., 2017***). Importantly, non-proliferative interstitial cytotrophoblasts (npiCTB) were identified for the first time in the placental villi as forming a distinct cluster. This new cluster was also observed in the basal plate, but not in the chorioamniotic membranes, suggesting that this type of trophoblast has specific functions in the placental tree. Lineage reconstruction by slingshot (***Street et al., 2018***) revealed that npiCTBs are likely derived from conventional CTBs (***Figure 1—figure supplement 2***). The non-proliferative nature of npiCTBs was evidenced by the reduced expression of genes involved in cell proliferation such as *XIST*, *DDX3X*, and *EIF1AX* (***Figure 1—figure supplement 3***). npiCTBs displayed an increased expression of *PAGE4* (***Figure 1—figure supplement 3***), a gene expressed by CTBs isolated from pregnancy terminations (***Genbacev et al., 2011***), suggesting that this type of trophoblast cell is present earlier in gestation. As expected, trophoblast cell types were of fetal origin, and decidual cells present in the basal plate (including the decidua basalis) and chorioamniotic membranes (including the decidua parietalis) were of maternal origin (***Figure 1E*** and ***Figure 1—figure supplement 4***).

In terms of immune cell types, the chorioamniotic membranes largely contained lymphoid and myeloid cells of maternal origin, including T cells (mostly in a resting state), NK cells, and macrophages (***Figure 1C and E*** and ***Figure 1—figure supplement 4***). In contrast, the basal plate included immune cells of both maternal and fetal origin, such as T cells (mostly in an active state), NK cells, and macrophages. The placental villi contained more fetal than maternal immune cells, namely monocytes, macrophages, T cells, and NK cells. Two macrophage subsets were found in placenta compartments: macrophage 1 of maternal origin that was predominant in the chorioamniotic membranes, and macrophage 2 of fetal origin that was mainly present in the basal plate and placental villi. Together with previous single cell studies of early pregnancy (***Vento-Tormo et al., 2018***), these results highlight the complexity and dynamics of the immune cellular composition of the placental tissues, including the maternal-fetal interface (i.e. decidua), from early gestation to term or preterm delivery.

Importantly, a new lymphatic endothelial decidual (LED) cell type of maternal origin was identified in the chorioamniotic membranes, forming a distinct transcriptional cluster that was separate from other endothelial cell-types (***Figure 1C and E***). LED cells were rarely observed in the basal plate and were completely absent in the placental villous. Similar to other endothelial cell types, LED cells highly expressed *CD34*, *CDH5*, *EDNRB*, *PDPN*, and *TIE1* (***Figure 2—figure supplement 1***). The signature genes of this novel cell type were enriched for pathways involving cell-cell and cell-surface interactions at the vascular wall, extracellular matrix organization (***Figure 2—figure supplement 2***), tight junction, and focal adhesion (***Figure 2—figure supplement 3***), indicating that LEDs possess the machinery required to mediate the influx of immune cells into the chorioamniotic membranes. Immunostaining confirmed the co-expression of LYVE1 (lymphatic marker) and CD31 (endothelial molecule marker) in the vessels of the decidua parietalis of the chorioamniotic membranes, but not in the basal plate or placenta (***Figure 2A***). The co-localization of LYVE1 and CD31 proteins (i.e. LED cells) in the chorioamniotic membranes is shown in ***Figure 2B*** and ***Figure 2—video 1***. LED cells also expressed the common endothelial cell marker *CD34* (***Figure 2C***, green arrow). *LYVE1* was also expressed by the fetal macrophages present in the placental villi and basal plate (***Figure 2C***, red arrow), yet the protein was only visualized by immunostaining in immune cells located in the villous tree (***Figure 2A***, red arrows). This finding conclusively shows the presence of lymphatic vessels in the decidua parietalis of the chorioamniotic membranes, providing a major route for maternal lymphocytes (e.g. T cells) infiltrating the maternal-fetal interface (***Arenas-Hernandez et al., 2019***).

For cell types that were present in more than one placental compartment, major differences in gene expression were identified across locations, indicative of further specialization of cells depending on the unique physiological functions of each microenvironment (***Figure 3—figure supplement 1*** and ***Supplementary file 4***). Differences in the transcriptional profiles were particularly large for

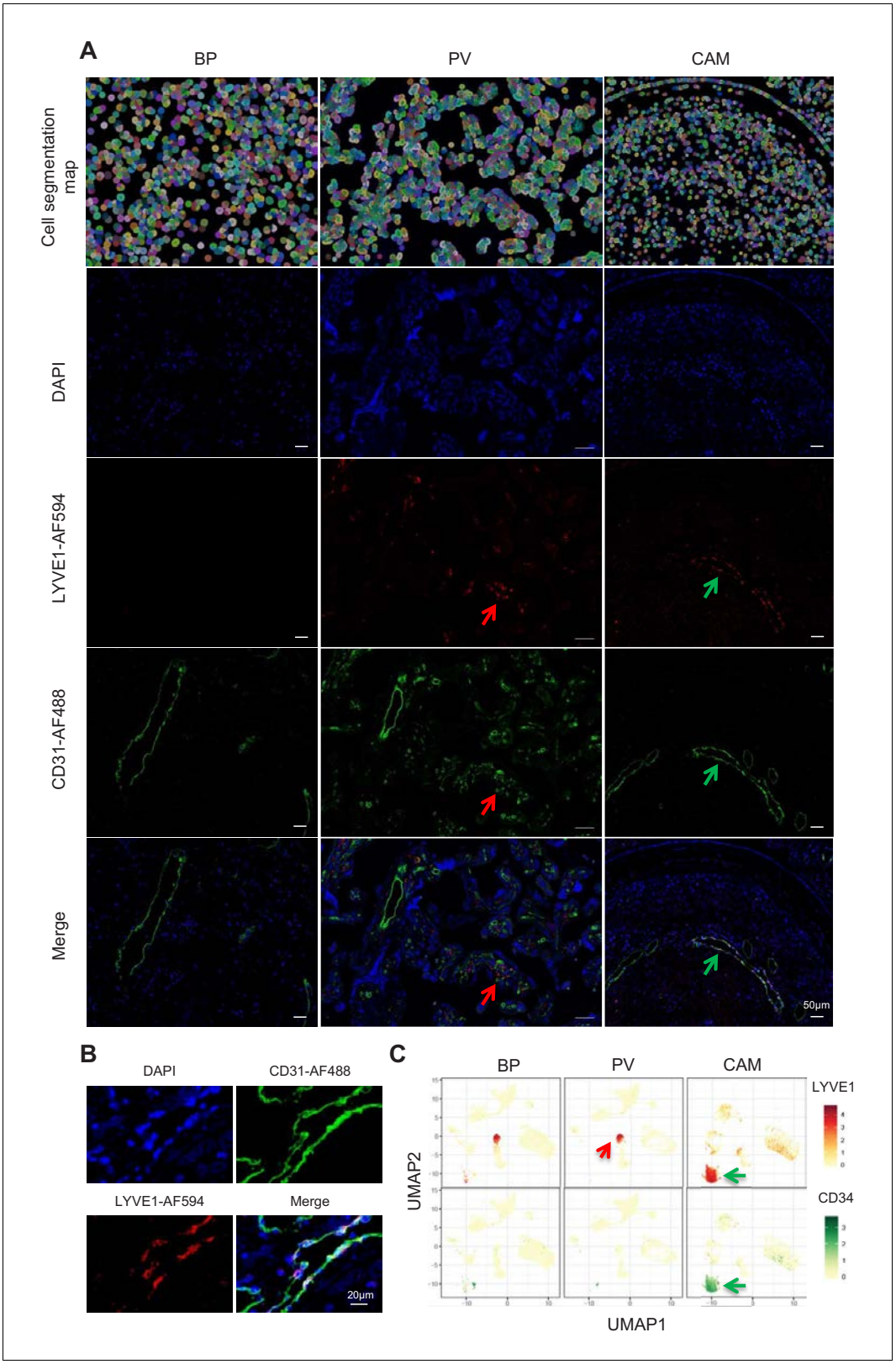


Figure 2. Identification of LED cells in the chorioamniotic membranes. (**A**) Cell segmentation map (built using the DAPI nuclear staining) and immunofluorescence detection of LYVE-1 (red) and CD31 (green) in the basal plate (BP), placental villi (PV), and chorioamniotic membranes (CAM). Red arrows point to fetal macrophages expressing LYVE1 but not CD31, and green arrows indicate lymphatic endothelial decidual cells (LED cells)

Figure 2 continued on next page

Figure 2 continued

expressing both LYVE1 and CD31. (**B**) Co-expression of LYVE1 and CD31 (i.e. LED cells) in the chorioamniotic membranes. (**C**) Single-cell expression UMAP of LYVE-1 (red) and CD34 (green) in the placental compartments.

The online version of this article includes the following video and figure supplement(s) for figure 2:

Figure supplement 1. Single marker gene expression UMAP plot for genes that are more highly expressed in lymphatic endothelial decidual (LED) cells.

Figure supplement 2. Clusterprofiler dot plot showing the ReactomeDB Pathways enriched for genes that define each cell-type.

Figure supplement 3. Clusterprofiler dot plot showing the Kegg Pathways enriched for genes that define each cell-type.

Figure 2—video 1. Video with the 3D reconstruction of the lymphatic endothelium in the decidua present in the CAM compartment.
https://elifesciences.org/articles/52004#fig2video1

maternal macrophages as well as EVTs, NK cells, and T cells in the chorioamniotic membranes compared to the other compartments. Genes differentially expressed in the chorioamniotic membranes were enriched for interleukin and Toll-like receptor signaling as well as for the NF-κB and TNF pathways (***Figure 3—figure supplements 2–4***). These results are consistent with previous reports showing a role for these mediators in the inflammatory process of labor (***Romero et al., 1989a***; ***Romero et al., 1990b***; ***Romero et al., 1992a***; ***Romero et al., 1990a***; ***Santhanam et al., 1991***; ***Romero et al., 1993***; ***Romero et al., 1991***; ***Hsu et al., 1998***; ***Keelan et al., 1999***; ***Young et al., 2002***; ***Osman et al., 2003***; ***Kim et al., 2004***; ***Abrahams et al., 2004***; ***Kumazaki et al., 2004***; ***Koga et al., 2009***; ***Belt et al., 1999***; ***Yan et al., 2002***; ***Lindström and Bennett, 2005***; ***Vora et al., 2010***; ***Romero, 1989b***; ***Romero et al., 1992b***; ***Lonergan et al., 2003***). Conversely, the placental villous and basal plate were more similar to each other, with most differentially expressed genes (DEG) between these compartments being noted in fibroblasts (335 DEG, $q < 0.1$ and fold change >2) (***Figure 3—figure supplements 1*** and ***5–10***). DEGs in the placental villous fibroblasts showed enrichment in smooth muscle contraction, the apelin and oxytocin signaling pathways (***Figure 3—figure supplement 9***), while DEGs in CAM fibroblasts were enriched in elastic fiber formation and extracellular matrix pathways (***Figure 3—figure supplement 2***). The latter finding indicates that the same cell type (e.g. fibroblasts) may have distinct functions in different microenvironments of the placenta.

Next, we assessed changes due to term and preterm labor in each cell type (***Supplementary file 5***). The largest number of DEGs between the term labor and term no labor groups were observed in the maternal macrophages (macrophage 1), followed by the EVT (144 and 37, respectively, $q < 0.1$; ***Figure 3A***). The largest number of DEGs between the preterm labor and term labor groups were observed in EVT and CTB (37 and 33, respectively, $q < 0.1$; ***Figure 3A***). ***Figure 3B*** displays the gene expression changes between TIL and TNL or PTL and TNL that are shared between the two labor groups, representing the common pathway of parturition (defined as the anatomical, physiological, biochemical, endocrinological, immunological, and clinical events that occur in the mother and/or fetus in both term and preterm labor ***Romero et al., 2006b***). Non-shared differences in gene expression with labor at term and in preterm labor were mostly observed in trophoblast cell types such as CTB and EVT as well as in stromal cells (***Figure 3C***). Some of these changes may be explained by the unavoidable confounding effect of gestational age since placentas from women without labor in preterm gestation cannot be obtained in the absence of pregnancy complications. Specifically, the expression of *NFKB1* by maternal macrophages was higher in women with term labor compared to non-labor controls, and this increase was further accentuated in preterm labor (***Figure 3D***). Consistent with the induction of the NFκB pathway, the labor-associated DEGs in macrophages involved biological processes such as activation of immune response and regulation of pro-inflammatory cytokine production (***Figure 3—figure supplement 11A***). These results are in line with previous studies showing that decidual macrophages undergo an M1-like macrophage polarization (i.e. pro-inflammatory phenotype) during term and preterm labor (***Xu et al., 2016***).

When comparing the effect sizes between the PTL/TNL and TIL/TNL juxtapositions on the same gene and cell type, positive correlations were observed for most of the placental cell types (***Figure 3E***). Genes displaying differential effects in term and preterm labor are mostly found in trophoblast cell types (see off-diagonal points in the scatter plot), which may be explained by the phenomenon of gene expression decoherence (***Lea et al., 2019***). This lack of proper correlation between biomarkers to their expected normal relationships is commonly found in pathological

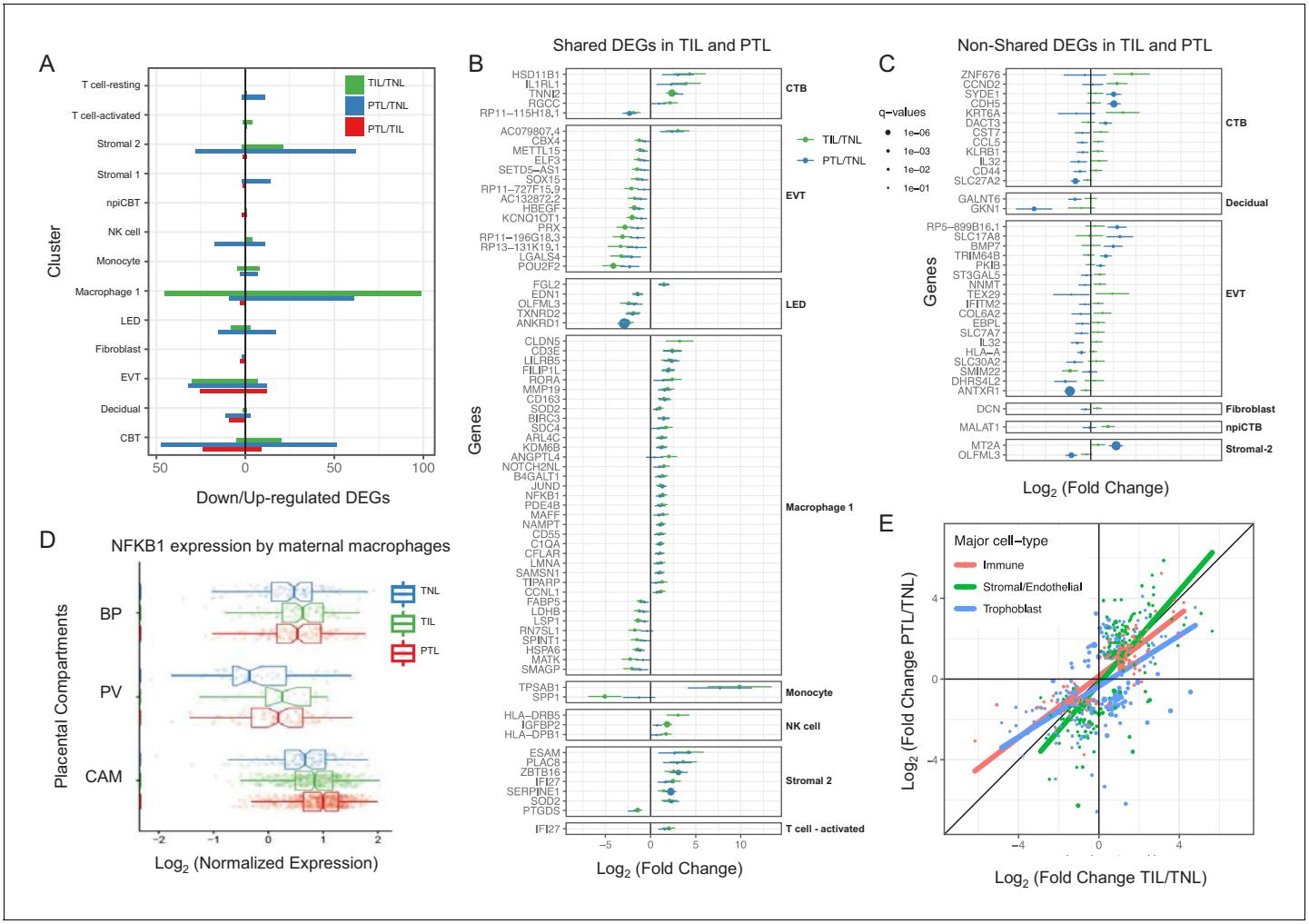


Figure 3. Cell type specific expression changes in term and preterm labor. (**A**) Number of differentially expressed genes (DEGs) among study groups (TNL, term no labor; TIL, term in labor; PTL, preterm labor) by direction of change. Shared (**B**) and non-shared (**C**) expression changes in term labor and preterm labor relative to the term no labor group ($q < 0.01$). The length of each whisker represents the 95% confidence interval. (**D**) The expression of *NFKB1* by maternal macrophages in the placental compartments (BP, basal plate; PV, placental villous; CAM, chorioamniotic membranes) and study groups. The notch represents the 95% confidence interval of the median. (**E**) Differences and similarities in expression changes with preterm labor and term labor by three major cell types (immune, stromal/endothelial, and trophoblast cells).

The online version of this article includes the following figure supplement(s) for figure 3:

Figure supplement 1. Stacked bar plot summarizing differentially expressed genes across compartments for a cell types that are present on all three of them.

Figure supplement 2. Clusterprofiler dot plot showing the ReactomeDB Pathways enriched for genes that are significantly more highly expressed in the CAM compartment relative to the other compartments for each cell-type.

Figure supplement 3. Clusterprofiler dot plot showing the Kegg Pathways enriched for genes that are significantly more highly expressed in the CAM compartment relative to the other compartments for each cell-type.

Figure supplement 4. Clusterprofiler dot plot showing gene ontology (GO) terms enriched for genes that are significantly more highly expressed in the CAM compartment relative to the other compartments for each cell-type.

Figure supplement 5. Clusterprofiler dot plot showing the ReactomeDB Pathways enriched for genes that are significantly more highly expressed in the BP compartment relative to the other compartments for each cell-type.

Figure supplement 6. Clusterprofiler dot plot showing the Kegg Pathways enriched for genes that are significantly more highly expressed in the BP compartment relative to the other compartments for each cell-type.

Figure supplement 7. Clusterprofiler dot plot showing gene ontology (GO) terms enriched for genes that are significantly more highly expressed in the BP compartment relative to the other compartments for each cell-type.

Figure supplement 8. Clusterprofiler dot plot showing the ReactomeDB Pathways enriched for genes that are significantly more highly expressed in the PV compartment relative to the other compartments for each cell-type.

Figure 3 continued on next page

Figure 3 continued

Figure supplement 9. Clusterprofiler dot plot showing the Kegg Pathways enriched for genes that are significantly more highly expressed in the PV compartment relative to the other compartments for each cell-type.

Figure supplement 10. Clusterprofiler dot plot showing gene ontology (GO) terms enriched for genes that are significantly more highly expressed in the PV compartment relative to the other compartments for each cell-type.

Figure supplement 11. Clusterprofiler dot plot showing ReactomeDB pathways enriched using gene set enrichment analysis (GSEA) for genes differentially expressed in term labor compared to term no labor condition.

conditions. Lastly, in EVT the DEGs with labor were enriched for genes implicated in cellular response to stress, including the WNT and NOTCH pathways, as well as cell cycle checkpoints (***Figure 3—figure supplement 11B***), further supporting the hypothesis that the cellular senescence pathway (i.e. cell cycle arrest) is implicated in the physiologic (***Behnia et al., 2015***; ***Polettini et al., 2015***) and pathologic (***Hirota et al., 2010***; ***Gomez-Lopez et al., 2017***) processes of labor.

To demonstrate the translational value of single-cell RNA signatures derived from the placenta, we conducted an in silico analysis in public datasets (***Tarca et al., 2019***; ***Paquette et al., 2018***) to test whether the single-cell signatures could be non-invasively monitored in the maternal circulation throughout gestation (***Figure 4A***). Previous studies have correlated bulk mRNA expression in the maternal circulation with gestational age at blood draw (***Tarca et al., 2019***; ***Al-Garawi et al., 2016***), risk for preterm birth (***Paquette et al., 2018***; ***Heng et al., 2014***; ***Sirota et al., 2018***; ***Knijnenburg et al., 2019***), or both (***Heng et al., 2016***; ***Ngo et al., 2018***). First, using whole blood bulk RNAseq data, we quantified the expression of single-cell signatures in the maternal circulation. We found that the expression of the single-cell signatures of macrophages, monocytes, NK cells, T cells, npiCTB, and fibroblasts is modulated with advancing gestational age (***Figure 4B–C***, ***Figure 4—figure supplement 1A***). These results validate the T-cell and monocyte signature changes with gestational age that were previously reported (***Tsang et al., 2017***; ***Tarca et al., 2019***); yet, here we show that novel placental single-cell signatures (e.g., npiCTB and fibroblast) can also be non-invasively monitored in maternal circulation (***Figure 4—figure supplement 1A***). In addition, for the first time, we report that the expression of the single-cell signatures of NK-cells and activated T-cells were upregulated in women with spontaneous labor at term compared to gestational-age matched controls without labor (***Figure 4D***). Furthermore, we found that the average expression of the single-cell signatures of macrophages, monocytes, activated T cells, and fibroblasts were increased in the circulation of women with preterm labor and delivery compared to gestational age-matched controls (24–34 weeks of gestation) (***Figure 4E*** and ***Figure 4—figure supplement 1B***). These findings are in line with previous reports indicating a role for these immune cell types in the pathophysiology of preterm labor (***Arenas-Hernandez et al., 2019***; ***Hamilton et al., 2012***; ***Shynlova et al., 2013***; ***Gomez-Lopez et al., 2016***).

Conclusion

In summary, this study provides evidence of differences in cell type composition and transcriptional profiles among the basal plate, placental villi, and chorioamniotic membranes, as well as between the pathologic and physiologic processes of labor at single-cell resolution. Using scRNAseq technology, two novel cell types were identified in the chorioamniotic membranes and placental villi. In addition, we showed that maternal macrophages and extravillous trophoblasts are the cell types with the most transcriptional changes during the process of labor. Importantly, many of the genes differentially expressed in these cell-types replicate for both conditions of labor. This result shows that we have enough statistical power to detect the changes in gene expression with a large effect size that are general or a common molecular pathway in parturition; yet, additional studies are needed to characterize the different etiologies of the preterm labor syndrome. Lastly, we report that maternal and fetal transcriptional signatures derived from placental scRNA-seq are modulated with advancing gestation and are markedly perturbed with term and preterm labor in the maternal circulation. These results highlight the potential of single-cell signatures as biomarkers to non-invasively monitor the cellular dynamics during pregnancy and to predict obstetrical disease. The current study represents the most comprehensive single-cell analysis of the human placental transcriptome in physiologic and pathologic parturition.

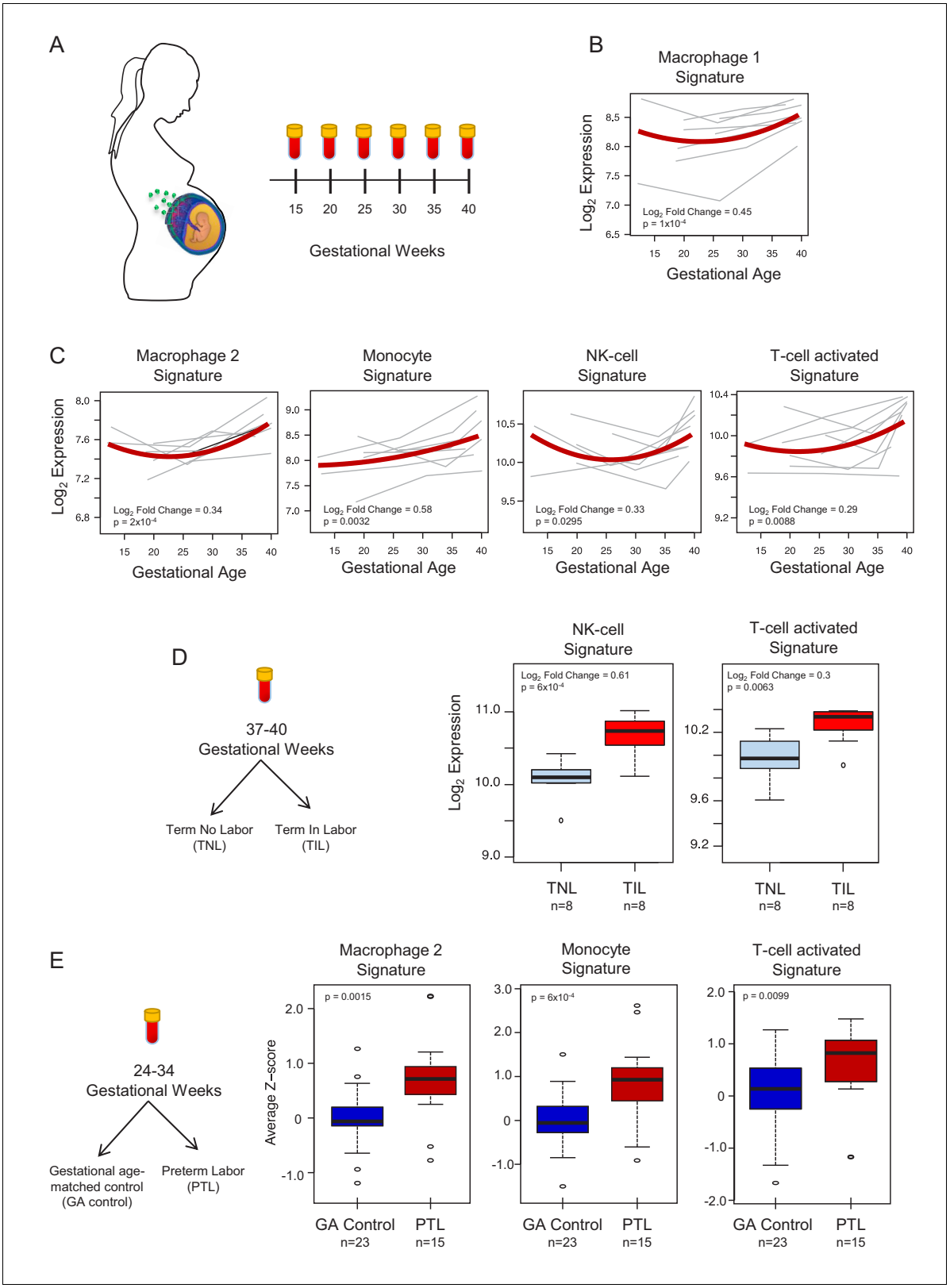


Figure 4. In silico analysis to quantify scRNA-seq signatures in the maternal circulation. (**A**) Diagram of the longitudinal study used to generate bulk RNAseq data (GSE114037) (***Tarca et al., 2019***) to evaluate changes in scRNA-seq signatures with advancing gestation. Whole blood samples were collected throughout gestation from women who delivered at term. (**B and C**) Variation of scRNA-seq signature expression in the maternal circulation with advancing gestation. (**D**) Diagram of the cross-sectional study used to generate bulk RNAseq data (GSE114037) to evaluate changes in scRNA-seq

Figure 4 continued on next page

Figure 4 continued

signatures with labor at term (***Tarca et al., 2019***). Differences in the expression of scRNA-seq signatures between women with spontaneous labor at term (TIL) and term no labor controls (TNL). (**E**) Diagram of the cross-sectional study used to generate bulk RNAseq data (GSE96083) to evaluate changes in scRNA-seq signatures in preterm labor (***Paquette et al., 2018***). Differences in the expression of scRNA-seq signatures between women with spontaneous preterm labor (PTL) and gestational-age matched controls (GA control).
The online version of this article includes the following figure supplement(s) for figure 4:

Figure supplement 1. Quantification of scRNA-seq signatures in maternal circulation (continued from main ***Figure 4***).

Materials and methods

Sample collection and processing, single-cell preparation, library preparation, and sequencing

Human subjects

Immediately after delivery, placental samples [the villi, basal plate (including the decidua basalis) and chorioamniotic membranes (including the decidua parietalis)] were collected from women with or without labor at term or preterm labor at the Detroit Medical Center, Wayne State University School of Medicine (Detroit, MI). Labor was defined by the presence of regular uterine contractions at a frequency of at least two contractions every 10 min with cervical changes resulting in delivery. Women with preterm labor delivered between 33–35 weeks of gestation whereas those with term labor delivered between 38–40 weeks of gestation (***Supplementary file 6***). The collection and use of human materials for research purposes were approved by the Institutional Review Boards of the Wayne State University School of Medicine. All participating women provided written informed consent prior to sample collection.

Single-cell preparation

Cells from placental villi, basal plate, and chorioamniotic membranes were isolated by enzymatic digestion, using previously described protocols with modifications (***Tsang et al., 2017***; ***Xu et al., 2015***). Briefly, placental tissues were homogenized using a gentleMACS Dissociator (Miltenyi Biotec, San Diego, CA) either in an enzyme cocktail from the Umbilical Cord Dissociation Kit (Miltenyi Biotec) or in collagenase A (Sigma Aldrich, St. Louis, MO). After digestion, homogenized tissues were washed with ice-cold 1X phosphate-buffered saline (PBS) and filtered through a cell strainer (Fisher Scientific, Durham, NC). Cell suspensions were then collected and centrifuged at 300 x g for 5 min. at 4˚C. Red blood cells were lysed using a lysing buffer (Life Technologies, Grand Island, NY). Next, cells were washed with ice-cold 1X PBS and resuspended in 1X PBS for cell counting, which was performed using an automatic cell counter (Cellometer Auto 2000; Nexcelom Bioscience, Lawrence, MA). Lastly, dead cells were removed from the cell suspensions using the Dead Cell Removal Kit (Miltenyi Biotec) and cells were counted again using an automatic cell counter.

Single-cell preparation using the 10x genomics platform

Viable cells were used for single-cell RNAseq library construction using the Chromium Controller and Chromium Single Cell 3' version two kit (10x Genomics, Pleasanton, CA), following the manufacturer's instructions. Briefly, viable cell suspensions were loaded into the Chromium Controller to generate gel beads in emulsion (GEM) with each GEM containing a single cell as well as barcoded oligonucleotides. Next, the GEMs were placed in the Veriti 96-well Thermal Cycler (Thermo Fisher Scentific, Wilmington, DE) and reverse transcription was performed in each GEM (GEM-RT). After the reaction, the complementary cDNA was cleaned using Silane DynaBeads (Thermo Fisher Scentific) and the SPRIselect Reagent kit (Beckman Coulter, Indianapolis, IN). Next, the cDNAs were amplified using the Veriti 96-well Thermal Cycler and cleaned using the SPRIselect Reagent kit. Indexed sequencing libraries were then constructed using the Chromium Single Cell 3' version two kit, following the manufacturer's instructions.

Library preparation

cDNA was fragmented, end-repaired, and A-tailed using the Chromium Single Cell 3' version two kit, following the manufacturer's instructions. Next, adaptor ligation was performed using the

Chromium Single Cell 3' version two kit followed by post-ligation cleanup using the SPRIselect Reagent kit to obtain the final library constructs, which were then amplified using PCR. After performing a post-sample index double-sided size selection using the SPRIselect Reagent kit, the quality and quantity of the DNA were analyzed using the Agilent Bioanalyzer High Sensitivity chip (Agilent Technologies, Wilmington, DE). The Kapa DNA Quantification Kit for Illumina platforms (Kapa Biosystems, Wilmington, MA) was used to quantify the DNA libraries, following the manufacturer's instructions.

Sequencing

Sequencing of the single-cell libraries was performed by NovoGene (Sacramento, CA) using the Illumina Platform (HiSeq X Ten System).

Immunofluorescence

Samples of the chorioamniotic membranes, placenta villi, and decidua basal plate were embedded in Tissue-Tek Optimum Cutting Temperature (OCT) compound (Miles, Elkhart, IN) and snap-frozen in liquid nitrogen. Ten-µm-thick sections of each OCT-embedded tissue were cut using the Leica CM1950 (Leica Biosystems, Buffalo Grove, IL). Frozen slides were thawed to room temperature, fixed with 4% paraformaldehyde (Electron Microscopy Sciences, Hatfield, PA), and washed with 1X PBS. Non-specific background signals were blocked using Image-iT FX Signal Enhancer (Life Technologies) followed by blocking with antibody diluent/blocker (Perkin Elmer, Waltham, MA) for 30 min. at room temperature. Slides were then incubated with the rabbit anti-LYVE-1 antibody (Novus Biologicals, Centennial, CO) and the Flex mouse anti-human CD31 antibody (clone JC70A, Dako North America, Carpinteria, CA) for 90 min. at room temperature. Following washing with 1X PBS and blocking with 10% goat serum (SeraCare, Milford, MA), the slides were incubated with secondary goat anti-rabbit IgG–Alexa Fluor 594 (Life Technologies) and goat anti-mouse IgG–Alexa Fluor 488 (Life Technologies) for 30 min. at room temperature. Finally, the slides were washed and coverslips were mounted using ProLong Gold Antifade Mountant with DAPI (Life Technologies). Immunofluorescence was visualized using a confocal fluorescence microscope (Zeiss LSM 780; Carl Zeiss Microscopy GmbH, Jena, Germany) at the Microscopy, Imaging, and Cytometry Resources Core at the Wayne State University School of Medicine. Tile scans were performed from the chorioamniotic membranes, placental villi, and basal plate and the complete imaging fields were divided into six-by-six quadrants.

scRNA-seq data analyses

Raw fastq files obtained from Novogene were processed using Cell Ranger version 2.1.1 from 10X Genomics. First, sequence reads for each library (sample) were aligned to the hg19 reference genome using the STAR (***Dobin et al., 2013***) aligner, and expression of gene transcripts documented in the ENSEMBL database (Build 82) were determined for each cell. Gene expression was determined by the number of unique molecular identifiers (UMI) observed per gene (QC metrics are shown in ***Supplementary file 7***). Second, data were aggregated and down-sampled to take into account differences in sequencing depth across libraries using Cell Ranger Aggregate to obtain gene by cell expression data. Third, Seurat (***Butler et al., 2018***) was used to further clean and normalize the data. Then, only data from cells with a minimum of 200 detected genes, and from genes expressed in at least 10 cells were retained. Cells expressing mitochondrial genes at a level of >10% of total gene counts were also excluded, resulting in 77,906 cells and 25,803 genes (summary in ***Supplementary file 1***). Gene read counts were normalized with the Seurat 'NormalizeData' function (normalization.method = LogNormalize, scale.factor = 10,000). Genes showing significant variation across cells were selected based on 'LogVMR' dispersion function and 'FindVariableGenes'. Ribosomal and mitochondrial genes were next removed, yielding 3147 highly variable genes which were subsequently analyzed using Seurat 'RunPCA' function to obtain the first 20 principal components. Clustering was done using Seurat 'FindClusters' function based on the 20 PCAs (resolution of 0.7). Visualization of the cells was performed using Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) algorithm as implemented by the Seurat 'runUMAP' function and using the first 20 principal components.

Assigning cell type labels to single-cell clusters (Appendix 1)

Multiple methods were utilized to label the cell clusters identified by Seurat. First, marker genes showing distinct expression in individual cell clusters compared to all others were identified using the Seurat FindAllMarkers function with default parameters (***Supplementary file 3***). Marker genes with significant specificity to each cluster (***Figure 1—figure supplement 1*** and ***Supplementary file 3***) were compared to those reported elsewhere (***Tsang et al., 2017***; ***Pavličev et al., 2017***). We also used previous known markers used by our group and https://www.proteinatlas.org/ to manually curate the labels. Further, the xCell (http://xcell.ucsf.edu/#) (***Aran et al., 2017***) tool was utilized to compare the pseudo-bulk expression signatures of the initial clusters to those of known cell types.

Additionally, we compared our manually curated cluster cell type labels to those derived from two automated cell labeling methods: SingleR (***Aran et al., 2019***) and Seurat (***Stuart et al., 2019***), using a human cell atlas reference and the placenta single-cell data in early pregnancy (***Vento-Tormo et al., 2018***) (see Appendix 1 for more details, ***Figure 1—figure supplements 5–7***). Finally, we used the R package DoubletFinder (***McGinnis et al., 2019***) (https://github.com/chris-mcginnis-ucsf/DoubletFinder) to identify potential doublets. None of our clusters were impacted by doublets (***Figure 1—figure supplement 8***).

Identification of cell-type maternal/fetal origin

We used two complementary approaches to determine the maternal/fetal origin of each cell-type. First, we used the samples derived from pregnancies where the neonate was male (3/9 cases, 8/25 samples) and we derived a fetal index based on the sum of all the reads mapping to genes on the Y chromosome relative to the total number of reads mapping to genes on the autosomes (***Figure 1—figure supplement 4***). The second method was based on genotype information derived from the scRNA-seq reads that overlap to known genetic variants from the 1000 Genomes reference panel using the freemuxlet approach implemented in popscle (***Figure 1E***). The freemuxlet approach extends the demuxlet (***Kang et al., 2018***) method, which can be useful for cases in which separate genotype information for each individual is not available. The software available at https://github.com/statgen/popscle/ was used with the '–nsample 2' option to map each cell barcode to one of the two possible genomes: fetal or maternal. The trophoblast cells are of fetal origin; therefore, we used this information to determine the fetal genome.

Trophoblast trajectory analysis

We used the slingshot R package (***Street et al., 2018***) to reconstruct the trophoblast cell lineages from our single-cell gene expression data. This method works by building a minimum spanning tree across clusters of cells and has been reviewed as one of the most accurate tools for this task (***Saelens et al., 2019***). This analysis focused on the trophoblast cell-types (STB, CTB, EVT, and npiCTB), in which we used as input the unmerged cluster labels (i.e., four sub-clusters for CTB, and two for EVT) and the matrix of cell embedding in UMAP (see ***Figure 1—figure supplement 2***).

Differential gene expression

To identify genes differentially expressed among locations (independent of study group), we created a pseudo-bulk aggregate of all the cells of the same cell-type. Only cell types with a minimum of 100 cell in each location were considered in this analysis. Differences in cell type specific expression were estimated using negative binomial models implemented in DESeq2 (***Love et al., 2014***), including a fixed effect for each individual and location. The distribution of p-values for DEGs between pairs of compartments was assessed using a qq-plot to ensure the statistical models were well calibrated (***Supplementary file 3***). To detect DEGs across study groups we aggregated read counts across locations for each cell-type cluster, excluding cell-types with less than 100 cells in each study group (15 clusters). Differences in cell-type specific expression among study groups were estimated using negative binomial models implemented in Deseq2. Differential gene expression was inferred based on FDR adjusted p-value (q-value <0.1) and fold change >2.0.

Gene ontology and pathway enrichment analyses

The clusterProfiler (***Yu et al., 2012***) package in R was utilized for the identification and visualization of enriched pathways among differentially expressed genes identified as described above. The

functions 'enrichGO', 'enrichKEGG', and 'enrichPathway' were used to identify over-represented pathways based on the Gene Ontology (GO), KEGG, and Reactome databases, respectively. Similar enrichment analyses were also conducted using Gene Set Enrichment Analysis (GSEA) (***Subramanian et al., 2005***) which does not require selection of differentially expressed genes as a first step. Significance in all enrichment analyses were based on $q < 0.05$.

In silico quantification of single-cell signatures in maternal whole blood mRNA

Analysis of transcriptional signatures with advancing gestation and with labor at term

Whole-blood samples collected longitudinally (12 to 40 weeks of gestation) from women with a normal pregnancy who delivered at term with (TIL) (n = 8) or without (TNL) (n = 8) spontaneous labor, were profiled using DriverMap and RNA-Seq, as previously described (***Tarca et al., 2019***) and data were available as GSE114037 dataset in the Gene Expression Omnibus. The $\log_2$ normalized read counts were averaged over the top genes (up to 20, ranked by decreasing fold change) distinguishing each cluster from all others as described above (single-cell signature). Whole blood single-cell signature expression in patients with three longitudinal samples was modeled using linear mixed-effects models with quadratic splines in order to assess the significance of changes with gestational age. Differences in single-cell signature expression associated with labor at term (TIL vs. TNL) were assessed using two-tailed equal variance t-tests. In both analyses, adjustment for multiple signature testing was performed using the false discovery rate method, with $q < 0.1$ being considered significant.

Analysis of transcriptional signatures in preterm labor

Whole blood RNAseq gene expression profiles from samples collected at 24–34 weeks of gestation were previously described (***Paquette et al., 2018***) and data were available as GSE96083 dataset in the Gene Expression Omnibus. The study included samples from 15 women with preterm labor who delivered preterm, and 23 gestational age matched controls. Log_2 transformed pseudo read count data were next transformed into Z-scores based on mean and standard deviation estimated in the control group. Single cell signatures were quantified as the average of Z-scores of member genes and compared between groups using a two-tailed Wilcoxon test. Adjustment for multiple signature testing was performed using the false discovery rate method, with $q < 0.1$ being considered a significant result.

Data and materials availability

The scRNA-seq data reported in this study has been submitted to NIH dbGAP repository (accession number phs001886.v1.p1). All other data used in this study are already available through Gene Expression Omnibus (accession identifiers GSE114037 and GSE96083) and through ArrayExpress (E-MTAB-6701). All software and R packages used herein are detailed in the Materials and methods. Scripts detailing the analyses are also available at https://github.com/piquelab/sclabor. To enable further exploration of the results we have also provided a Shiny App in Rstudio available at: http://placenta.grid.wayne.edu/.

Additional information

Funding

Funder	Grant reference number	Author
Eunice Kennedy Shriver National Institute of Child Health and Human Development	HHSN275201300006C	Roberto Romero
Wayne State University	Perinatal Research Initiative	Nardhy Gomez-Lopez Adi L Tarca

The funders had no role in study design, data collection and interpretation.

Author contributions

Roger Pique-Regi, Resources, Data curation, Software, Formal analysis, Supervision, Investigation, Visualization, Methodology, Project administration; Roberto Romero, Conceptualization, Resources, Supervision, Funding acquisition, Project administration; Adi L Tarca, Resources, Software, Formal analysis, Investigation, Visualization, Methodology; Edward D Sendler, Formal analysis, Investigation, Visualization; Yi Xu, Investigation, Methodology, Project administration; Valeria Garcia-Flores, Methodology, Experiments; Yaozhu Leng, Validation, Investigation, Visualization, Methodology; Francesca Luca, Resources, Methodology; Sonia S Hassan, Resources, Funding acquisition, Project administration; Nardhy Gomez-Lopez, Conceptualization, Resources, Data curation, Supervision, Funding acquisition, Validation, Investigation, Visualization, Methodology, Project administration

Author ORCIDs

Roger Pique-Regi https://orcid.org/0000-0002-1262-2275
Roberto Romero http://orcid.org/0000-0002-4448-5121
Adi L Tarca https://orcid.org/0000-0003-1712-7588
Francesca Luca http://orcid.org/0000-0001-8252-9052
Nardhy Gomez-Lopez https://orcid.org/0000-0002-3406-5262

Ethics

Human subjects: The collection and use of human materials for research purposes were approved by the Institutional Review Boards of the Wayne State University School of Medicine 040302M1F. All participating women provided written informed consent prior to sample collection. Data sharing certification (dbGaP phs001886.v1.p1) has been provided (see data availibility section).

Decision letter and Author response

Decision letter https://doi.org/10.7554/eLife.52004.sa1
Author response https://doi.org/10.7554/eLife.52004.sa2

Additional files

Supplementary files

- Supplementary file 1. Summary of the scRNA-seq libraries prepared. Each row summarizes each 10X Genomics scRNA-seq library prepared and processed in this study: sample ID, number of cells detected after filtering, location of the tissue (BP = basal plate, PV = Placental Villi, CAM = chorioamniotic membranes), pregnancy condition (TNL = term no labor, TIL = term in labor, PTL = preterm labor), gender of the neonate, and total number of UMIs detected.
- Supplementary file 2. Summary of cell count by cell-type, location and condition. Each row summarizes the total number of cells of each cell-type as determined by Seurat and split by pregnancy condition (TNL = term no labor, TIL = term in labor, PTL = preterm labor), or location of the tissue (BP = basal plate, PV = Placental Villi, CAM = chorioamniotic membranes).
- Supplementary file 3. Marker Genes identified for each cell-type. The columns represent: 1) Cluster or cell-type name, 2) Ensembl gene identifier, 3) Gene symbol, 4) pct.1: percentage of cells in this cluster where the feature is detected, 5) pct.2: percentage of cells in other clusters where the feature is detected, 6) log fold-change of the average expression between this cluster and the rest, 7) Nominal p-value, 8) Adjusted p-value (Bonferroni).
- Supplementary file 4. Genes differentially expressed across compartments for each common cell-type. The columns represent: 1) Cluster or cell-type name, 2) Comparison groups or contrast (i.e., BP vs PV, BP vs CAM, and CAM vs PV), 3) Ensembl gene identifier, 4) Gene symbol, 5) baseMean gene baseline expression as calculated by DESeq2, 6) log2 Fold Change of the first group in column two versus the second group, 7) Standard error estimated for the log2 Fold Change, 8) Nominal p-value, 9) q-value or adjusted p-value to control for FDR. Only rows with $q < 0.2$ are reported.

- Supplementary file 5. Genes differentially expressed across conditions for each cell-type. The columns represent: 1) Cluster or cell-type name, 2) Comparison groups or contrast (i.e., TNL vs TIL, TIL vs PTL), 3) Ensembl gene identifier, 4) Gene symbol, 5) baseMean gene baseline expression as calculated by DESeq2, 6) log2 Fold Change of the first group in column two versus the second group, 7) Standard error estimated for the log2 Fold Change, 8) Nominal p-value, 9) q-value or adjusted p-value to control for FDR. Only rows with q < 0. two are reported.
- Supplementary file 6. Summary of the sample demographics included in this study. Data are given as medians with interquartile ranges (IQR) or as percentages (n/N). [a]One sample missing data.
- Supplementary file 7. Summary of the QC metrics for the scRNA-seq libraries prepared. Each row represents a library, and each column a QC metric reported by the 10X Cellranger software.
- Transparent reporting form

Data availability

Protected Human subjects data deposited in dbGaP phs001886.v1.p1 Data from other sources detailed in manuscript.

The following dataset was generated:

Author(s)	Year	Dataset title	Dataset URL	Database and Identifier
Pique-Regi R, Romero R, Tarca AL, Sendler ED, Xu Y, Garcia-Flores V, Leng Y, Luca F, HassanSS, Gomez-Lopez N	2019	Single Cell Transcriptional Signatures of the Human Placenta in Term and Preterm Parturition	https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001886.v1.p1	dbGaP, phs001886.v1.p1

The following previously published datasets were used:

Author(s)	Year	Dataset title	Dataset URL	Database and Identifier
Tarca AL, Romero R, Gomez-Lopez N, Hassan SS, Chenchik A	2018	Targeted sequencing based maternal whole blood expression changes with gestational age and labor in normal pregnancy	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE114037	NCBI Gene Expression Omnibus, GSE114037
Paquette AG, Shynlova O, Kibschull M, Price ND, Lye SJ	2017	Genome Scale Analysis of miRNA and mRNA regulation during preterm labor	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE96083	NCBI Gene Expression Omnibus, GSE96083
Vento-Tormo R, Efremova M, Botting RA, Turco MY, Vento-Tormo M, Meyer KB, Park JE, Stephenson E, Polański K, Goncalves A, Gardner L, Holmqvist S, Henriksson J, Zou A, Sharkey AM, Millar B, Innes B, Wood L, Wilbrey-Clark A, Payne RP, Ivarsson MA, Lisgo S, Filby A, Rowitch DH, Bulmer JN, Wright GJ, Stubbington MJT, Haniffa M, Moffett A, Teichmann SA	2018	Reconstructing the human first trimester fetal-maternal interface using single cell transcriptomics - 10x data	https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-6701/	ArrayExpress, E-MTAB-6701

References

Abrahams VM, Bole-Aldo P, Kim YM, Straszewski-Chavez SL, Chaiworapongsa T, Romero R, Mor G. 2004. Divergent trophoblast responses to bacterial products mediated by TLRs. *The Journal of Immunology* **173**: 4286–4296. DOI: https://doi.org/10.4049/jimmunol.173.7.4286, PMID: 15383557

Al-Garawi A, Carey VJ, Chhabra D, Mirzakhani H, Morrow J, Lasky-Su J, Qiu W, Laranjo N, Litonjua AA, Weiss ST. 2016. The role of vitamin D in the transcriptional program of human pregnancy. *PLOS ONE* **11**:e0163832. DOI: https://doi.org/10.1371/journal.pone.0163832, PMID: 27711190

Aran D, Hu Z, Butte AJ. 2017. xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biology* **18**:220. DOI: https://doi.org/10.1186/s13059-017-1349-1, PMID: 29141660

Aran D, Looney AP, Liu L, Wu E, Fong V, Hsu A, Chak S, Naikawadi RP, Wolters PJ, Abate AR, Butte AJ, Bhattacharya M. 2019. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nature Immunology* **20**:163–172. DOI: https://doi.org/10.1038/s41590-018-0276-y, PMID: 30643263

Arenas-Hernandez M, Romero R, Xu Y, Panaitescu B, Garcia-Flores V, Miller D, Ahn H, Done B, Hassan SS, Hsu CD, Tarca AL, Sanchez-Torres C, Gomez-Lopez N. 2019. Effector and activated T cells induce preterm labor and birth that is prevented by treatment with progesterone. *The Journal of Immunology* **202**:2585–2608. DOI: https://doi.org/10.4049/jimmunol.1801350, PMID: 30918041

Becht E, McInnes L, Healy J, Dutertre C-A, Kwok IWH, Ng LG, Ginhoux F, Newell EW. 2019. Dimensionality reduction for visualizing single-cell data using UMAP. *Nature Biotechnology* **37**:38–44. DOI: https://doi.org/10.1038/nbt.4314

Behnia F, Taylor BD, Woodson M, Kacerovsky M, Hawkins H, Fortunato SJ, Saade GR, Menon R. 2015. Chorioamniotic membrane senescence: a signal for parturition? *American Journal of Obstetrics and Gynecology* **213**:359.e1–35359. DOI: https://doi.org/10.1016/j.ajog.2015.05.041, PMID: 26025293

Belt AR, Baldassare JJ, Molnár M, Romero R, Hertelendy F. 1999. The nuclear transcription factor NF-kappaB mediates interleukin-1beta-induced expression of cyclooxygenase-2 in human myometrial cells. *American Journal of Obstetrics and Gynecology* **181**:359–366. DOI: https://doi.org/10.1016/S0002-9378(99)70562-4, PMID: 10454683

Blencowe H, Cousens S, Oestergaard MZ, Chou D, Moller A-B, Narwal R, Adler A, Vera Garcia C, Rohde S, Say L, Lawn JE. 2012. National, regional, and worldwide estimates of preterm birth rates in the year 2010 with time trends since 1990 for selected countries: a systematic analysis and implications. *The Lancet* **379**:2162–2172. DOI: https://doi.org/10.1016/S0140-6736(12)60820-4

Bollopragada S, Youssef R, Jordan F, Greer I, Norman J, Nelson S. 2009. Term labor is associated with a core inflammatory response in human fetal membranes, Myometrium, and cervix. *American Journal of Obstetrics and Gynecology* **200**:104.e1–10104. DOI: https://doi.org/10.1016/j.ajog.2008.08.032

Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. 2018. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology* **36**:411–420. DOI: https://doi.org/10.1038/nbt.4096

Chan YW, van den Berg HA, Moore JD, Quenby S, Blanks AM. 2014. Assessment of myometrial transcriptome changes associated with spontaneous human labour by high-throughput RNA-seq. *Experimental Physiology* **99**: 510–524. DOI: https://doi.org/10.1113/expphysiol.2013.072868, PMID: 24273302

Charpigny G, Leroy MJ, Breuiller-Fouché M, Tanfin Z, Mhaouty-Kodja S, Robin P, Leiber D, Cohen-Tannoudji J, Cabrol D, Barberis C, Germain G. 2003. A functional genomic study to identify differential gene expression in the preterm and term human myometrium. *Biology of Reproduction* **68**:2289–2296. DOI: https://doi.org/10.1095/biolreprod.102.013763, PMID: 12606369

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**:15–21. DOI: https://doi.org/10.1093/bioinformatics/bts635, PMID: 23104886

Dobyns AE, Goyal R, Carpenter LG, Freeman TC, Longo LD, Yellon SM. 2015. Macrophage gene expression associated with remodeling of the prepartum rat cervix: microarray and pathway analyses. *PLOS ONE* **10**: e0119782. DOI: https://doi.org/10.1371/journal.pone.0119782, PMID: 25811906

Garrido-Gomez T, Ona K, Kapidzic M, Gormley M, Simón C, Genbacev O, Fisher SJ. 2017. Severe pre-eclampsia is associated with alterations in cytotrophoblasts of the smooth chorion. *Development* **144**:767–777. DOI: https://doi.org/10.1242/dev.146100, PMID: 28232601

Genbacev O, Donne M, Kapidzic M, Gormley M, Lamb J, Gilmore J, Larocque N, Goldfien G, Zdravkovic T, McMaster MT, Fisher SJ. 2011. Establishment of human trophoblast progenitor cell lines from the chorion. *Stem Cells* **29**:1427–1436. DOI: https://doi.org/10.1002/stem.686, PMID: 21755573

Genbačev O, Vićovac L, Larocque N. 2015. The role of chorionic cytotrophoblasts in the smooth chorion fusion with parietal decidua. *Placenta* **36**:716–722. DOI: https://doi.org/10.1016/j.placenta.2015.05.002, PMID: 26003500

Gomez-Lopez N, Romero R, Arenas-Hernandez M, Ahn H, Panaitescu B, Vadillo-Ortega F, Sanchez-Torres C, Salisbury KS, Hassan SS. 2016. In vivo T-cell activation by a monoclonal αcd3ε antibody induces preterm labor and birth. *American Journal of Reproductive Immunology* **76**:386–390. DOI: https://doi.org/10.1111/aji.12562, PMID: 27658719

Gomez-Lopez N, Romero R, Plazyo O, Schwenkel G, Garcia-Flores V, Unkel R, Xu Y, Leng Y, Hassan SS, Panaitescu B, Cha J, Dey SK. 2017. Preterm labor in the absence of acute histologic chorioamnionitis is

characterized by cellular senescence of the chorioamniotic membranes. *American Journal of Obstetrics and Gynecology* **217**:592.e1–59592. DOI: https://doi.org/10.1016/j.ajog.2017.08.008

Haddad R, Tromp G, Kuivaniemi H, Chaiworapongsa T, Kim YM, Mazor M, Romero R. 2006. Human spontaneous labor without histologic chorioamnionitis is characterized by an acute inflammation gene expression signature. *American Journal of Obstetrics and Gynecology* **195**:394–405. DOI: https://doi.org/10.1016/j.ajog.2005.08.057, PMID: 16890549

Hamilton S, Oomomian Y, Stephen G, Shynlova O, Tower CL, Garrod A, Lye SJ, Jones RL. 2012. Macrophages infiltrate the human and rat decidua during term and preterm labor: evidence that decidual inflammation precedes labor. *Biology of Reproduction* **86**:39. DOI: https://doi.org/10.1095/biolreprod.111.095505, PMID: 22011391

Hassan SS, Romero R, Haddad R, Hendler I, Khalek N, Tromp G, Diamond MP, Sorokin Y, Malone J. 2006. The transcriptome of the uterine cervix before and after spontaneous term parturition. *American Journal of Obstetrics and Gynecology* **195**:778–786. DOI: https://doi.org/10.1016/j.ajog.2006.06.021, PMID: 16949412

Hassan SS, Romero R, Tarca AL, Draghici S, Pineles B, Bugrim A, Khalek N, Camacho N, Mittal P, Yoon BH, Espinoza J, Kim CJ, Sorokin Y, Malone J. 2007. Signature pathways identified from gene expression profiles in the human uterine cervix before and after spontaneous term parturition. *American Journal of Obstetrics and Gynecology* **197**:250.e1–25250. DOI: https://doi.org/10.1016/j.ajog.2007.07.008

Hassan SS, Romero R, Tarca AL, Nhan-Chang CL, Vaisbuch E, Erez O, Mittal P, Kusanovic JP, Mazaki-Tovi S, Yeo L, Draghici S, Kim JS, Uldbjerg N, Kim CJ. 2009. The transcriptome of cervical ripening in human pregnancy before the onset of labor at term: identification of novel molecular functions involved in this process. *The Journal of Maternal-Fetal & Neonatal Medicine* **22**:1183–1193. DOI: https://doi.org/10.3109/14767050903353216, PMID: 19883264

Heng YJ, Pennell CE, Chua HN, Perkins JE, Lye SJ. 2014. Whole blood gene expression profile associated with spontaneous preterm birth in women with threatened preterm labor. *PLOS ONE* **9**:e96901. DOI: https://doi.org/10.1371/journal.pone.0096901, PMID: 24828675

Heng YJ, Pennell CE, McDonald SW, Vinturache AE, Xu J, Lee MW, Briollais L, Lyon AW, Slater DM, Bocking AD, de Koning L, Olson DM, Dolan SM, Tough SC, Lye SJ. 2016. Maternal whole blood gene expression at 18 and 28 weeks of gestation associated with spontaneous preterm birth in asymptomatic women. *PLOS ONE* **11**:e0155191. DOI: https://doi.org/10.1371/journal.pone.0155191, PMID: 27333071

Hirota Y, Daikoku T, Tranguch S, Xie H, Bradshaw HB, Dey SK. 2010. Uterine-specific p53 deficiency confers premature uterine senescence and promotes preterm birth in mice. *Journal of Clinical Investigation* **120**:803–815. DOI: https://doi.org/10.1172/JCI40051, PMID: 20124728

Hsu CD, Meaddough E, Aversa K, Copel JA. 1998. The role of amniotic fluid L-selectin, GRO-alpha, and interleukin-8 in the pathogenesis of intraamniotic infection. *American Journal of Obstetrics and Gynecology* **178**:428–432. DOI: https://doi.org/10.1016/S0002-9378(98)70414-4, PMID: 9539502

Kang HM, Subramaniam M, Targ S, Nguyen M, Maliskova L, McCarthy E, Wan E, Wong S, Byrnes L, Lanata CM, Gate RE, Mostafavi S, Marson A, Zaitlen N, Criswell LA, Ye CJ. 2018. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nature Biotechnology* **36**:89–94. DOI: https://doi.org/10.1038/nbt.4042, PMID: 29227470

Keelan JA, Marvin KW, Sato TA, Coleman M, McCowan LM, Mitchell MD. 1999. Cytokine abundance in placental tissues: evidence of inflammatory activation in gestational membranes with term and preterm parturition. *American Journal of Obstetrics and Gynecology* **181**:1530–1536. DOI: https://doi.org/10.1016/S0002-9378(99)70400-X, PMID: 10601939

Kim YM, Romero R, Chaiworapongsa T, Kim GJ, Kim MR, Kuivaniemi H, Tromp G, Espinoza J, Bujold E, Abrahams VM, Mor G. 2004. Toll-like receptor-2 and -4 in the chorioamniotic membranes in spontaneous labor at term and in preterm parturition that are associated with chorioamnionitis. *American Journal of Obstetrics and Gynecology* **191**:1346–1355. DOI: https://doi.org/10.1016/j.ajog.2004.07.009, PMID: 15507964

Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, Peshkin L, Weitz DA, Kirschner MW. 2015. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**:1187–1201. DOI: https://doi.org/10.1016/j.cell.2015.04.044, PMID: 26000487

Knijnenburg TA, Vockley JG, Chambwe N, Gibbs DL, Humphries C, Huddleston KC, Klein E, Kothiyal P, Tasseff R, Dhankani V, Bodian DL, Wong WSW, Glusman G, Mauldin DE, Miller M, Slagel J, Elasady S, Roach JC, Kramer R, Leinonen K, et al. 2019. Genomic and molecular characterization of preterm birth. *PNAS* **116**:5819–5827. DOI: https://doi.org/10.1073/pnas.1716314116, PMID: 30833390

Koga K, Cardenas I, Aldo P, Abrahams VM, Peng B, Fill S, Romero R, Mor G. 2009. Activation of TLR3 in the trophoblast is associated with preterm delivery. *American Journal of Reproductive Immunology* **61**:196–212. DOI: https://doi.org/10.1111/j.1600-0897.2008.00682.x, PMID: 19239422

Kumazaki K, Nakayama M, Yanagihara I, Suehara N, Wada Y. 2004. Immunohistochemical distribution of Toll-like receptor 4 in term and preterm human placentas from normal and complicated pregnancy including chorioamnionitis. *Human Pathology* **35**:47–54. DOI: https://doi.org/10.1016/j.humpath.2003.08.027, PMID: 14745724

Lea A, Subramaniam M, Ko A, Lehtimäki T, Raitoharju E, Kähönen M, Seppälä I, Mononen N, Raitakari OT, Ala-Korpela M, Pajukanta P, Zaitlen N, Ayroles JF. 2019. Genetic and environmental perturbations lead to regulatory decoherence. *eLife* **8**:e40538. DOI: https://doi.org/10.7554/eLife.40538, PMID: 30834892

Lindström TM, Bennett PR. 2005. The role of nuclear factor kappa B in human labour. *Reproduction* **130**:569–581. DOI: https://doi.org/10.1530/rep.1.00197, PMID: 16264088

Lonergan M, Aponso D, Marvin KW, Helliwell RJ, Sato TA, Mitchell MD, Chaiworapongsa T, Romero R, Keelan JA. 2003. Tumor necrosis factor-related apoptosis-inducing ligand (TRAIL), TRAIL receptors, and the soluble receptor osteoprotegerin in human gestational membranes and amniotic fluid during pregnancy and labor at term and preterm. *The Journal of Clinical Endocrinology & Metabolism* **88**:3835–3844. DOI: https://doi.org/10.1210/jc.2002-021905, PMID: 12915677

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**:550. DOI: https://doi.org/10.1186/s13059-014-0550-8, PMID: 25516281

Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, Trombetta JJ, Weitz DA, Sanes JR, Shalek AK, Regev A, McCarroll SA. 2015. Highly parallel Genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**:1202–1214. DOI: https://doi.org/10.1016/j.cell.2015.05.002, PMID: 26000488

McGinnis CS, Murrow LM, Gartner ZJ. 2019. DoubletFinder: doublet detection in Single-Cell RNA sequencing data using artificial nearest neighbors. *Cell Systems* **8**:329–337. DOI: https://doi.org/10.1016/j.cels.2019.03.003, PMID: 30954475

Mittal P, Romero R, Mazaki-Tovi S, Tromp G, Tarca AL, Kim YM, Chaiworapongsa T, Kusanovic JP, Erez O, Than NG, Hassan SS. 2009. Fetal membranes as an interface between inflammation and metabolism: increased aquaporin 9 expression in the presence of spontaneous labor at term and chorioamnionitis. *The Journal of Maternal-Fetal & Neonatal Medicine* **22**:1167–1175. DOI: https://doi.org/10.3109/14767050903019692, PMID: 19916714

Mittal P, Romero R, Tarca AL, Gonzalez J, Draghici S, Xu Y, Dong Z, Nhan-Chang CL, Chaiworapongsa T, Lye S, Kusanovic JP, Lipovich L, Mazaki-Tovi S, Hassan SS, Mesiano S, Kim CJ. 2010. Characterization of the myometrial transcriptome and biological pathways of spontaneous human labor at term. *Journal of Perinatal Medicine* **38**:617–643. DOI: https://doi.org/10.1515/jpm.2010.097, PMID: 20629487

Mittal P, Romero R, Tarca AL, Draghici S, Nhan-Chang C-L, Chaiworapongsa T, Hotra J, Gomez R, Kusanovic JP, Lee D-C, Kim CJ, Hassan SS. 2011. A molecular signature of an arrest of descent in human parturition. *American Journal of Obstetrics and Gynecology* **204**:177.e15. DOI: https://doi.org/10.1016/j.ajog.2010.09.025

Muglia LJ, Katz M. 2010. The enigma of spontaneous preterm birth. *New England Journal of Medicine* **362**:529–535. DOI: https://doi.org/10.1056/NEJMra0904308, PMID: 20147718

Nelson AC, Mould AW, Bikoff EK, Robertson EJ. 2016. Single-cell RNA-seq reveals cell type-specific transcriptional signatures at the maternal–foetal interface during pregnancy. *Nature Communications* **7**:e11414. DOI: https://doi.org/10.1038/ncomms11414

Ngo TTM, Moufarrej MN, Rasmussen MH, Camunas-Soler J, Pan W, Okamoto J, Neff NF, Liu K, Wong RJ, Downes K, Tibshirani R, Shaw GM, Skotte L, Stevenson DK, Biggio JR, Elovitz MA, Melbye M, Quake SR. 2018. Noninvasive blood tests for fetal development predict gestational age and preterm delivery. *Science* **360**:1133–1136. DOI: https://doi.org/10.1126/science.aar3819, PMID: 29880692

Nhan-Chang C-L, Romero R, Tarca AL, Mittal P, Kusanovic JP, Erez O, Mazaki-Tovi S, Chaiworapongsa T, Hotra J, Than NG, Kim J-S, Hassan SS, Kim CJ. 2010. Characterization of the transcriptome of chorioamniotic membranes at the site of rupture in spontaneous labor at term. *American Journal of Obstetrics and Gynecology* **202**:462.e1–46462. DOI: https://doi.org/10.1016/j.ajog.2010.02.045

Norwitz ER, Robinson JN, Challis JR. 1999. The control of labor. *New England Journal of Medicine* **341**:660–666. DOI: https://doi.org/10.1056/NEJM199908263410906, PMID: 10460818

Norwitz ER, Bonney EA, Snegovskikh VV, Williams MA, Phillippe M, Park JS, Abrahams VM. 2015. Molecular regulation of parturition: the role of the decidual clock. *Cold Spring Harbor Perspectives in Medicine* **5**:a023143. DOI: https://doi.org/10.1101/cshperspect.a023143, PMID: 25918180

Osman I, Young A, Ledingham MA, Thomson AJ, Jordan F, Greer IA, Norman JE. 2003. Leukocyte density and pro-inflammatory cytokine expression in human fetal membranes, decidua, cervix and myometrium before and during labour at term. *Molecular Human Reproduction* **9**:41–45. DOI: https://doi.org/10.1093/molehr/gag001, PMID: 12529419

Paquette AG, Shynlova O, Kibschull M, Price ND, Lye SJ. 2018. Comparative analysis of gene expression in maternal peripheral blood and monocytes during spontaneous preterm labor. *American Journal of Obstetrics and Gynecology* **218**:345.e1–34345. DOI: https://doi.org/10.1016/j.ajog.2017.12.234

Pavličev M, Wagner GP, Chavan AR, Owens K, Maziarz J, Dunn-Fletcher C, Kallapur SG, Muglia L, Jones H. 2017. Single-cell transcriptomics of the human placenta: inferring the cell communication network of the maternal-fetal interface. *Genome Research* **27**:349–361. DOI: https://doi.org/10.1101/gr.207597.116, PMID: 28174237

Polettini J, Behnia F, Taylor BD, Saade GR, Taylor RN, Menon R. 2015. Telomere fragment induced amnion cell senescence: a contributor to parturition? *PLOS ONE* **10**:e0137188. DOI: https://doi.org/10.1371/journal.pone.0137188, PMID: 26397719

Romero R, Brody DT, Oyarzun E, Mazor M, Wu YK, Hobbins JC, Durum SK. 1989a. Infection and labor. III. Interleukin-1: a signal for the onset of parturition. *American Journal of Obstetrics and Gynecology* **160**:1117–1123. DOI: https://doi.org/10.1016/0002-9378(89)90172-5, PMID: 2786341

Romero R. 1989b. Infection and labor. IV. Cachectin-tumor necrosis factor in the amniotic fluid of women with intraamniotic infection and preterm labor. *Am J Obstet Gynecol* **161**:336–341.

Romero R, Avila C, Santhanam U, Sehgal PB. 1990a. Amniotic fluid interleukin 6 in preterm labor. Association with infection. *Journal of Clinical Investigation* **85**:1392–1400. DOI: https://doi.org/10.1172/JCI114583

Romero R, Parvizi ST, Oyarzun E, Mazor M, Wu YK, Avila C, Athanassiadis AP, Mitchell MD. 1990b. Amniotic fluid interleukin-1 in spontaneous labor at term. *The Journal of Reproductive Medicine* **35**:235–238. PMID: 2325034

Romero R, Ceska M, Avila C, Mazor M, Behnke E, Lindley I. 1991. Neutrophil attractant/activating peptide-1 / interleukin-8 in term and preterm parturition. *American Journal of Obstetrics and Gynecology* **165**:813–820. DOI: https://doi.org/10.1016/0002-9378(91)90422-N

Romero R, Mazor M, Brandt F, Sepulveda W, Avila C, Cotton DB, Dinarello CA. 1992a. Interleukin-1 alpha and interleukin-1 beta in preterm and term human parturition. *American Journal of Reproductive Immunology* **27**: 117–123. DOI: https://doi.org/10.1111/j.1600-0897.1992.tb00737.x, PMID: 1418402

Romero R, Mazor M, Sepulveda W, Avila C, Copeland D, Williams J. 1992b. Tumor necrosis factor in preterm and term labor. *American Journal of Obstetrics and Gynecology* **166**:1576–1587. DOI: https://doi.org/10.1016/0002-9378(92)91636-O

Romero R, Yoon BH, Kenney JS, Gomez R, Allison AC, Sehgal PB. 1993. Amniotic fluid interleukin-6 determinations are of diagnostic and prognostic value in preterm labor. *American Journal of Reproductive Immunology* **30**:167–183. DOI: https://doi.org/10.1111/j.1600-0897.1993.tb00618.x, PMID: 8311926

Romero R, Tarca AL, Tromp G. 2006a. Insights into the physiology of childbirth using transcriptomics. *PLOS Medicine* **3**:e276. DOI: https://doi.org/10.1371/journal.pmed.0030276, PMID: 16752954

Romero R, Espinoza J, Kusanovic JP, Gotsch F, Hassan S, Erez O, Chaiworapongsa T, Mazor M. 2006b. The preterm parturition syndrome. *BJOG: An International Journal of Obstetrics & Gynaecology* **113**:17–42. DOI: https://doi.org/10.1111/j.1471-0528.2006.01120.x

Romero R, Dey SK, Fisher SJ. 2014a. Preterm labor: one syndrome, many causes. *Science* **345**:760–765. DOI: https://doi.org/10.1126/science.1251816, PMID: 25124429

Romero R, Tarca AL, Chaemsaithong P, Miranda J, Chaiworapongsa T, Jia H, Hassan SS, Kalita CA, Cai J, Yeo L, Lipovich L. 2014b. Transcriptome interrogation of human myometrium identifies differentially expressed sense-antisense pairs of protein-coding and long non-coding RNA genes in spontaneous labor at term. *The Journal of Maternal-Fetal & Neonatal Medicine* **27**:1397–1408. DOI: https://doi.org/10.3109/14767058.2013.860963, PMID: 24168098

Saelens W, Cannoodt R, Todorov H, Saeys Y. 2019. A comparison of single-cell trajectory inference methods. *Nature Biotechnology* **37**:547–554. DOI: https://doi.org/10.1038/s41587-019-0071-9, PMID: 30936559

Santhanam U, Avila C, Romero R, Viguet H, Ida N, Sakurai S, Sehgal PB. 1991. Cytokines in normal and abnormal parturition: elevated amniotic fluid interleukin-6 levels in women with premature rupture of membranes associated with intrauterine infection. *Cytokine* **3**:155–163. DOI: https://doi.org/10.1016/1043-4666(91)90037-E, PMID: 1888885

Shynlova O, Nedd-Roderique T, Li Y, Dorogin A, Nguyen T, Lye SJ. 2013. Infiltration of myeloid cells into decidua is a critical early event in the labour cascade and post-partum uterine remodelling. *Journal of Cellular and Molecular Medicine* **17**:311–324. DOI: https://doi.org/10.1111/jcmm.12012, PMID: 23379349

Sirota M, Thomas CG, Liu R, Zuhl M, Banerjee P, Wong RJ, Quaintance CC, Leite R, Chubiz J, Anderson R, Chappell J, Kim M, Grobman W, Zhang G, Rokas A, England SK, Parry S, Shaw GM, Simpson JL, Thomson E, et al. 2018. Enabling precision medicine in neonatology, an integrated repository for preterm birth research. *Scientific Data* **5**:e180219. DOI: https://doi.org/10.1038/sdata.2018.219, PMID: 30398470

Stanfield Z, Lai PF, Lei K, Johnson MR, Blanks AM, Romero R, Chance MR, Mesiano S, Koyutürk M. 2019. Myometrial transcriptional signatures of human parturition. *Frontiers in Genetics* **10**:185. DOI: https://doi.org/10.3389/fgene.2019.00185, PMID: 30988671

Street K, Risso D, Fletcher RB, Das D, Ngai J, Yosef N, Purdom E, Dudoit S. 2018. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* **19**:477. DOI: https://doi.org/10.1186/s12864-018-4772-0, PMID: 29914354

Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, Hao Y, Stoeckius M, Smibert P, Satija R. 2019. Comprehensive integration of Single-Cell data. *Cell* **177**:1888–1902. DOI: https://doi.org/10.1016/j.cell.2019.05.031

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *PNAS* **102**:15545–15550. DOI: https://doi.org/10.1073/pnas.0506580102, PMID: 16199517

Tarca AL, Romero R, Xu Z, Gomez-Lopez N, Erez O, Hsu CD, Hassan SS, Carey VJ. 2019. Targeted expression profiling by RNA-Seq improves detection of cellular dynamics during pregnancy and identifies a role for T cells in term parturition. *Scientific Reports* **9**:848. DOI: https://doi.org/10.1038/s41598-018-36649-w, PMID: 30696862

Tsang JCH, Vong JSL, Ji L, Poon LCY, Jiang P, Lui KO, Ni YB, To KF, Cheng YKY, Chiu RWK, Lo YMD. 2017. Integrative single-cell and cell-free plasma RNA transcriptomics elucidates placental cellular dynamics. *PNAS* **114**:E7786–E7795. DOI: https://doi.org/10.1073/pnas.1710470114, PMID: 28830992

Vento-Tormo R, Efremova M, Botting RA, Turco MY, Vento-Tormo M, Meyer KB, Park JE, Stephenson E, Polański K, Goncalves A, Gardner L, Holmqvist S, Henriksson J, Zou A, Sharkey AM, Millar B, Innes B, Wood L, Wilbrey-Clark A, Payne RP, et al. 2018. Single-cell reconstruction of the early maternal-fetal interface in humans. *Nature* **563**:347–353. DOI: https://doi.org/10.1038/s41586-018-0698-6, PMID: 30429548

Vora S, Abbas A, Kim CJ, Summerfield TL, Kusanovic JP, Iams JD, Romero R, Kniss DA, Ackerman WE. 2010. Nuclear factor-kappa B localization and function within intrauterine tissues from term and preterm labor and

cultured fetal membranes. *Reproductive Biology and Endocrinology* **8**:8. DOI: https://doi.org/10.1186/1477-7827-8-8, PMID: 20100341

Xu Y, Plazyo O, Romero R, Hassan SS, Gomez-Lopez N. 2015. Isolation of leukocytes from the human Maternal-fetal interface. *Journal of Visualized Experiments*:e52863. DOI: https://doi.org/10.3791/52863, PMID: 26067211

Xu Y, Romero R, Miller D, Kadam L, Mial TN, Plazyo O, Garcia-Flores V, Hassan SS, Xu Z, Tarca AL, Drewlo S, Gomez-Lopez N. 2016. An M1-like macrophage polarization in decidual tissue during spontaneous preterm labor that is attenuated by rosiglitazone treatment. *The Journal of Immunology* **196**:2476–2491. DOI: https://doi.org/10.4049/jimmunol.1502055

Yan X, Sun M, Gibb W. 2002. Localization of nuclear factor-kappa B (NF kappa B) and inhibitory factor-kappa B (I kappa B) in human fetal membranes and decidua at term and preterm delivery. *Placenta* **23**:288–293. DOI: https://doi.org/10.1053/plac.2002.0789, PMID: 11969339

Young A, Thomson AJ, Ledingham M, Jordan F, Greer IA, Norman JE. 2002. Immunolocalization of proinflammatory cytokines in Myometrium, Cervix, and fetal membranes during human parturition at Term1. *Biology of Reproduction* **66**:445–449. DOI: https://doi.org/10.1095/biolreprod66.2.445

Yu G, Wang L-G, Han Y, He Q-Y. 2012. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS: A Journal of Integrative Biology* **16**:284–287. DOI: https://doi.org/10.1089/omi.2011.0118

Appendix 1

Cell type labeling procedures

Multiple methods and resources were utilized to label the clusters identified by Seurat. First, we used the function FindAllMarkers to identify the genes with significant changes in expression between each cluster and the rest of the cells using min.pct=0.33 and requiring FDR adjusted q < 10% and a $\log$(FC)> 0.5 to determine significance. Clusters with no significant differences at this threshold were merged resulting in a total of 19 clusters. For each cluster, we generated a pseudo-bulk gene expression profile and xCell (http://xcell.ucsf.edu/) (***Aran et al., 2017***) was used to compare the gene expression signatures of our clusters with those of known cell types to the default n = 64 xCell reference which includes immune cells, progenitor, epithelial, and extracellular matrix cells. Eight of the original clusters clearly identified with known cell types in the xCell reference panel that includes T-cell, B-cell, Macrophage, HSC, Fibroblast and Monocyte.

The next method we used is by comparing the marker genes identified by Seurat FindAllMarkers that passed the threshold to previously published scRNAseq marker genes (***Tsang et al., 2017***; ***Pavličev et al., 2017***; ***Vento-Tormo et al., 2018***) and common known markers used by our group and others https://www.proteinatlas.org/search/placenta (***Figure 1—figure supplement 1***). This resolved many of our placental (non-immune) cell clusters in the following cell types: cytotrophoblast, extravillous trophoblast, syncytiotrophoblast, decidual, endothelial, and stromal cells. To further resolve genes differentially expressed between clusters in close proximity to each other (e.g., T-cell subtypes), we ran Seurat FindMarkers function to contrast gene expression between each cluster pair, and determined as differentially expressed genes those showing a minimum logFC> 0.25 and q < 0.1. Using this analysis, we were able to label two subgroups of T cells as activated or resting. Clusters that were distinct but could not be clearly separated into well-known cell sub-types or cellular states were assigned a number (e.g., Stromal-1, Stromal-2). Some of the differences between these clusters are likely due to the maternal/fetal origin of each cell type as shown in ***Figure 1B*** (i.e., Macrophage 1 is likely maternal and Macrophage 2 is likely fetal) as shown by genotype analysis freemuxlet (see Materials and methods). Additionally, we used DoubletFinder https://github.com/chris-mcginnis-ucsf/DoubletFinder (***McGinnis et al., 2019***) to identify doublet cells and to ensure that none of our clusters were confounded by doublets (***Figure 1—figure supplement 8***).

Finally, we also compared our manually curated cell type identification to that derived from automated cell labeling methods SingleR (***Aran et al., 2019***) and Seurat (***Stuart et al., 2019***), (see ***Figure 1—figure supplement 5*** and ***Figure 1—figure supplement 6***). Automated annotation provides a convenient way of transferring biological knowledge across datasets, thereby reducing the burden of interpreting clusters, but it is important to manually curate the cell labels using well established biological knowledge. If the reference database is not specific for the same tissue or similar conditions, this could lead to incorrect assignments. For SingleR, we used the vignette detailed in https://bioconductor.org/packages/devel/bioc/vignettes/SingleR/inst/doc/SingleR.html using the human primary cell atlas (HPCA) reference provided by SingleR and the human placenta first trimester (HPFT) single cell data made available by another group (***Vento-Tormo et al., 2018***) downloaded from https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-6701/. For Seurat, we used only the latter reference and the standard workflow detailed in https://satijalab.org/seurat/v3.1/integration.html, and we removed any cell label with a max score > 0.001 (with almost identical results if the threshold was 0.01 or 0.0001). Similarly, we only used the pruned labels provided by SingleR.

Lymphoid cell types

Four of our clusters correspond to lymphocyte cell-types: B-cell, NK-cell, T-cell activated, and T-cell resting. The cluster labeled as B-cell has an xCell score of 0.88 and express very highly *CD79A*. The automated labeling methods also clearly identify this cluster as B-cell when using the HPCA reference, while it was identified as Plasma cell when using the HPFT reference

panel, as no cell type is labeled as B-cell in HPFT and Plasma cell would make sense as a close match. The cluster labeled NK-cell express very highly *GNLY* and *NKG7* genes, and is also very well matched to NK-cell in the HPCA reference or one of the many NK cell types in HPFT (***Vento-Tormo et al., 2018***), which was a major focus of that study that also enriched for more rare NK cell types as they have a very important role in first trimester pregnancy, but here we only see evidence for one NK-cell cluster in (***Figure 1—figure supplement 7***). Our two clusters labeled as T-cells also closely matched the T-cell types for both reference panels and had xCell scores > 0.5, but only one T cell type is provided by those reference panels. Here, our two clusters differed in some of the genes being expressed that showed that one of the clusters was more active as indicated by signaling factors such as pro-inflammatory cytokine *TNF* and AP-1 factors such as *FOSL* and *JUNB*.

Myeloid cell types

Three of our clusters closely matched myeloid cell types: Macrophage 1, Macrophage 2, and Monocyte. Each of these clusters closely matched to their respective cell types (xCell score > 0.8) and also when using SingleR and Seurat automated label transfer from both reference panels. Macrophage 2, which seemed to be of fetal origin, matched the Hofbauer cell type from the HPFT reference (***Vento-Tormo et al., 2018***), which are fetal resident macrophages found in the human placenta.

Trophoblasts and other cell types

The major trophoblast cell types (CTB, EVT, and STB) expressed the markers that were defined in ***Tsang et al. (2017)***. The newly identified npiCTB also expressed the canonical CTB markers, but had a significantly higher expression of PAGE4 and decreased expression of *DDX3X*, *EIF1AX*, and *XIST* that indicate a non-proliferative state. Using automated cell labeling methods, CTB matched with VCT as defined in HPFT (***Vento-Tormo et al., 2018***), except for a small proportion that matched the SCT profile in HPFT (***Vento-Tormo et al., 2018***) (***Figure 1—figure supplement 7***). This finding may be due to differences in the expression profile of the trophoblast cells types between early and late pregnancy. The SCT in the reference panel (first trimester placental scRNA-seq data) may also include the profile of the transient stage between CTB and STB. This is supported by the trajectory analysis shown in Figure (***Figure 1—figure supplement 2***). Our EVT and STB clusters matched the labels from the automated method using the HPFT reference panel. Other small clusters showing stromal cells matched related cell types described in HPFT (***Vento-Tormo et al., 2018***).

scientific reports

Check for updates

OPEN

Preterm birth buccal cell epigenetic biomarkers to facilitate preventative medicine

Paul Winchester[1], Eric Nilsson[2], Daniel Beck[2] & Michael K. Skinner[2]✉

Preterm birth is the major cause of newborn and infant mortality affecting nearly one in every ten live births. The current study was designed to develop an epigenetic biomarker for susceptibility of preterm birth using buccal cells from the mother, father, and child (triads). An epigenome-wide association study (EWAS) was used to identify differential DNA methylation regions (DMRs) using a comparison of control term birth versus preterm birth triads. Epigenetic DMR associations with preterm birth were identified for both the mother and father that were distinct and suggest potential epigenetic contributions from both parents. The mother (165 DMRs) and female child (136 DMRs) at $p < 1e{-}04$ had the highest number of DMRs and were highly similar suggesting potential epigenetic inheritance of the epimutations. The male child had negligible DMR associations. The DMR associated genes for each group involve previously identified preterm birth associated genes. Observations identify a potential paternal germline contribution for preterm birth and identify the potential epigenetic inheritance of preterm birth susceptibility for the female child later in life. Although expanded clinical trials and preconception trials are required to optimize the potential epigenetic biomarkers, such epigenetic biomarkers may allow preventative medicine strategies to reduce the incidence of preterm birth.

Abbreviations

EWAS	Epigenome-wide association study
DMRs	Differential DNA methylation regions
PTB	Preterm birth
ms-AFP	Maternal serum levels of alpha-fetoprotein
ms-hCG	Human chorionic gonadotropin
FSH	Follicle stimulating hormone
MeDIP	Methylated DNA immunoprecipitation
MeDIP-Seq	Methylated DNA immunoprecipitation followed by next generation sequencing
FT	Full term
FDR	False discovery rate
PCA	Principal component analysis
CTL	Control
Aopep	Aminopeptidase O
DOHAD	Developmental Origins of Health and Disease
PSA	Prostate Specifi Antigen

Preterm birth (PTB) is childbirth that occurs at less than 37 weeks of gestation. Worldwide, preterm birth rates are estimated at 11%, accounting for about 14.8 million of the live births of 2014[1]. Complications of being born preterm were the leading cause of mortality in children less than five years of age in 2015[2]. Children who survive preterm birth are at increased risk of developing future adverse health outcomes, including cognitive disabilities, seizures, visual and hearing impairment, and cardiovascular problems[3–11]. Although there are many risk factors associated with preterm birth including genetic variants, exposure to environmental toxicants, presence of multiple fetuses, preeclampsia and ethnicity, more than half of premature birth cases have an unknown etiology[9,10,12,13]. Reliable biomarkers for preterm birth could greatly help in predicting which pregnancies are at risk and would improve clinical management and health outcomes for the children.

[1]Department of Pediatrics, St. Franciscan Hospital, School of Medicine, Indiana University, Indianapolis, IN 46202-5201, USA. [2]Center for Reproductive Biology, School of Biological Sciences, Washington State University, Pullman, WA 99164-4236, USA. ✉email: skinner@wsu.edu

A number of potential biomarkers for preterm birth have been identifi d. Maternal serum levels of alpha-fetoprotein (ms-AFP) and human chorionic gonadotropin (ms-hCG) have been used clinically[6–8]. Although many associations between mid-trimester ms-hCG and/or ms-AFP levels and adverse pregnancy outcomes are statistically significant, the sensitivity and positive predictive value are too low for them to be clinically useful as screening tests for preterm birth[3,14]. Other proposed biomarkers of PTB risk include selected inflammatory cytokines[15–18], metabolic lipid products[17,19], specific gene mRNA transcripts[20,21], cervicovaginal proteome[22], and micro-RNA transcripts[20,23,24]. Urinary oxidative stress metabolites have also been proposed as biomarkers of preterm birth[25,26]. These biomarkers are not extensively used and are not considered efficient or ideal[27]. Either the assays for proteins and metabolites are technically challenging and expensive, or the specific ty and sensitivity of the assays in predicting preterm birth need to be improved[27].

Previous studies have proposed that epigenetic alterations should be considered for use as biomarkers to predict preterm birth[28–31]. Epigenetics is defi ed as "molecular factors and processes around DNA that regulate genome activity, independent of DNA sequence, and that are mitotically stable"[32]. Epigenetic factors and processes include DNA methylation, histone modifications, non-coding RNA, and chromatin structure changes[33]. Assays for DNA methylation have the advantage of using smaller sample size due to high sensitivity of the assays, as well as being less expensive and technically demanding than assays for proteins and metabolic products. DNA methylation changes can also be detected in easily obtained surrogate samples (i.e., marker cells not directly associated with the etiology of the pathology), such as cheek buccal epithelial cells[34]. This is due to the fact that epigenetic differences can be heritable, so all somatic cells derived from the embryo of an individual have cell-specific epigenetic changes derived from the germline[33]. Altered DNA methylation sites caused by fetal toxicant exposure, abnormal nutrition, or stress have been found in previous studies to be associated with increased risk of disease in exposed offspring and their descendants (i.e., epigenetic transgenerational inheritance)[35–37].

There is evidence that epigenetic differences are associated with preterm birth in the placenta[38] and tissues of children born preterm. Studies that compared DNA methylation in umbilical cord blood between preterm and full-term children found from 31 to 296 differentially methylated sites[38–40]. One study found DNA methylation differences in umbilical cord tissue between preterm and full-term children[39]. These results indicate that DNA methylation changes may occur with preterm birth and suggest that DNA methylation changes are worth investigating as a viable biomarker for predicting preterm birth. Although all cell types have the same DNA sequence present, a limitation of examining DNA methylation changes in a mixed cell population, such as blood with over 20 different cell types, is that each cell type has a unique epigenome and DNA methylation profile driving the cell type specific ty[32]. Thus, small changes in the relative numbers of different cell types in a mixed population can suggest an epigenetic difference, but are in fact due to the changes in cell population numbers[32,33]. Therefore, purifi d individual cell types are preferred to effectively assess epigenetic differences and potential disease biomarkers[41,42].

Changes in DNA methylation at particular genomic loci have been reported as biomarkers associated with human diseases. Sperm samples from men with idiopathic infertility (i.e. infertility from no known cause, and not related to low sperm count or motility) were found to have 217 differential DNA methylation regions (DMRs) at a p value of $p < 1e{-}05$ compared to sperm samples from fertile men[43]. In addition, 56 DMRs were found between initially infertile men who responded to follicle stimulating hormone (FSH) therapy versus those who did not, suggesting that DNA methylation may be used as a biomarker of responsiveness to this therapy[43]. Recently it was reported that a set of 805 DMRs in sperm was potentially associated with men having an increased risk of having a child with autism[44]. Previously, it has been shown that DNA methylation at the SLC9B1 gene in blood samples from pregnant women between 24 and 32 weeks gestation can predict whether the fetus is at risk for fetal intolerance of labor, which can cause fetal hypoxia, and is an indication for performing a Caesarean section[45]. In a recent study, we have used buccal cells as an easily obtained purified cell population to identify epigenetic (i.e., DNA methylation) biomarkers for female rheumatoid arthritis[46]. Although sperm epigenetic biomarkers refl ct epigenetic inheritance of disease in off pring and subsequent generations, a surrogate cell such as buccal cells can reflect early embryo impacts on all somatic cells to be used for disease assessment[46,47]. Together, these studies indicate that epigenetic biomarkers of preterm birth susceptibility or pathology potentially exist and are worthy of further development. Identification of maternal biomarkers associated with preterm birth could help in the prediction and clinical management of at-risk pregnancies and allow for better preventative care for preterm birth children. Clinical management protocols that could be used to reduce the incidence of preterm birth and infant morbidity include: enhanced surveillance of at-risk pregnancies, timely use of prenatal steroids and tocolytics, application of protective uterine monitoring, hospitalization and operative delivery. Epigenetics may also point the way to specific gene targets for future pharmaceutical agents where epigenetically identified "at risk" women could be given gene-specific therapeutics.

The current study was designed to develop epigenetic biomarkers for preterm birth that could be used in a clinical setting to predict preterm birth susceptibility. Buccal cells were obtained from the mother, father, and child from control (> 37 week gestation) and premature (< 37 week gestation) populations and compared. The goal was to find in maternal and paternal buccal cells DMRs which could distinguish preterm from term birth. Clearly the infant epigenetic biomarker is not used to predict potential preterm birth, but can potentially be used to assess later life disease susceptibility in the individual. These epigenetic biomarkers identifi d can now be prospectively tested for their positive and negative predictive power in subsequent investigations. The generational study presented suggests potential epigenetic inheritance aspects for preterm birth.

Results

The objective of the study was to develop an epigenetic (i.e., DNA methylation) biomarker for preterm birth (PTB). One of the least invasive and easiest purifi d cell types to collect is a buccal swab from the cheek, which is greater than 90% pure squamous epithelial cells[48]. Any contaminating bacterial molecular data can be removed during the analysis. Buccal cells were obtained from participants with a home collection swab kit and sent directly to the lab for storage and analysis. The participants were recruited prior to collection or analysis from Indiana University (IU) Health Hospitals (Riley Hospital for Children, IUH Methodist, IUH North) and Franciscan Health, Indianapolis, Indiana. Approvals to conduct the study were obtained from Indiana University Institutional Review Board (IRB) #1901985132 and the Franciscan Institutional Review Board (IRB), #1489434-5. Informed consent and HIPAA authorization was obtained from all participants and from a parent and/or guardian for participants that were minors prior to the clinical sample collection. The buccal cells were collected from the mother, father, and newborn child (triads) to assess epigenetic biomarkers in each group separately. The triad samples were collected, approximately nine days following delivery. Th s period was used to allow the PTB case child to mature and allow an effective buccal cell collection. The full term (FT) birth controls had 21 triad participants and the pre-term birth (PTB) cases had 19 triad participants. Although the majority were of non-Hispanic white Caucasian backgrounds, a number of triads in each population were of African American descent, Supplemental Table S1. The presence of the African American participants did not appear to affect the analysis and similar methylation data was observed in these samples, as assessed with a principal component analysis (PCA), Supplemental Figure S1. The samples were collected in 2019 and early 2020, Supplemental Table S1. The mean maternal age was 28.1 years (controls) and 28.7 years (PTB cases) and mean paternal age 30.8 years (controls) and 30.4 years (PTB cases) with no statistical difference between the control or PTB case groups, with no statistical difference between the groups, Supplemental Table S1. The newborn gestational age at birth, mean ± SD was 38.8 ± 0.94 weeks for the control group and 30.2 ± 3.24 weeks for the PTB case group, with statistical difference ($p \leq 0.001$), Supplemental Table S1. The Supplemental Table S1B presents the clinical demographics for the populations. The preterm pregnancies were found to be signifi antly more likely to be multiparous and less likely to be primiparous. Therefore, PTB occurrences were more likely to have had one or more of the following clinical conditions: (1) to have had a previous preterm birth or pregnancy loss; (2) more likely to have preeclampsia; (3) to have a medically indicated delivery; and/or (4) to have a delivery accompanied by fetal distress and lower APGAR scores. Preterm infants naturally would have had lower birth weights, shorter gestation, and longer hospital stay. Other maternal characteristics were not signifi antly different between groups (i.e., maternal age, paternal age, BMI, insurance source, substance use, diabetes, thyroid placental disorders, cervical disorders, infections, neuropsychiatric disorders), Supplemental Table S1B. Since there were no major outliers in the PCA analysis, the various clinical parameters within the PTB group appear not to be variables for the DMRs, but expanded studies are required to thoroughly assess, Supplemental Figure S1. Buccal cells were collected from each group as outlined in the Methods. All samples were stored at − 80 °C until DNA preparation and analysis.

DNA was isolated from the buccal cell collections and analyzed with a methylated DNA immunoprecipitation (MeDIP) procedure to obtain methylated DNA for subsequent sequencing (Seq) for an MeDIP-Seq protocol[49], as described in the Methods. Th s procedure can provide a genome-wide assessment of greater than 90% of the genome, compared to approximately 50–70% for bisulfite sequencing or less than 1% for array analysis[50]. Differential DNA methylation regions (DMRs) were identifi d by comparing the control and PTB case samples for each mother, father, or child triad. DMRs identifi d were obtained for each group and presented in Fig. 1a for the mother, Fig. 1b for the father, Fig. 1c for the female child, and Fig. 1d for the male child. The DMRs at various edgeR p-value statistical thresholds are presented, and $p < 1e{-}04$ was used for all subsequent data analysis, which was selected as it also provided a reasonable false discovery rate (FDR). The number of adjacent DMR 1 kb windows are shown at a signifi ance level of $p < 1e{-}04$ and the majority of DMR for each group had a single 1 kb window with some higher numbers of significant adjacent windows, Fig. 1a–d. Maternal buccal cells had 165 DMRs, paternal 73 DMRs, female child 136 DMRs, and male child 61 DMRs. The FDR p-value was less than 0.1 for 100% of the mother DMRs, 75% for the father DMRs, 50% for the female child, and 25% (i.e., 14 DMRs) for the male child. Therefore, the male child had less signifi ant DMRs, Fig. 1d. Approximately 50% of DMRs showed an increase and 50% a decrease in DNA methylation in each group, Fig. 1e and f and Supplemental Figure S2. An overlap of the DMRs demonstrated each group was primarily distinct at $p < 1e{-}04$, except for the mother and female child, which shared 31 DMRs in common, Fig. 2a. Further analysis of potential overlaps used an extended overlap analysis with a comparison of the $p < 1e{-}04$ DMRs with the other groups at a $p < 0.05$ threshold. Th s extended overlap demonstrated much higher levels of overlaps with maternal DMRs having a 49% overlap with the paternal, 58% with the female child, and 30% with the male child. Paternal DMRs had a 75% overlap with the mother, 64% with the female child, and 47% with the male child. The female child overlaps were higher and ranged from 34 to 58%, while the male child overlap ranged from 18 to 28%, Fig. 2b. Therefore, preterm birth DMR were identifi d in the buccal cells of the mother and father, as well as in the female children following a preterm birth.

The lists of DMRs and genomic information are presented in Supplemental Table S2 for the mother, Supplemental Table S3 for the father, Supplemental Table S4 for the female child, and Supplemental Table S5 for the male child. These tables present for each group the DMR name, chromosomal location, start and stop nucleotide number, statistics information (p value and FDR), log-fold methylation change (increase positive or decrease negative) for each DMR, gene associations (within 10 kb of gene) and functional categories for the associated genes. The chromosomal locations of the DMRs (red arrowheads) for each group are presented in Fig. 3. The DMRs are present on most chromosomes throughout the genome. The black boxes indicate clusters of DMRs at similar regions. Although some individual DMR overlaps at a 1 kb level are observed, Fig. 2, no obvious gross (Mb size) chromosomal regions or sites are in common between the mother, father or female child genomes, Fig. 3.

DMR Identification

a Mother DMRs

p-value	All Window	Multiple Window
0.001	601	53
1e-04	**165**	**28**
1e-05	56	17
1e-06	32	12
1e-07	20	9

Significant windows (1 kb)	1	2	3	4	≥ 5
Number of DMR (p<1e-04)	137	18	5	2	3

b Father DMRs

p-value	All Window	Multiple Window
0.001	312	30
1e-04	**73**	**10**
1e-05	25	5
1e-06	12	4
1e-07	8	2

Significant windows (1 kb)	1	2	3
Number of DMR (p<1e-04)	63	9	1

c Female Child DMRs

p-value	All Window	Multiple Window
0.001	629	50
1e-04	**136**	**18**
1e-05	32	5
1e-06	10	2
1e-07	3	0

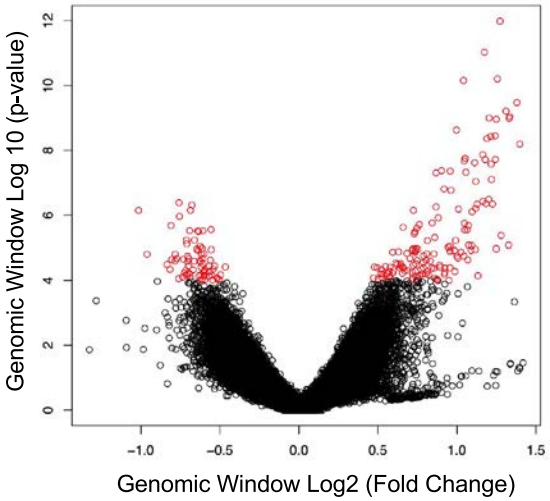
Significant windows (1 kb)	1	2	3
Number of DMR (p<1e-04)	118	16	2

d Male Child DMRs

p-value	All Window	Multiple Window
0.001	317	37
1e-04	**61**	**7**
1e-05	14	1
1e-06	4	0
1e-07	1	0

Significant windows (1 kb)	1	2	3
Number of DMR (p<1e-04)	54	6	1

e Mother DMR Alterations



f Father DMR Alterations

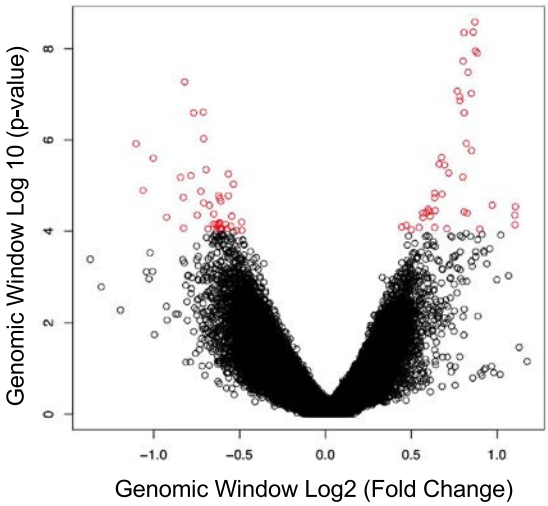
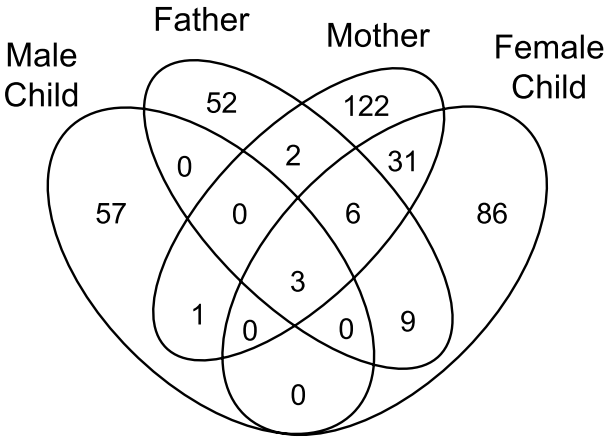


Figure 1. DMR identifi ation and numbers. The number of DMRs found using different p-value cutoff thresholds. The All-Window column shows all DMRs. The Multiple Window column shows the number of DMRs containing at least two nearby signifi ant windows (1 kb each). The number of DMRs with the number of signifi ant windows (1 kb per window) at a p value threshold of $p < 1e{-}04$ for DMR is bolded. (**a**) Mother DMRs; (**b**) Father DMRs; (**c**) Female child DMRs; (**d**) Male child DMRs; (**e**) Mother; and (**f**) Father log-fold-change DMR alterations. The red circles are statistically signifi ant DMRs showing log-fold change distribution (i.e., increase or decrease DNA methylation).

The size of the DMRs for each group is 1 or 2 kb with a CpG density less than 3 CpG/100 bp, Supplemental Figure S3. These regions with low CpG density are considered CpG desserts[51], which represents the majority (> 90%) of the genome, but some DMRs are observed at higher 8–10 CpG/100 bp density associated with CpG islands[50].

A principal component analysis (PCA) of the DMRs for the control and case comparison for each group are presented in Supplemental Figure S1. Generally, the case and control DMR principal component 1 and 2 separated samples by treatment group, Supplemental Figure S1A–D. The African American samples circled generally clustered with the appropriate case or control groups, Supplemental Figure S1. Therefore, the racial background did not appear to have major impacts. As previously mentioned, the various clinical parameters in Supplemental Table S1b did not correlate with outliers in the PCA analysis, Supplemental Figure S1. Therefore, the DMRs identified appear to reflect PTB rather than specific pathology parameters or race.

a DMR Overlap (p<1e-04)



b Extended Overlaps (p<1e-04 vs. p<0.05)

p<1e-04 \ p<0.05	Mother	Father	Female Child	Male Child
Mother	165 (100%)	81 (49%)	96 (58%)	50 (30%)
Father	54 (74%)	73 (100%)	47 (64%)	34 (47%)
Female Child	79 (58%)	62 (46%)	136 (100%)	46 (34%)
Male Child	17 (28%)	15 (25%)	11 (18%)	61 (100%)

Figure 2. DMR group overlaps. (**a**) DMR $p < 1e{-}04$ Venn diagram overlap. (**b**) Extended overlaps with $p < 1e{-}04$ and $p < 0.05$ comparisons. DMR number and percent (%) overlap presented within the rows.

A blinded test set of samples were collected to help validate the predictive ability of the PTB samples identified. Five triads for control and five triads for PTB case were collected for analysis. These samples were blinded to the WSU investigators during the analysis and prediction. This test set was analyzed and the data used in dendrogram, machine learning and PCA analysis, as previously described[44]. The accuracy for the test set mother was 50%, father was 40%, and female child 60%. However, after the analyses of the unblinded samples, a very heterogenous equal mixture of moderate, very, and extreme PTB were present. In addition, some batch effects within the assay were detected. Due to the low sample size (n = 5) of the test set and heterogeneity of the samples, this blinded test set analysis was potentially compromised and marginally successful, so not utilized for further analysis. As now discussed in the Discussion section, expanded clinical trials with larger sample size and larger test sample size are required to optimize and validate the epigenetic biomarkers (DMRs) identifi d.

The fi al analysis investigated the DMR associated genes with each mother, father, and child DMR sets. The DMRs within 10 kb of a gene were considered to include proximal and distal promoter regions, as well as the gene. The DMR associated genes listed in Supplemental Tables S2–S5 were identified for gene functional category, Fig. 4a. The cytoskeleton, transport, transcription, and signaling categories were prominent in each group. The DMR associated gene groups were analyzed for KEGG pathways with ≥ 3 genes in the pathway, and the pathways and genes presented for each group, Fig. 4b. The mother DMR associated genes had the highest number of pathways with metabolism, synaptic vesicle cycle, and a number of signaling pathways prominent. The father had metabolism pathway, and male child no pathways. Interestingly, both the mother and female child had microRNA pathways represented (highlighted), Fig. 4b. Th s refl cts DMRs shared between them that contain a cluster of genes and non-coding RNA, including Aopep (aminopeptidase O) and the micro-RNAs Mir 24-1, Mir 27b, Mir 23b, and Mir 3074. Therefore, an additional epigenetic mechanism altered in preterm birth appears to involve ncRNA that was common between the mother and daughter DMRs.

A presentation of the mother, father, and child DMR associated genes with network links, as determined by Pathway Studio (Elsevier, Inc.), are presented in Fig. 5. For each group the three disease states most over-represented in the list of DMR-associated genes are presented. Also included are any DMR associated genes with known associations with disease terms Premature Birth, Very Premature Birth, Preterm Labor, and Premature Rupture of Membranes. The mother, father, and female child groups all had DMR-associated genes previously shown to be linked to preterm birth. These known genes include Rock1, Ghrl1, Fkbp5, Sigirr, Kdr, Mir24-1,

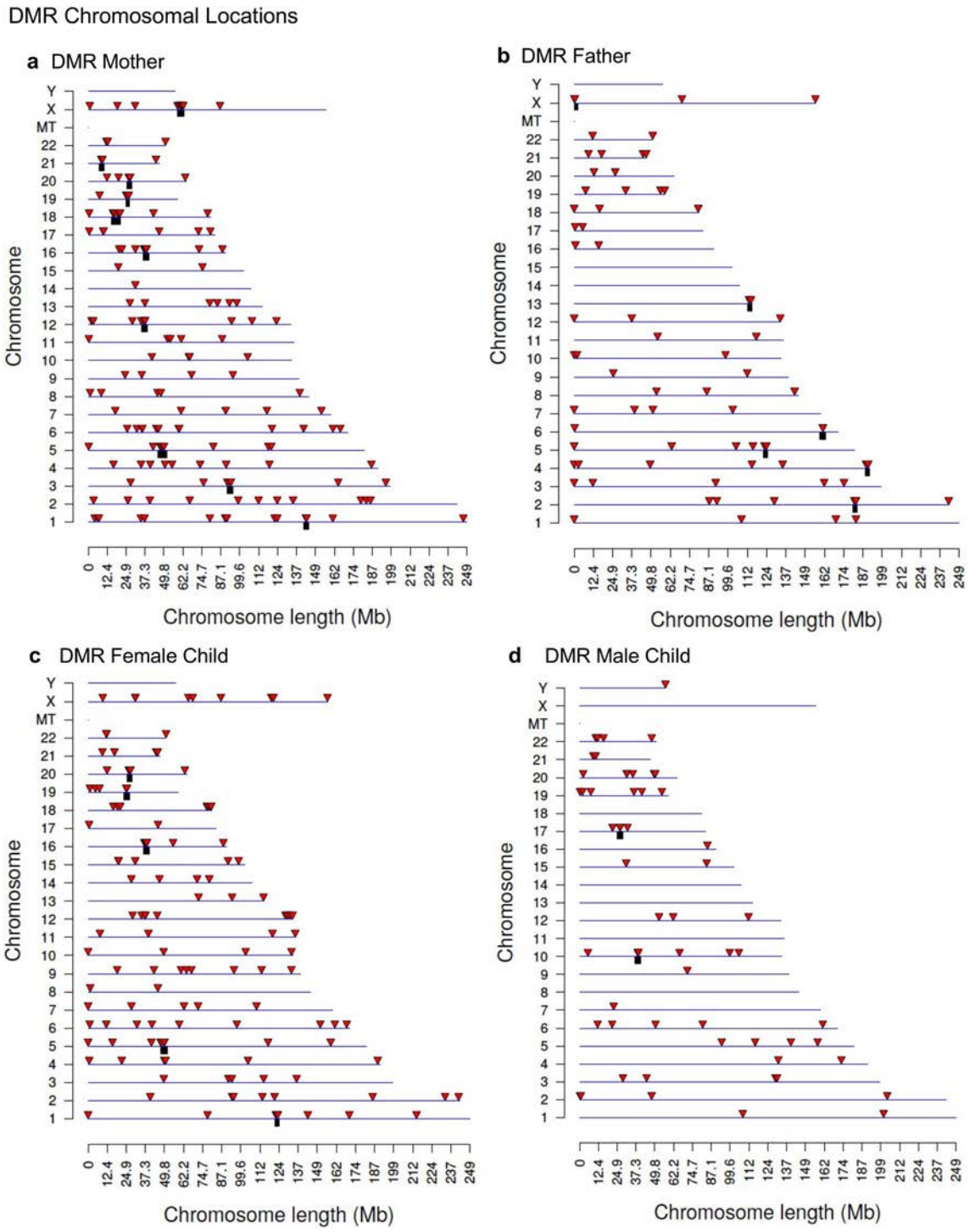


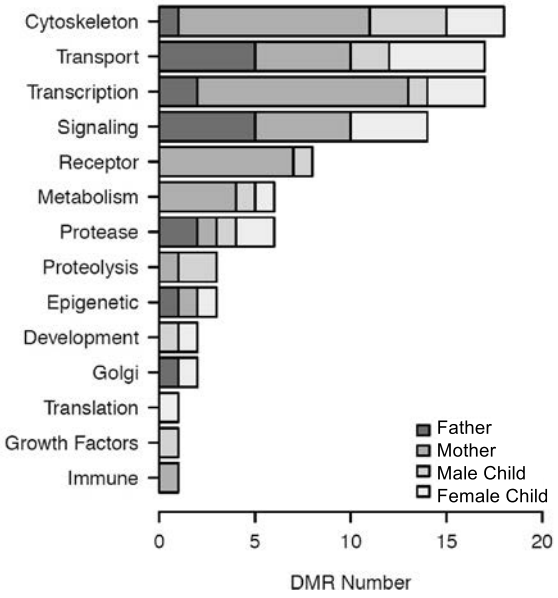
Figure 3. DMR chromosomal locations. The DMR locations on the individual chromosomes is represented with an arrowhead and a cluster of DMRs with a black box. All DMRs containing at least one signifi ant window at a *p* value threshold of $p < 1e{-}04$ for DMR are shown. (**a**) Mother DMRs; (**b**) Father DMRs; (**c**) Female child DMRs; and (**d**) Male child DMRs. The chromosome number versus size (megabase) is presented.

Cacna1c, Neu1, Nlrp1, F7 and F10, Fig. 5. This helps validate the potential PTB DMR biomarkers identifi d, as well as identify potential new DMRs and associated genes for PTB to consider.

Discussion

Preterm birth is a major health concern worldwide, affecting more than one in 10 pregnancies[1]. Even when preterm children survive, they are at higher risk of developing chronic disease conditions[3–5]. These include hypertension, diabetes, metabolic and lipid disorders, heart disease, kidney disease, sleep apnea, and all cause

a DMR Associated Gene Category



b DMR Associated Gene Pathways (≥ 3 Genes)

Mother p<1e-04 DMRs

hsa01100 Metabolic pathways (human) (4)
hsa:55256 ADI1; acireductone dioxygenase 1
hsa:23545 ATP6V0A2; ATPase H+ transporting V0 subunit a2
hsa:3631 INPP4A; inositol polyphosphate-4-phosphatase type I A
hsa:10846 PDE10A; phosphodiesterase 10A

hsa04721 Synaptic vesicle cycle (human) (4)
hsa:23545 ATP6V0A2; ATPase H+ transporting V0 subunit a2
hsa:4905 NSF; N-ethylmaleimide sensitive factor, vesicle fusing ATPase
hsa:112755 STX1B; syntaxin 1B
hsa:10497 UNC13B; unc-13 homolog B

hsa05200 Pathways in cancer (human) (3)
hsa:29119 CTNNA3; catenin alpha 3
hsa:5467 PPARD; peroxisome proliferator activated receptor delta
hsa:6093 ROCK1; Rho associated coiled-coil containing protein kinase 1

hsa04024 cAMP signaling pathway (human) (3)
hsa:775 CACNA1C; calcium voltage-gated channel subunit alpha1 C
hsa:10846 PDE10A; phosphodiesterase 10A
hsa:6093 ROCK1; Rho associated coiled-coil containing protein kinase 1

hsa04022 cGMP-PKG signaling pathway (human) (3)
hsa:775 CACNA1C; calcium voltage-gated channel subunit alpha1 C
hsa:6093 ROCK1; Rho associated coiled-coil containing protein kinase 1
hsa:6546 SLC8A1; solute carrier family 8 member A1

hsa05412 Arrhythmogenic right ventricular cardiomyopathy (human) (3)
hsa:775 CACNA1C; calcium voltage-gated channel subunit alpha1 C
hsa:29119 CTNNA3; catenin alpha 3
hsa:6546 SLC8A1; solute carrier family 8 member A1

hsa04020 Calcium signaling pathway (human) (3)
hsa:775 CACNA1C; calcium voltage-gated channel subunit alpha1 C
hsa:3791 KDR; kinase insert domain receptor
hsa:6546 SLC8A1; solute carrier family 8 member A1

hsa05206 MicroRNAs in cancer (human) (3)
hsa:407011 MIR23B; microRNA 23b
hsa:407019 MIR27B; microRNA 27b
hsa:6093 ROCK1; Rho associated coiled-coil containing protein kinase 1

Father p<1e-04 DMR

hsa01100 Metabolic pathways (human) (5)
hsa:64579 NDST4; N-deacetylase and N-sulfotransferase 4
hsa:29922 NME7; NME/NM23 family member 7
hsa:5136 PDE1A; phosphodiesterase 1A
hsa:93166 PRDM6; PR/SET domain 6
hsa:7357 UGCG; UDP-glucose ceramide glucosyltransferase

Female Child p<1e-04 DMR

hsa05206 MicroRNAs in cancer (human) (4)
hsa:407011 MIR23B; microRNA 23b
hsa:407019 MIR27B; microRNA 27b
hsa:407045 MIR7-3; microRNA 7-3
hsa:6093 ROCK1; Rho associated coiled-coil containing protein kinase 1

hsa01100 Metabolic pathways (human) (3)
hsa:64409 GALNT17; polypeptide N-acetylgalactosaminyltransferase 17
hsa:3632 INPP5A; inositol polyphosphate-5-phosphatase A
hsa:4758 NEU1; neuraminidase 1

Male Child p<1e-04 DMR
None

Figure 4. DMR gene associations (**a**) DMR ($p < 1e{-}04$) associated gene function category frequency. (**b**) DMR associated gene pathways for mother, father, and female child.

mortality[5]. Th s is in part due to the stresses placed on the late-stage fetus, impacting their normal development. These impacts are studied in light of the Developmental Origins of Health and Disease (DOHAD) hypothesis. Previous studies have correlated many adult-onset diseases with fetal and early life developmental stresses[52–54]. The potential to predict preterm birth, and provide interventions to reduce its incidence, would have a signifi ant impact on human health.

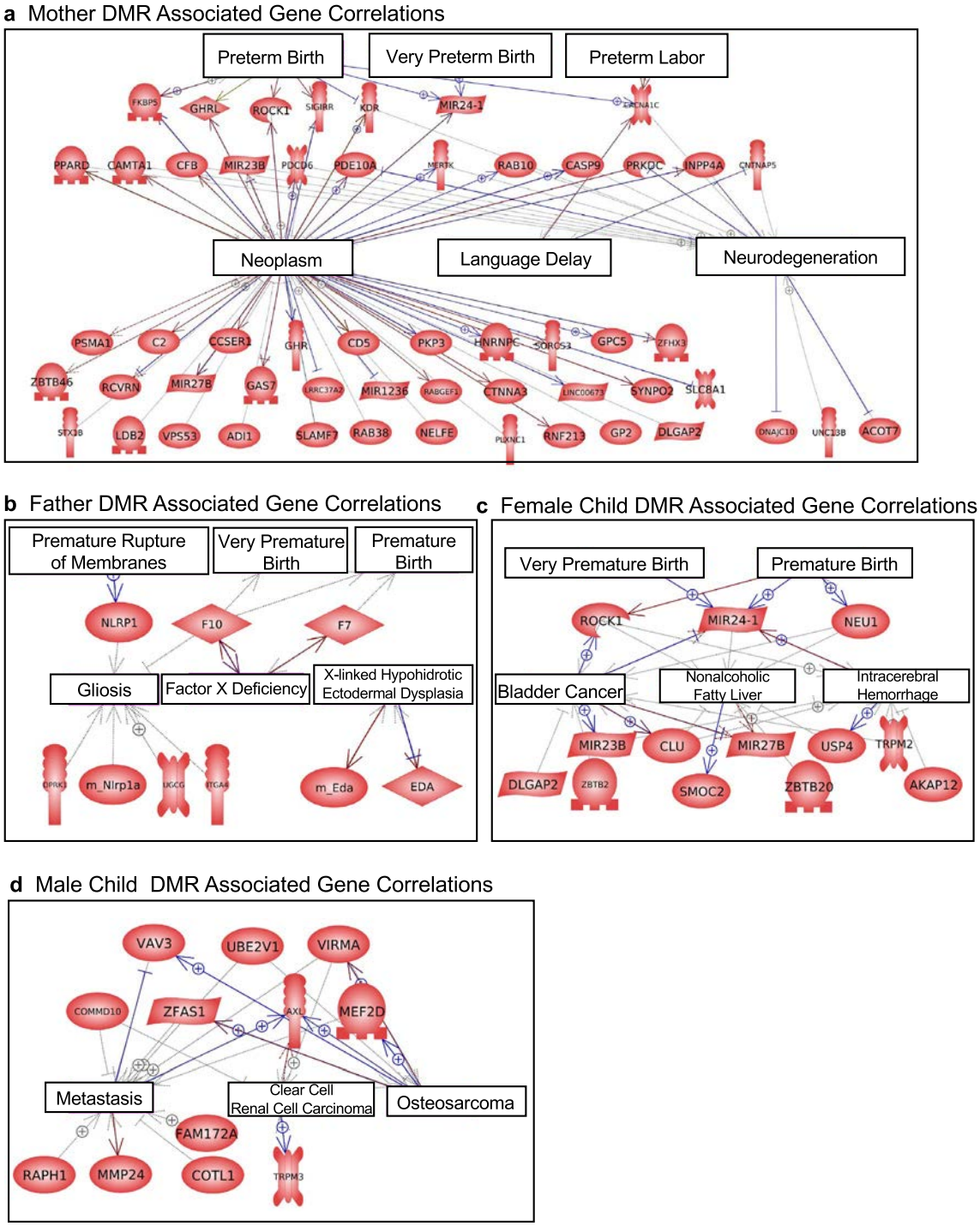


Figure 5. Associated gene networks and correlations. (**a**) Mother DMR associated gene correlations. (**b**) Father DMR associated gene correlations. The gene correlations and associated genes are presented for each disease group. (**c**) Female child DMR associated gene correlations. (**d**) Male child DMR associated gene correlations. The gene correlations and associated genes are presented for each disease pathology.

In this study buccal swab samples were collected from mothers, fathers and newborn infants approximately nine days following birth in cases where preterm birth occurred, and similarly in control full-term births. The buccal epithelial cells were analyzed for sites of DNA methylation in genomic regions when differences in methylation (DMRs) were detected. Mothers, fathers, and children all showed DMR signatures related to preterm birth (Figs. 1, 2). Male children had negligible DMR and a lower false discovery rate confid nce than the other groups. The results of this study suggest that potential epigenetic tests of mothers, as well as fathers, could help predict the risk of preterm birth. However, extended prospective longitudinal pre-conception trials are required to optimize the potential biomarkers and assess the associations with different clinical parameters for preterm birth such as preeclampsia or obesity. Although the infant buccal analyses are not predictive of PTB, the epigenetic differences

seen in children who have experienced preterm birth could potentially be used to assess later life disease (e.g., preterm birth) susceptibility and improve future preventative clinical management approaches. While it has been reported that paternal exposure to phenols is associated with increased incidence of preterm birth[55], most previous studies have found that paternal lifestyle factors do not predict gestation length[56]. The current study identifi d epigenetic changes in both the mothers and fathers of children born preterm, suggesting potential maternal and paternal epigenetic components. Future expanded epigenetic analysis applied to both the mother and father may better assess risk of preterm birth, compared to assays of the mother alone.

The statistical confidence and accuracy of the prediction needs to be improved with expanded clinical trials with larger numbers of samples and trials monitoring individuals prior to conception of the child. Although, the current study demonstrates that epigenetic biomarkers in maternal and paternal buccal cells may be useful, larger studies are needed for predicting preterm birth. In the current study, buccal samples were collected from mothers and fathers immediately after the birth of their child. In the future, prospective studies with sample collection during pregnancy and prior to birth will be needed in order to develop a more clinically relevant predictive assay. Although a prospective study is anticipated to have similar DMR biomarkers, this remains to be confirmed.

In considering the accuracy of the epigenetic biomarkers observed, it is important to optimize with expanded clinical trials that include subpopulations of various sources of PTB such as obesity or preeclampsia. Interestingly some major disease biomarkers work approximately within a 50% accuracy range with either false positives or false negatives to consider. For example, for the major male prostate disease biomarker of Prostate Specifi Antigen (PSA) for prostate cancer, the standard PSA cut-off of 4 ng/mL has low sensitivity. With this cut-off only 20.5% of the prostate cancer cases test positive and nearly 80% of prostate cancer cases are missed. The specific ty at this cut-off is high (93.6%), meaning only 6.2% of men who do not have prostate cancer falsely test positive[57]. Another example is the ovarian cancer Ca125 biomarker which has a low accuracy for screening with both false positives and false negatives being problematic. However, for both PSA and Ca125, monitoring an individual over time does improve the accuracy of the assay to over 70% for monitoring, but not screening[58,59]. In addition, use of additional biomarkers in concert with the PSA and Ca125 has been found to improve the accuracy of screening to approximately 50%[60]. Due to the general low accuracy of such disease biomarkers, there have been a number of qualification and verification parameters put in place to improve and allow greater discovery efforts to be made for disease biomarkers[61,62]. Clearly disease biomarkers are essential for future medicine, but the current major protein-based biomarkers developed have limited use for general screening due to low accuracy. The current study provides large numbers of unique epigenetic-based DMR sites throughout the genome, which appear to relate to preterm birth. Th s is a unique molecular approach that may improve biomarker development. The study's observations are encouraging and support the concept that epigenetic biomarkers derived from surrogate marker cells may be used as a biomarker for preterm birth. However, like PSA and Ca125, further clinical trials are needed to refine and validate the use of epigenetic biomarkers to predict preterm birth.

Previous studies have attempted to identify changes in DNA methylation in pregnant women that could be used as biomarkers of preterm birth. Parets et al.[63] collected peripheral blood leukocyte samples from African American women at the start of labor that delivered either preterm (24–34 weeks; n = 16) or at term (39–41 weeks; n = 24). DNA methylation was assessed using the HumanMethylation450 BeadChip by Illumina. No DNA methylation biomarkers for preterm birth were identifi d, but these researchers did report that there were many DNA methylation changes that were shared between mothers that delivered preterm and their infants[63]. In a larger study of African American women, Hong et al.[64] collected peripheral blood leukocyte samples at the time of labor from 150 women who delivered preterm, and 150 who delivered at term. DNA methylation was assessed using the Illumina HumanOmni2.5-4v1 array. Forty-five DMR were identifi d, of which two were found to be retained in a follow-up replication analysis[64]. Knijnenburg et al.[65] performed a study that evaluated genomic variants, gene expression and DNA methylation simultaneously in whole blood samples taken in the day or two after birth. Two hundred seventy preterm and five hundred twenty-one full term maternal samples were evaluated. DNA methylation was assessed using the Illumina Methylation 450K array. No genomic variants were associated with preterm birth. However, 215 differentially expressed genes and two DMRs were found to be associated with preterm birth. There were greater numbers of molecular differences associated with very early preterm birth (< 28 weeks of gestation). Analysis of the 44 cases of these very early births showed that 217 genetic variants, 838 differentially expressed genes and 811 DMRs were associated[65]. A combined approach like this that uses multiple types of biomarkers shows promise for developing accurate clinical assays to predict preterm birth in the future. As previously mentioned, a limitation of all these studies is the use of mixed cell populations, which can suggest the presence of an epigenetic change, but which is in fact due to alterations in cell population numbers[32,33]. Purified individual cell types are more effective to both identify and assess epigenetic differences as disease biomarkers[41,42].

A number of the previous studies have used the Illumina array platform to identify DMRs as biomarkers of preterm birth[63–65]. These array platforms are biased toward detection of DMR in high density CpG islands, which constitute less than 1% of the genome[50]. However, the majority of the genome has a low density (1–3%) 1–3 CpG/100 bp density [50]. The MeDIP procedure used in the current study is biased toward detection of DNA methylation in regions of lower CpG density of < 5 CpG/100 bp, which corresponds to > 95% of the genome[50]. Using the genome-wide MeDIP procedure to identify DNA methylation alterations increases the feasibility of finding robust epigenetic biomarkers of preterm birth.

In the current study, only approximately half of the DMRs had nearby associated genes. Although the regulatory role of the DMRs to alter gene expression requires further investigation, the potential functional correlations of the DMR-associated genes for each group were evaluated. Genes involved in cytoskeleton, transcription and signaling were prominent in the gene sets (Figs. 4, 5). Among the disease states associated with these gene sets, the cancer pathways were frequently seen, possibly due to signaling abnormalities being prominent in cancer pathways. The mother, father and female child gene sets included DMR associated genes that have been

previously associated with preterm birth (Fig. 5). This occurred even though cheek buccal cells are not directly involved in gestation, which suggests surrogate marker cell samples can be useful to detect epigenetic biomarkers of disease. Th s is supported by a recent study that used buccal cells to identify epigenetic biomarkers for female rheumatoid arthritis[46].

Conclusions

In conclusion, genome-wide differential DNA methylation regions for preterm birth were detected in buccal cells of mothers, fathers, and female children. This provides a "proof of concept" that DNA methylation analysis of buccal swabs of parents may be used to potentially predict preterm birth. However, the accuracy and predictive ability of the biomarker needs to be improved with future clinical trials, as discussed. Such a preterm birth risk or susceptibility biomarker would allow for better obstetrical management to prevent preterm birth, mitigate morbidity in unprevented preterm births (through timely administration of prenatal steroids, magnesium sulfate, tocolytics and optimal delivery procedures), and thus improve the health and long-term outcomes for many children. Unanticipated preterm births continue to catch providers by surprise, and often lead to major morbidities such as intraventricular hemorrhage, severe lung disease and other irreversible injuries. The presence of preterm birth associated DMRs in parental buccal cells suggests potential parental early life exposures and/or ancestral impacts are involved in the etiology of preterm birth. Rodent models have shown that environmental exposures in early pregnancy when epigenetic programming occurs in the fetus impact DMRs in every somatic cell type in the body across the life span of the exposed fetus and its descendants. Parents' buccal cells, thus, may have the epigenetic changes resulting from ancestral exposure and can potentially be used as biomarkers for risk of preterm birth. This assay could also potentially be used in the future to identify environmental exposures and risk factors that promote preterm birth.

Methods

Clinical sample collection and analysis. St. Franciscan Hospital and Indiana University School of Medicine. IU Health Hospitals (Riley Hospital for Children, IUH Methodist, IUH North) and Franciscan Health, Indianapolis, Indiana, USA provided samples for the current study. Informed consent and HIPAA authorization was obtained from all participants prior to the clinical sample collection. The study protocol was approved by the Indiana University Institutional Review Board (IRB) #1901985132 and the Franciscan Institutional Review Board (IRB), #1489434-5. All research was performed in accordance with relevant guidelines/regulations. Informed consent and HIPAA authorization was obtained from all participants prior to sample collection. For sample collection involving human participants that are minors, informed consent from a parent and/or legal guardian for study participation was obtained prior to sample collection. Buccal samples were collected from the mother, father, and child in instances where pre-term birth occurred (case), or where term birth occurred (control), approximately nine days following birth. Th s period was used to allow the case PTB child to mature and allow and effective buccal cell collection. The demographic data for these subjects is presented in Supplemental Table S1. Buccal swabs were stored at -80 C until use.

DNA preparation. Frozen human buccal samples were thawed for analysis. Genomic DNA from buccal samples was prepared as follows: The buccal brush was suspended in 750 µL of cell lysis solution and 3.5 µL of Proteinase K (20 mg/mL). Th s suspension was incubated at 55 °C for 3 h, then vortexed and centrifuged briefly. The lysis solution was then transferred to a new 1.5 µL microcentrifuge tube. The microcentrifuge tube with the buccal brush was centrifuged again to retain any remaining solution which was combined with the transferred lysis solution. The buccal brush was discarded and 300 µL of protein precipitation solution (Promega, A795A, Madison, WI) was added to the lysis solution. The sample was incubated on ice for 15 min, then centrifuged at 4C for 30 min. The supernatant was transferred to a fresh 2 mL microcentrifuge tube and 1000 µL ice cold isopropanol was added along with 2 µL glycoblue. Th s suspension was mixed thoroughly and incubated at − 20 °C overnight. The suspension was then centrifuged at 4°C for 20 min, the supernatant was discarded, and the pellet was washed with 75% ethanol, then air-dried and resuspended in 100 µL H2O. DNA concentration was measured using the Nanodrop (Thermo Fisher, Waltham, MA).

Methylated DNA immunoprecipitation (MeDIP). Methylated DNA Immunoprecipitation (MeDIP) with genomic DNA was performed as follows: individual DNA samples (2–4 ug of total DNA) were diluted to 130 µL with 1 × Tris–EDTA (TE, 10 mM Tris, 1 mM EDTA) and sonicated with the Covaris M220 using the 300 bp setting. Fragment size was verifi d on a 2% E-gel agarose gel. The sonicated DNA was transferred from the Covaris tube to a 1.7 mL microfuge tube, and the volume was measured. The sonicated DNA was then diluted with TE buffer (10 mM Tris HCl, pH7.5; 1 mM EDTA) to 400 µL, heat-denatured for 10 min at 95 C, then immediately cooled on ice for 10 min. Then 100 µL of 5X IP buffer and 5 µg of antibody (monoclonal mouse anti 5-methyl cytidine; Diagenode #C15200006) were added to the denatured sonicated DNA. The DNA-antibody mixture was incubated overnight on a rotator at 4 C. The following day magnetic beads (Dynabeads M-280 Sheep anti-Mouse IgG; 11201D) were pre-washed as follows: The beads were resuspended in the vial, then the appropriate volume (50 µL per sample) was transferred to a microfuge tube. The same volume of Washing Buffer (at least 1 mL 1XPBS with 0.1% BSA and 2 mM EDTA) was added and the bead sample was resuspended. The tube was then placed into a magnetic rack for 1–2 min and the supernatant was discarded. The tube was removed from the magnetic rack and the beads were washed once. The washed beads were resuspended in the same volume of 1xIP buffer (50 mM sodium phosphate ph7.0, 700 mM NaCl, 0.25% TritonX-100) as the initial volume of beads. 50 µL of beads were added to the 500 µL of DNA-antibody mixture from the overnight incubation, then incubated for 2 h on a rotator at 4 C. After the incubation, the bead-antibody-DNA complex was washed

three times with 1X IP buffer as follows: The tube was placed into a magnetic rack for 1–2 min and the supernatant was discarded, then the magnetic bead antibody pellet was washed with 1xIP buffer 3 times. The washed bead antibody DNA pellet was then resuspended in 250 µL digestion buffer with 3.5 µL Proteinase K (20 mg/mL). The sample was incubated for 2–3 h on a rotator at 55 C, then 250 µL of buffered Phenol–Chloroform-Isoamylalcohol solution was added to the sample, and the tube was vortexed for 30 s and then centrifuged at 14,000 rpm for 5 min at room temperature. The aqueous supernatant was carefully removed and transferred to a fresh microfuge tube. Then 250 µL chloroform were added to the supernatant from the previous step, vortexed for 30 s and centrifuged at 14,000 rpm for 5 min at room temperature. The aqueous supernatant was removed and transferred to a fresh microfuge tube. To the supernatant 2 µL of glycoblue (20 mg/mL), 20 µL of 5 M NaCl and 500 µL ethanol were added and mixed well, then precipitated in -20 C freezer for 1 h to overnight. The precipitate was centrifuged at 14,000 rpm for 20 min at 4 C and the supernatant was removed, while not disturbing the pellet. The pellet was washed with 500 µL cold 70% ethanol in − 20 C freezer for 15 min then centrifuged again at 14,000 rpm for 5 min at 4 C and the supernatant was discarded. The tube was spun again briefly to collect residual ethanol to the bottom of the tube and as much liquid as possible was removed with gel loading tip. The pellet was air-dried at RT until it looked dry (about 5 min) then resuspended in 20 µL H2O or TE. DNA concentration was measured in Qubit (Life Technologies) with ssDNA kit (Molecular Probes Q10212).

MeDIP-Seq analysis. The MeDIP DNA samples (50 ng of each) were used to create libraries for next generation sequencing (NGS) using the NEBNext Ultra RNA Library Prep Kit for Illumina (San Diego, CA) starting at step 1.4 of the manufacturer's protocol to generate double stranded DNA. After this step the manufacturer's protocol was followed. Each sample received a separate index primer. NGS was performed at WSU Spokane Genomics Core using the Illumina HiSeq 2500 with a PE50 application, with a read size of approximately 50 bp and approximately 5–35 million reads per sample, and 6–7 sample libraries each were run in one lane.

Molecular bioinformatics and statistics. Basic read quality was verifi d using information produced by the FastQC program[66]. Reads were filtered and trimmed to remove low quality base pairs using Trimmomatic[67]. The reads for each sample were mapped to the GRCh38 human genome using Bowtie2[68] with default parameter options. The mapped read files were then converted to sorted BAM files using SAMtools[69]. To identify DMR, the reference genome was broken into 1000 bp windows. The MEDIPS R package[70] was used to calculate differential coverage between control and exposure sample groups. The edgeR *p* value[71] was used to determine the relative difference between the two groups for each genomic window. Windows with an edgeR p-value less than 10^{-4} were considered DMRs. The DMR edges were extended until no genomic window with an edgeR p-value less than 0.1 remained within 1000 bp of the DMR. CpG density and other information was then calculated for the DMR based on the reference genome. DMR were annotated using the NCBI provided annotations. The genes that overlapped with DMR were then input into the KEGG pathway search[72,73] to identify associated pathways. The DMR associated genes were then sorted into functional groups by reducing Panther[74] protein classifi ations into more general categories. All MeDIP-Seq genomic data obtained in the current study have been deposited in the NCBI public GEO database (GEO #: GSE194227).

Blinded test set analysis was performed to classify test samples into case or control groups. Samples from ten novel trios were collected to evaluate the efficacy of using the DMR sets identified as a biomarker for preterm birth. The test samples were processed identically to the samples used in the main analysis. PCA and cluster dendrogram analyses were used to search for test samples that clustered with the known samples when only DMR sites were considered. Additionally, linear discriminant analysis (LDA) and random forest (RF) classification was performed to identify which blinded samples were preterm birth, as previously described[44].

Ethics approval and consent to participate. Approvals to conduct the study were obtained from Indiana University Institutional Review Board (IRB) #1901985132 and the Franciscan Institutional Review Board (IRB), #1489434-5.

Data availability

All molecular data have been deposited into the public database at NCBI (GEO # GSE194227), and R code computational tools are available at GitHub (https://github.com/skinnerlab/MeDIP-seq) and www.skinner.wsu.edu.

Received: 13 November 2021; Accepted: 11 February 2022
Published online: 01 March 2022

References

1. Chawanpaiboon, S. *et al.* Global, regional, and national estimates of levels of preterm birth in 2014: a systematic review and modelling analysis. *Lancet Glob. Health* **7**, e37–e46. https://doi.org/10.1016/S2214-109X(18)30451-0 (2019).
2. You, D., New, J. R. & Wardlaw, T. Levels and trends in child mortality. Report 2015. Estimates developed by the UN Inter-agency Group for Child Mortality Estimation. (United Nations Children's Fund, 2017).
3. Soleimani, F., Zaheri, F. & Abdi, F. Long-term neurodevelopmental outcomes after preterm birth. *Iran Red Crescent Med J* **16**, e17965. https://doi.org/10.5812/ircmj.17965 (2014).
4. Tanz, L. J. *et al.* Preterm delivery and maternal cardiovascular disease in young and middle-aged adult women. *Circulation* **135**, 578–589. https://doi.org/10.1161/CIRCULATIONAHA.116.025954 (2017).
5. Crump, C. An overview of adult health outcomes after preterm birth. *Early Hum. Dev.* **150**, 105187. https://doi.org/10.1016/j.earlhumdev.2020.105187 (2020).
6. Jelliffe-Pawlowski, L. L. *et al.* Maternal characteristics and mid-pregnancy serum biomarkers as risk factors for subtypes of preterm birth. *BJOG Int. J. Obstet. Gynaecol.* **122**, 1484–1493. https://doi.org/10.1111/1471-0528.13495 (2015).

7. Smith, G. C. *et al.* Maternal and biochemical predictors of spontaneous preterm birth among nulliparous women: A systematic analysis in relation to the degree of prematurity. *Int. J. Epidemiol.* **35**, 1169–1177. https://doi.org/10.1093/ije/dyl154 (2006).
8. Tancrede, S. *et al.* Mid-trimester maternal serum AFP and hCG as markers of preterm and term adverse pregnancy outcomes. *J. Obstet. Gynaecol. Can.* **37**, 111–116. https://doi.org/10.1016/s1701-2163(15)30331-5 (2015).
9. Liu, X. *et al.* Effects of prenatal exposure to air particulate matter on the risk of preterm birth and roles of maternal and cord blood LINE-1 methylation: A birth cohort study in Guangzhou, China. *Environ. Int.* **133**, 105177. https://doi.org/10.1016/j.envint.2019.105177 (2019).
10. Romero, R., Dey, S. K. & Fisher, S. J. Preterm labor: One syndrome, many causes. *Science* **345**, 760–765. https://doi.org/10.1126/science.1251816 (2014).
11. Bolton, C. E., Bush, A., Hurst, J. R., Kotecha, S. & McGarvey, L. Lung consequences in adults born prematurely. *Thorax* **70**, 574–580. https://doi.org/10.1136/thoraxjnl-2014-206590 (2015).
12. Zhang, G. *et al.* Genetic associations with gestational duration and spontaneous preterm birth. *N. Engl. J. Med.* **377**, 1156–1167. https://doi.org/10.1056/NEJMoa1612665 (2017).
13. Ferrero, D. M. *et al.* Cross-country individual participant analysis of 4.1 million singleton births in 5 countries with very high human development index confirms known associations but provides no biologic explanation for 2/3 of all preterm births. *PLoS ONE* **11**, e0162506. https://doi.org/10.1371/journal.pone.0162506 (2016).
14. Dugoff, L., Society for Maternal-Fetal Medicine. First- and second-trimester maternal serum markers for aneuploidy and adverse obstetric outcomes. *Obstet. Gynecol.* **115**, 1052–1061. https://doi.org/10.1097/AOG.0b013e3181da93da (2010).
15. Menon, R., Bhat, G., Saade, G. R. & Spratt, H. Multivariate adaptive regression splines analysis to predict biomarkers of spontaneous preterm birth. *Acta Obstet. Gynecol. Scand.* **93**, 382–391. https://doi.org/10.1111/aogs.12344 (2014).
16. Wallenstein, M. B. *et al.* Inflammatory biomarkers and spontaneous preterm birth among obese women. *J. Matern. Fetal Neonatal Med.* **29**, 3317–3322. https://doi.org/10.3109/14767058.2015.1124083 (2016).
17. Jelliffe-Pawlowski, L. L. *et al.* Combined elevated midpregnancy tumor necrosis factor alpha and hyperlipidemia in pregnancies resulting in early preterm birth. *Am. J. Obstet. Gynecol.* **211**(141), e141-149. https://doi.org/10.1016/j.ajog.2014.02.019 (2014).
18. Sorokin, Y. *et al.* Maternal serum interleukin-6, C-reactive protein, and matrix metalloproteinase-9 concentrations as risk factors for preterm birth <32 weeks and adverse neonatal outcomes. *Am. J. Perinatol.* **27**, 631–640. https://doi.org/10.1055/s-0030-1249366 (2010).
19. Aung, M. T. *et al.* Maternal lipidomic signatures in relation to spontaneous preterm birth and large-for-gestational age neonates. *Sci. Rep.* **11**, 8115. https://doi.org/10.1038/s41598-021-87472-9 (2021).
20. Manuck, T. A., Eaves, L. A., Rager, J. E. & Fry, R. C. Mid-pregnancy maternal blood nitric oxide-related gene and miRNA expression are associated with preterm birth. *Epigenomics* **13**, 667–682. https://doi.org/10.2217/epi-2020-0346 (2021).
21. Chim, S. S. C., Chan, T. F. & Leung, T. Y. Whole-transcriptome analysis of maternal blood for identification of RNA markers for predicting spontaneous preterm birth among preterm labour women: Abridged secondary publication. *Hong Kong Med. J.* **26**(Suppl 6), 20–23 (2020).
22. Heng, Y. J. *et al.* Human cervicovaginal fluid biomarkers to predict term and preterm labor. *Front. Physiol.* **6**, 151. https://doi.org/10.3389/fphys.2015.00151 (2015).
23. Winger, E. E. *et al.* MicroRNAs isolated from peripheral blood in the fi st trimester predict spontaneous preterm birth. *PLoS ONE* **15**, e0236805. https://doi.org/10.1371/journal.pone.0236805 (2020).
24. Zhou, G., Holzman, C., Heng, Y. J., Kibschull, M. & Lye, S. J. Maternal blood EBF1-based microRNA transcripts as biomarkers for detecting risk of spontaneous preterm birth: A nested case-control study. *J. Matern. Fetal Neonatal Med.* https://doi.org/10.1080/14767058.2020.1745178 (2020).
25. Rosen, E. M. *et al.* Urinary oxidative stress biomarkers and accelerated time to spontaneous delivery. *Free Radic. Biol. Med.* **130**, 419–425. https://doi.org/10.1016/j.freeradbiomed.2018.11.011 (2019).
26. Millan, I. *et al.* Oxidative stress in the newborn period: Useful biomarkers in the clinical setting. *Antioxidants (Basel)* https://doi.org/10.3390/antiox7120193 (2018).
27. Ronde, E. *et al.* The potential of metabolomic analyses as predictive biomarkers of preterm delivery: A systematic review. *Front. Endocrinol. (Lausanne)* **12**, 668417. https://doi.org/10.3389/fendo.2021.668417 (2021).
28. Menon, R., Conneely, K. N. & Smith, A. K. DNA methylation: An epigenetic risk factor in preterm birth. *Reprod. Sci.* **19**, 6–13. https://doi.org/10.1177/1933719111424446 (2012).
29. Knight, A. K. & Smith, A. K. Epigenetic biomarkers of preterm birth and its risk factors. *Genes (Basel)* https://doi.org/10.3390/genes7040015 (2016).
30. Park, B. *et al.* Epigenetic biomarkers and preterm birth. *Environ. Epigenet.* **6**, dvaa005. https://doi.org/10.1093/eep/dvaa005 (2020).
31. Dumeige, L. *et al.* Preterm birth is associated with epigenetic programming of transgenerational hypertension in mice. *Exp. Mol. Med.* **52**, 152–165. https://doi.org/10.1038/s12276-020-0373-5 (2020).
32. Skinner, M. K. Environmental epigenetic transgenerational inheritance and somatic epigenetic mitotic stability. *Epigenetics* **6**, 838–842 (2011).
33. Nilsson, E., Sadler-Riggleman, I. & Skinner, M. K. Environmentally induced epigenetic transgenerational inheritance of disease. *Environ. Epigenet.* **4**, 1–13. https://doi.org/10.1093/eep/dvy016 (2018).
34. Hannon, E. *et al.* Assessing the co-variability of DNA methylation across peripheral cells and tissues: Implications for the interpretation of fi dings in epigenetic epidemiology. *PLoS Genet.* **17**, e1009443. https://doi.org/10.1371/journal.pgen.1009443 (2021).
35. Agarwal, P. *et al.* Maternal obesity, diabetes during pregnancy and epigenetic mechanisms that influence the developmental origins of cardiometabolic disease in the offspring. *Crit. Rev. Clin. Lab. Sci.* **55**, 71–101. https://doi.org/10.1080/10408363.2017.1422109 (2018).
36. Zhang, L., Lu, Q. & Chang, C. Epigenetics in health and disease. *Adv. Exp. Med. Biol.* **1253**, 3–55. https://doi.org/10.1007/978-981-15-3449-2_1 (2020).
37. Nilsson, E. E. & Skinner, M. K. Environmentally induced epigenetic transgenerational inheritance of disease susceptibility. *Transl. Res.* **165**, 12–17 (2015).
38. Wang, X. M. *et al.* Comparison of DNA methylation profiles associated with spontaneous preterm birth in placenta and cord blood. *BMC Med. Genom.* **12**, 1. https://doi.org/10.1186/s12920-018-0466-3 (2019).
39. Wu, Y. *et al.* Analysis of two birth tissues provides new insights into the epigenetic landscape of neonates born preterm. *Clin. Epigenet.* **11**, 26. https://doi.org/10.1186/s13148-018-0599-4 (2019).
40. Spada, E. *et al.* Epigenome wide association and stochastic epigenetic mutation analysis on cord blood of preterm birth. *Int. J. Mol. Sci.* https://doi.org/10.3390/ijms21145044 (2020).
41. Skinner, M. K. Differential DNA methylation analysis optimally requires purified cell populations. *Fertil. Steril.* **106**, 551. https://doi.org/10.1016/j.fertnstert.2016.06.008 (2016).
42. Lu, T. *et al.* Detecting cord blood cell type-specific epigenetic associations with gestational diabetes mellitus and early childhood growth. *Clin. Epigenet.* **13**, 131. https://doi.org/10.1186/s13148-021-01114-5 (2021).
43. Luján, S. *et al.* Sperm DNA methylation epimutation biomarkers for male infertility and FSH therapeutic responsiveness. *Sci. Rep.* **9**, 16786. https://doi.org/10.1038/s41598-019-52903-1 (2019).
44. Garrido, N. *et al.* Sperm DNA methylation epimutation biomarker for paternal off pring autism susceptibility. *Clin. Epigenet.* **13**, 6. https://doi.org/10.1186/s13148-020-00995-2 (2021).

45. Knight, A. K. *et al.* SLC9B1 methylation predicts fetal intolerance of labor. *Epigenet. Off. J. DNA Methylation Soc.* **13**, 33–39. https://doi.org/10.1080/15592294.2017.1411444 (2018).
46. Craig, G. *et al.* Epigenome association study for DNA methylation biomarkers in buccal and monocyte cells for female rheumatoid arthritis. *Sci. Rep.* **11**, 23789. https://doi.org/10.1038/s41598-021-03170-6 (2021).
47. Bearer, E. L. & Mulligan, B. S. Epigenetic changes associated with early life experiences: Saliva, a biospecimen for DNA methylation signatures. *Curr. Genom.* **19**, 676–698. https://doi.org/10.2174/1389202919666180307150508 (2018).
48. Turinsky, A. L., Butcher, D. T., Choufani, S., Weksberg, R. & Brudno, M. Don't brush off buccal data heterogeneity. *Epigenet. Off. J. DNA Methylation Soc.* **14**, 109–117. https://doi.org/10.1080/15592294.2019.1581592 (2019).
49. Ben Maamar, M., Sadler-Riggleman, I., Beck, D. & Skinner, M. K. Genome-wide mapping of DNA methylation 5mC by methylated DNA immunoprecipitation (MeDIP)-sequencing. *DNA Modif. Methods Mol. Biol.* **2198**, 301–310. https://doi.org/10.1007/978-1-0716-0876-0_23 (2021).
50. Beck, D., Ben Maamar, M. & Skinner, M. K. Genome-wide CpG density and DNA methylation analysis method (MeDIP, RRBS, and WGBS) comparisons. *Epigenet. Off. J. DNA Methylation Soc.* https://doi.org/10.1080/15592294.2021.1924970 (2021).
51. Skinner, M. K. & Guerrero-Bosagna, C. Role of CpG deserts in the epigenetic transgenerational inheritance of differential DNA methylation regions. *BMC Genom.* **15**, 692 (2014).
52. Suzuki, K. The developing world of DOHaD. *J. Dev. Orig. Health Dis.* **9**, 266–269. https://doi.org/10.1017/S2040174417000691 (2018).
53. Bianco-Miotto, T., Craig, J. M., Gasser, Y. P., van Dijk, S. J. & Ozanne, S. E. Epigenetics and DOHaD: From basics to birth and beyond. *J. Dev. Orig. Health Dis.* **8**, 513–519. https://doi.org/10.1017/S2040174417000733 (2017).
54. Goldstein, J. A., Gallagher, K., Beck, C., Kumar, R. & Gernand, A. D. Maternal-fetal inflammation in the placenta and the developmental origins of health and disease. *Front. Immunol.* **11**, 531543. https://doi.org/10.3389/fimmu.2020.531543 (2020).
55. Mustieles, V. *et al.* Maternal and paternal preconception exposure to phenols and preterm birth. *Environ. Int.* **137**, 105523. https://doi.org/10.1016/j.envint.2020.105523 (2020).
56. Oldereid, N. B. *et al.* The effect of paternal factors on perinatal and paediatric outcomes: A systematic review and meta-analysis. *Hum. Reprod. Update* **24**, 320–389. https://doi.org/10.1093/humupd/dmy005 (2018).
57. Ankerst, D. P. & Thompson, I. M. Sensitivity and specific ty of prostate-specific antigen for prostate cancer detection with high rates of biopsy verification. *Arch. Ital. Urol. Androl.* **78**, 125–129 (2006).
58. Visintin, I. *et al.* Diagnostic markers for early detection of ovarian cancer. *Clin. Cancer Res.* **14**, 1065–1072. https://doi.org/10.1158/1078-0432.CCR-07-1569 (2008).
59. Jin, W., Fei, X., Wang, X., Song, Y. & Chen, F. Detection and prognosis of prostate cancer using blood-based biomarkers. *Mediat. Inflamm.* **2020**, 8730608. https://doi.org/10.1155/2020/8730608 (2020).
60. Udagawa, Y. *et al.* Clinical characteristics of a newly developed ovarian tumour marker, galactosyltransferase associated with tumour (GAT). *Eur. J. Cancer* **34**, 489–495. https://doi.org/10.1016/s0959-8049(97)10079-x (1998).
61. Zhao, Y. & Brasier, A. R. Qualifi ation and verifi ation of protein biomarker candidates. *Adv. Exp. Med. Biol.* **919**, 493–514. https://doi.org/10.1007/978-3-319-41448-5_23 (2016).
62. Kraus, V. B. Biomarkers as drug development tools: Discovery, validation, qualifi ation and use. *Nat. Rev. Rheumatol.* **14**, 354–362. https://doi.org/10.1038/s41584-018-0005-9 (2018).
63. Parets, S. E., Conneely, K. N., Kilaru, V., Menon, R. & Smith, A. K. DNA methylation provides insight into intergenerational risk for preterm birth in African Americans. *Epigenet. Off. J. DNA Methylation Soc.* **10**, 784–792. https://doi.org/10.1080/15592294.2015.1062964 (2015).
64. Hong, X. *et al.* Genome-wide DNA methylation associations with spontaneous preterm birth in US blacks: Findings in maternal and cord blood samples. *Epigenet. Off. J. DNA Methylation Soc.* **13**, 163–172. https://doi.org/10.1080/15592294.2017.1287654 (2018).
65. Knijnenburg, T. A. *et al.* Genomic and molecular characterization of preterm birth. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 5819–5827. https://doi.org/10.1073/pnas.1716314116 (2019).
66. Andrews, S. *FastQC: A quality control tool for high throughput sequence data.* , https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (2010).
67. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flex ble trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170 (2014).
68. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359. https://doi.org/10.1038/nmeth.1923 (2012).
69. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352 (2009).
70. Lienhard, M., Grimm, C., Morkel, M., Herwig, R. & Chavez, L. MEDIPS: Genome-wide differential coverage analysis of sequencing data derived from DNA enrichment experiments. *Bioinformatics* **30**, 284–286. https://doi.org/10.1093/bioinformatics/btt650 (2014).
71. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140. https://doi.org/10.1093/bioinformatics/btp616 (2010).
72. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
73. Kanehisa, M. *et al.* Data, information, knowledge and principle: Back to metabolism in KEGG. *Nucleic Acids Res.* **42**, D199-205. https://doi.org/10.1093/nar/gkt1076 (2014).
74. Mi, H., Muruganujan, A., Casagrande, J. T. & Thomas, P. D. Large-scale gene function analysis with the PANTHER classifi ation system. *Nat. Protoc.* **8**, 1551–1566. https://doi.org/10.1038/nprot.2013.092 (2013).

Acknowledgements

We acknowledge Ms. Cathy Proctor, Ms. Donna Watkins, Ms. Leah Engelstad, Ms. Dianne Herron, and Mr. Jeffrey Joyce at Indiana University for clinical recruitment and sample collection assistance and, Dr. Jennifer L.M. Thorson, Dr. Millissia Ben Maamar, Mr. Ryan Thompson, Ms. Skylar Shea Davidson, Ms. Makena Horne, Ms. Emma Impala, and Ms. Rachel LaRosa for technical assistance. We acknowledge Ms. Amanda Quilty for editing and Ms. Heather Johnson for assistance in preparation of the manuscript. We thank the Genomics Core laboratory at WSU Spokane for sequencing data. Th s study was supported by the John Templeton Foundation (50183 and 61174) (https://templeton.org/) grants to MKS. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions

P.W. patient's recruitment, clinical and sample collection oversight, data analysis, editing manuscript. E.N. sample processing, data analysis, editing manuscript. D.B. bioinformatics, data analysis, editing manuscript. M.K.S. conceived, data analysis, funding acquisition, wrote and edited manuscript.

Funding
Th s study was supported by the John Templeton Foundation (50183 and 61174) (https://templeton.org/) grants to MKS. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-07262-9.

Correspondence and requests for materials should be addressed to M.K.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access Th s article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2022