All Dugoni School of Dentistry Faculty Articles

All Faculty Scholarship

1-1-2019

Multiomics modeling of the immunome, transcriptome, microbiome, proteome and metabolome adaptations during human pregnancy

Mohammad Sajjad Ghaemi
Stanford University School of Medicine

Daniel B. DiGiulio
Stanford University School of Medicine

Kévin Contrepois
Stanford University School of Medicine

Benjamin Callahan
Stanford University School of Medicine

Thuy T.M. Ngo
Stanford University

Follow this and additional works at: https://scholarlycommons.pacific.edu/dugoni-facarticles

Part of the Medicine and Health Sciences Commons

See next page for additional authors

Authors

Mohammad Sajjad Ghaemi, Daniel B. DiGiulio, Kévin Contrepois, Benjamin Callahan, Thuy T.M. Ngo, Brittany Lee-Mcmullen, Benoit Lehallier, Anna Robaczewska, David McIlwain, Yael Rosenberg-Hasson, Ronald J. Wong, Cecele Quaintance, Anthony Culos, Natalie Stanley, Athena Tanada, Amy Tsai, Dyani Gaudilliere, Edward Ganio, Xiaoyuan Han, Kazuo Ando, Leslie McNeil, Martha Tingle, Paul Wise, Ivana Maric, Marina Sirota, Tony Wyss-Coray, Virginia D. Winn, Maurice L. Druzin, and Ronald S. Gibbs

OXFORD

Systems biology

Multiomics modeling of the immunome, transcriptome, microbiome, proteome and metabolome adaptations during human pregnancy

Mohammad Sajjad Ghaemi[1,2,3,4]**, Daniel B. DiGiulio**[5,6]**,
Kévin Contrepois**[7]**, Benjamin Callahan**[5,8]**, Thuy T. M. Ngo**[9,10]**,
Brittany Lee-McMullen**[7]**, Benoit Lehallier**[11]**, Anna Robaczewska**[5,6]**,
David Mcilwain**[12]**, Yael Rosenberg-Hasson**[13]**, Ronald J. Wong**[14]**,
Cecele Quaintance**[14]**, Anthony Culos**[1]**, Natalie Stanley**[1]**,
Athena Tanada**[1]**, Amy Tsai**[1]**, Dyani Gaudilliere**[1]**, Edward Ganio**[1]**,
Xiaoyuan Han**[1]**, Kazuo Ando**[1]**, Leslie McNeil**[1]**, Martha Tingle**[1]**,
Paul Wise**[14]**, Ivana Maric**[14]**, Marina Sirota**[15,16]**, Tony Wyss-Coray**[11]**,
Virginia D. Winn**[17]**, Maurice L. Druzin**[17]**, Ronald Gibbs**[17]**,
Gary L. Darmstadt**[14]**, David B. Lewis**[14]**, Vahid Partovi Nia**[2,3]**,
Bruno Agard**[2,4]**, Robert Tibshirani**[18,19]**, Garry Nolan**[12]**,
Michael P. Snyder**[7]**, David A. Relman**[5,6,12]**, Stephen R. Quake**[9]**,
Gary M. Shaw**[14]**, David K. Stevenson**[14]**, Martin S. Angst**[1,†]**,
Brice Gaudilliere**[1,†] **and Nima Aghaeepour**[1,*,†]

[1]Department of Anesthesiology, Perioperative and Pain Medicine, Stanford University School of Medicine, Stanford, CA, 94305, USA, [2]Département de Mathématiques et de Génie Industriel, École Polytechnique de Montréal, QC, H3T 1J4, Canada, [3]Groupe d'Études et de Recherche en Analyse des Décision (GERAD), Montréal, QC, H3T 1J4, Canada, [4]Centre Interuniversitaire de Recherche sur les Réseaux d'Entreprise, la Logistique et le Transport (CIRRELT), Montréal, QC, H3T 1J4, Canada, [5]Department of Medicine, Stanford University School of Medicine, Stanford, CA, 94305-5101, USA, [6]Veterans Affairs Palo Alto Health Care System, Palo Alto, CA, 94304, USA, [7]Department of Genetics, Stanford University School of Medicine, Stanford, CA, 94305-5120, USA, [8]Department of Population Health and Pathobiology, College of Veterinary Medicine, North Carolina State University, Raleigh, NC, 27607, USA, [9]Department of Bioengineering, Stanford University, Stanford, CA, 94305, USA, [10]Cancer Early Detection Advanced Research Center, Knight Cancer Institute and Department of Molecular and Medical Genetics, Oregon Health Sciences University, Portland, OR, 97239-3098, USA, [11]Department of Neurology and Neurological Sciences, Stanford University School of Medicine, Stanford, CA, 94304, USA, [12]Department of Microbiology and Immunology, Stanford University, Stanford, CA, 94305-5124, USA, [13]Institute for Immunity, Transplantation and Infection, Human Immune Monitoring Center Stanford, CA, 94305, USA, [14]Division of Neonatology, Department of Pediatrics, Stanford University School of Medicine, Stanford, CA, 94304, USA, [15]Institute for Computational Health Sciences, University of California San Francisco, San Francisco, CA, 94158, USA, [16]Department of Pediatrics, University of California San Francisco, San Francisco, CA, 94143, USA, [17]Department of Obstetrics and Gynecology, Stanford University School of Medicine, Stanford, CA, 94304, USA, [18]Departments of Biomedical Data Sciences and Statistics, Stanford University, Stanford, CA, 94305-5464, USA and [19]Department of Statistics, Stanford University School of Medicine, Stanford, CA, 94305-4020, USA

*To whom correspondence should be addressed.

Abstract

Motivation: Multiple biological clocks govern a healthy pregnancy. These biological mechanisms produce immunologic, metabolomic, proteomic, genomic and microbiomic adaptations during the course of pregnancy. Modeling the chronology of these adaptations during full-term pregnancy provides the frameworks for future studies examining deviations implicated in pregnancy-related pathologies including preterm birth and preeclampsia.

Results: We performed a multiomics analysis of 51 samples from 17 pregnant women, delivering at term. The datasets included measurements from the immunome, transcriptome, microbiome, proteome and metabolome of samples obtained simultaneously from the same patients. Multivariate predictive modeling using the Elastic Net (EN) algorithm was used to measure the ability of each dataset to predict gestational age. Using stacked generalization, these datasets were combined into a single model. This model not only significantly increased predictive power by combining all datasets, but also revealed novel interactions between different biological modalities. Future work includes expansion of the cohort to preterm-enriched populations and in vivo analysis of immune-modulating interventions based on the mechanisms identified.

Availability and implementation: Datasets and scripts for reproduction of results are available through: https://nalab.stanford.edu/multiomics-pregnancy/.

Contact: naghaeep@stanford.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Physiological changes during pregnancy are highly dynamic and involve coordinated changes among multiple interconnected molecular and cellular systems from the fetus, the fetal-membrane and the mother (Diemert and Arck, 2018; Menon *et al.*, 2016). The simultaneous interrogation of these systems can reveal otherwise unrecognized crosstalk. Understanding such crosstalk can inform several lines of investigation. From a biological perspective, it can point to important disease mechanisms such as immune programming by the microbiome, or specific interactions between proteins and cellular elements (Aghaeepour *et al.*, 2017; Dethlefsen *et al.*, 2007). From a diagnostic perspective, it can reveal biomarkers from several biological domains that provide higher predictive power if combined. Alternatively, it can point to alternative biomarkers in an accessible biological compartment, which can replace biomarkers that are difficult to obtain or expensive to measure.

Recent technological advances in science provide novel opportunities to unravel the complex biology of pregnancy. A particularly pressing issue is to identify the biological pathways and the converging pathological processes that lead to preterm birth (Lackritz *et al.*, 2013). Preterm birth is the major cause of neonatal death, and the second leading cause of mortality in children under the age of 5 years (Liu *et al.*, 2012). An ongoing cohort study by the March of Dimes Prematurity Research Center at Stanford University exploits recent technological advances to examine an array of biological, demographic, clinical and environmental factors associated with normal and pathological pregnancies (Stevenson *et al.*, 2013; Shaw *et al.*, 2018; Wise *et al.*, 2017). From a biological perspective, this effort has so far produced two major lines of evidence. One line sheds light onto precisely tuned chronological changes that occur during normal pregnancy. For example, a highly multiplexed cell-based assay in whole blood revealed an 'immunological clock' of

human pregnancy that predicts gestational age at the time of sampling (Aghaeepour *et al.*, 2017). Similar results were reported in a longitudinal analysis of cell-free, maternal RNA (Pan *et al.*, 2017) and plasma proteins (Aghaeepour *et al.*, 2018). The primary objective of using gestational age as the clinical outcome in these studies is to extract molecular features that best capture normal chronological changes over the course of term pregnancy. Such knowledge will elucidate molecular deviations that are associated with pregnancy-related pathologies. The second line of this work points to important pathophysiological derangements. For example, dense longitudinal sampling of the vaginal microbiome revealed community composition profiles associated with preterm birth that were validated in an independent cohort (Callahan *et al.*, 2017; DiGiulio *et al.*, 2015). However, the important work of bringing these data modalities together has remained unexplored.

From a bioinformatics point of view, current multiomics efforts belong to two categories generally known as multi-staged and meta-dimensional (Ritchie *et al.*, 2015; Rohart *et al.*, 2017). In multi-staged analyses, measurements of the same biological factors (e.g. genes) are integrated at various biological levels and using different technological platforms (e.g. DNA and RNA sequencing, epigenetic analysis and proteomics assays—notable examples include (Emilsson *et al.*, 2008; Maynard *et al.*, 2008; Schadt *et al.*, 2005; Shabalin, 2012; Shen *et al.*, 2009)). However, recent biological studies extend well beyond just measurements of the same gene/protein and include various assays that cannot be mapped to a single gene. These include single cell analysis (Aghaeepour *et al.*, 2017), imaging (Woodward *et al.*, 2006), profile of metabolic profiling (Piening *et al.*, 2018), actigraphy using wearable sensors (Halilaj *et al.*, 2018) and clinical phenotypes (Ferrero *et al.*, 2016). Meta-dimensional multiomics approaches are now emerging that aim to combine heterogeneous datasets to identify key factors at various biological

levels, their interactions with each other, and with clinical outcomes. Some studies achieve this by simply merging all available datasets into a single matrix for joint modeling (Fridley *et al.*, 2012; Holzinger *et al.*, 2014; Mankoo *et al.*, 2011). These approaches are often susceptible to biases introduced by the differential sizes, modularities, scalings and batch effects of the included datasets. Various kernel (e.g. Borgwardt *et al.*, 2005) and graph (e.g. Kim *et al.*, 2012) transformations as well as latent space projections (Singh *et al.*, 2016) have been proposed to address these biases. In settings where analysis is performed against an external factor, an alternative is to use mixture-of-experts methods to combine the results of independent models produced using each dataset through various algorithms ranging from voting (e.g. Aghaeepour and Hoos, 2013) to integration of posterior Bayesian probabilities (Akavia *et al.*, 2010; Zhu *et al.*, 2008, 2012).

The main objective of this study was to test multiple strategies for integrating transcriptomic, immunological, microbiomic, metabolomic and proteomic datasets into different statistical models predicting gestational age in term pregnancy and identify the most accurate strategy. A final objective was to interrogate the derived model for novel and testable biological hypothesis.

2 Materials and methods

2.1 Study design

Pregnant women presenting to the obstetrics clinics of the Lucile Packard Children's Hospital at Stanford University for prenatal care were invited to participate in a cohort study to prospectively examine environmental and biological factors associated with normal and pathological pregnancies. Women were eligible if they were at least 18 years of age and in their first trimester of a singleton pregnancy. In 17 women, three samples were collected during pregnancy and a fourth one after deliver. The time points were chosen such that a peripheral blood sample (CyTOF analysis), a plasma sample (proteomic, cell-free transcriptomics, metabolomics analyses), a serum sample (luminex analyses) and a series of culture swabs (microbiome analysis) were simultaneously collected from each woman during the first (7–14 weeks), second (15–20 weeks) and third (24–32 weeks) trimester of pregnancy and 6-week postpartum. Repeated sampling during pregnancy allowed assessing important biological adaptations occurring continuously from the early phases of fetal development (first trimester) to the late phases of gestation (third trimester). The sample collected 6-week postpartum allowed for the assessment of the biological variables after the delivery of the fetus, a surrogate for the non-pregnant state which is not accessible in a prospective study of pregnant women.

2.2 Gestational age estimation

Gestational age was determined by best obstetrical estimate as recommended by the American College of Obstetricians and Gynecologists (Hershey, 2014).

2.3 Biological assays

Plasma and serum samples were assayed using the Luminex platform for cytokine levels. In addition, plasma samples were used for proteomics analysis, LC-MS metabolomics analysis, and cell-free transcriptomic analysis. Whole blood samples were analyzed using mass cytometry for single-cell characterization of the immune system. Finally, vaginal swabs, stool, saliva and tooth/gum samples were used for microbiomic profiling. See Supplementary Material for more detailed description of the assays. All timepoints of a given patient were analyzed simultaneously by all omics platforms to minimize systematic technical confounders (Supplementary Fig. S4).

2.4 Multivariate modeling

For a matrix $\mathbf{X}$ of all features from a given dataset, and a vector of estimated gestational ages at the time of each sampling, $\mathbf{Y}$, the EN algorithm calculates coefficients β to minimize the error term $L(\beta) = ||\mathbf{Y} - \mathbf{X}\beta||^2$. An L_1 regularization (Tibshirani, 1996) was used to increase model sparsity (which facilitates biological interpretation and validation). However, this approach is not ideal for the analysis of the highly interrelated biological datasets, because it only selects representatives of communities of highly correlated features. As a result, features correlated to these selected representatives are disregarded, despite the fact that they could be biologically relevant. This limitation is addressed by using an additional L_2 regularization penalty: $L(\alpha, \lambda, \beta) = ||\mathbf{Y} - \mathbf{X}\beta||_2 + \lambda[(1 - \alpha)||\beta||_2 + \alpha||\beta||_1]$, where $||\beta||_2 = \beta^\top \beta$ and $||\beta||_1 = \sum_{i=1}^{n} |\beta_i|$. The subset selecting factor λ controls the sparsity of the model and the smoothing factor α controls the smoothing of selection from correlated variables (Zou and Hastie, 2005).

2.5 Stack generalization

In the computer science literature, stacked generalization refers to the practice of combining several weak predictors for increased predictive power (Breiman, 1996; Sharkey, 1996; Wolpert, 1992). In life sciences, this often translates to analysis of a single dataset using multiple algorithms and then combining the results in a final multivariate modeling step (Ge and Wong, 2008; He *et al.*, 2013; Larranaga *et al.*, 2006; Yang *et al.*, 2010; Wang *et al.*, 2006). Here we expand this concept to multiomics analysis where a single multivariate analysis algorithm (EN) is used on a cohort of patients, and the variable factor is the biological assays used for developing the datasets. First, an EN model is constructed on each dataset from the same subjects. Then, all estimations of gestational age at time of sampling are used as features for a final EN model. This, essentially, is a weighted average of the individual models where the weights are the coefficients of the EN model.

2.6 Cross-validation

An underlying assumption of the EN algorithm is statistical independence between all observations. In this analysis, while the subjects are independent, the samples collected from various trimesters of the same subject are not. To account for this, we designed a leave-one-subject-out cross-validation strategy. In this setting, a model is trained on all available samples except for the three trimesters of a given subject. The model is then tested on all samples of the subject that it was blinded to. This process is repeated for all subjects until a blinded prediction has been produced for all samples. Final results are reported using these blinded predictions. This ensures complete independence from any intra-subject correlations.

A two-layer cross-validation strategy was implemented for simultaneous free-parameter optimization and analysis of the generalizability of the results (Fig. 2A). The inner layer selects the best values of α and λ (see Supplementary Fig. S1). The outer layer ensures that performance is reported on subjects that the models were blinded to during training.

A similar strategy was used for the stacked generalization step. Cross-validation folds where synchronized between the individual models from each dataset and the integrated model to leave out the same set of data points at all levels of the analysis. Importantly, this guarantees that not only the stacked generalization model, but also

its input features (i.e. the final predictions from each dataset) were blinded to the same subject during cross-validation.

2.7 Empirical evaluation

The procedure described above was empirically compared against a number of standard multivariate algorithms. The same algorithms were used for the individual datasets as well stacked generalization (Fig. 5). The algorithms included Random Forest (Breiman, 2001), Gaussian Process (Williams and Barber, 1998), Support Vector Regression (Chang and Lin, 2011; Hsu and Lin, 2002) and XGboost (Chen and Guestrin, 2016). The algorithms were compared using the default implementations provided in the following packages: (Chen and He, 2015; Karatzoglou *et al.*, 2004; Liaw and Wiener, 2002). All algorithms were evaluated using the same two-layer leave-one-patient-out CV strategy. The cross-validated parameter space for Gaussian process and Support Vector Regression included all available kernels [as described in (Karatzoglou *et al.*, (2004)] and initial noise variance between 0.001 and 10 000. EN predominantly outperforms the other methods on most datasets, followed by support vector regression. XGboost outperforms the other algorithms on the microbiome dataset.

2.8 Model reduction

A bootstrapping procedure was used to reduce the number of features used in each model. As described in Aghaeepour *et al.* (2017), one hundred bootstrap iterations were performed on each dataset where 57 samples were drawn randomly and with replacement. Piece-wise regression between the number of features (calculated by applying a range of thresholds to the mean coefficient of each measurement across all bootstrap iterations) and the final results of the models were used to select the number of features for each modality (Oosterbaan, 1994).

2.9 Correlation network

The features from the reduced models were visualized using a graph structure. Each feature was represented by a node. The correlation structure between the features was extracted using a Minimum Spanning Tree (MST) where the width of the edges were proportional to the spearman P-value of the correlation between the two nodes, on a $\log_{1}0$ scale. The graph was visualized using the Fruchterman-Reingolds layout (Fruchterman and Reingold, 1991).

2.10 *P*-value adjustment

All P-values were adjusted using Bonferroni's method ($adjusted\text{-}P\text{-}value = \min\{1, raw\text{-}P\text{-}value \times n\}$), where n is the number of features (Dunn, 1961).

2.11 Missing value interpolation

Missing values for all datasets were interpolated using a non-parametric multivariate model based on random forests. A model was trained for each feature of each dataset, and was subsequently used to estimate the missing values as described in Stekhoven and Bühlmann (2012).

3 Results

3.1 Modularity and size

Samples from 17 women for a total of 51 timepoints throughout pregnancy and 6 weeks postpartum were collected. Samples were analyzed for seven biological modalities: cell-free transcriptomics,
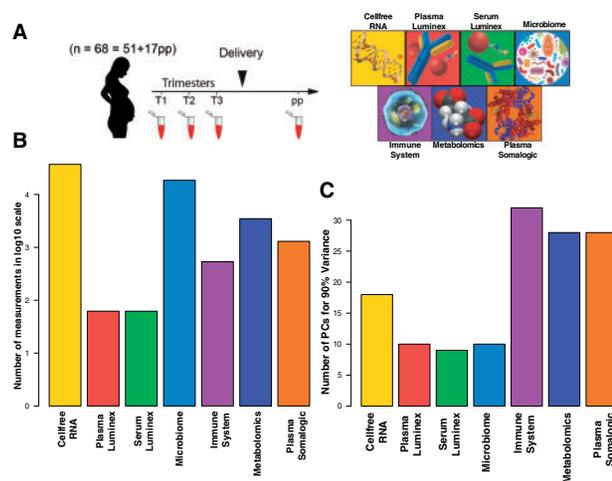

Fig. 1. (**A**) Overview of the study design. A total of 357 samples from 51 visits by 17 women were collected during three trimesters of pregnancy, as well as an additional 17 samples 6 weeks after delivery. Seven datasets were produced for each visit by each subject. (**B**) Data from each time point of each subject were analyzed using seven high-throughput assays, which produced different number of measurements. (**C**) The seven datasets had a range of correlations among the measured features. The internal correlation between features from each dataset was quantified using the number of Principle Components (PCs) needed to capture 90% variance (datasets in which most features are highly correlated would need fewer principal components)

antibody-based cytokine measurements in plasma and serum, microbiomic analyses (of vaginal swabs, stool, saliva and tooth/gum), mass cytometric analyses of whole blood, untargeted metabolomics and targeted proteomics analysis of plasma. These datasets produced different levels of modularity (as measured by the number of principal components needed to account for 90% variance of each dataset—Fig. 1C). The modularity of the datasets (Fig. 1C) was not correlated with the number of measurements available (Fig. 1B).

3.2 Per-dataset analysis

An Elastic Net (EN) model was developed to predict the gestational age of pregnancy of each subject at each visit. A two layer Cross-Validation (CV) procedure was used to both optimize the free parameters of the EN model (see Supplementary Fig. S1) and to ensure that predictions were made on samples that were not used for training model coefficients (see Fig. 2A and Section 2). Supplementary Figure S2 visualizes the predictions on the test samples for each modality versus the clinical estimations of gestational age. P-values of correlation with gestational age at time of sampling for the training and testing procedures are presented in Figure 2B and C, respectively. Plasma proteomics analysis using the SomaLogic platform produced the strongest predictive power (Fig. 2B and Supplementary Fig. S6). Results remained generally consistent between training and test sets (Fig. 2C). The datasets with a higher degree of independence between features (Fig. 1C) had a higher predictive power regardless of their size.

Due to the absence of true pre-pregnancy samples, we applied these models to postpartum samples collected 6 weeks postpartum as a surrogate for a non-pregnant state. At that time, some models (e.g. the immunologic and metabolomic models) recovered towards a state similar to a non-pregnant state, while others more closely reflected an early pregnant state or remain stable after delivery. This finding indicates that not all biological factors involved in pregnancy recover at a similar rates (Fig. 2D).
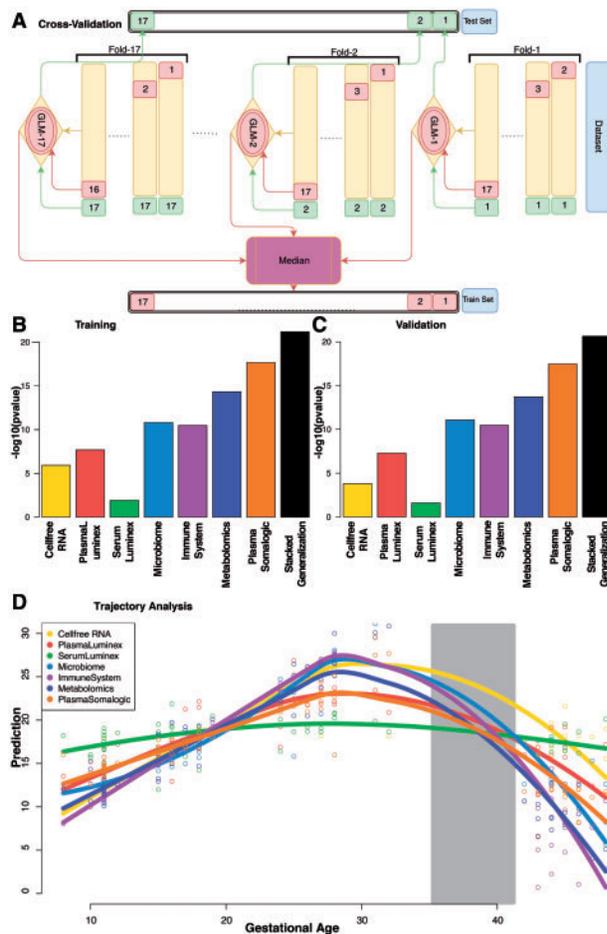
Fig. 2. (**A**) Overview of the two-layer CV procedure. On the outer layer, a modified leave-one-out procedure is used in which all samples from the same subject (as opposed to just one sample) are left out as a blinded dataset. Within each fold, a second CV procedure is performed to optimize the free parameters of the EN model. Test samples for the inner and outer layers are visualized in red and green, respectively. The final training prediction is the median of predictions from all models that included that patient during their training (bottom), and the final blinded test set prediction comes from the only model that was blinded to it (top). See Section 2 for details. (**B**) and (**C**) The Spearman correlation *P*-values of the (B) training set and (C) test set results of the CV procedure for each dataset. (**D**) The models for each dataset applied to all samples including the postpartum visit 6 weeks after delivery. The average trend for each platform is visualized using kernel density estimation for smoothing. The delivery range is highlighted in gray. Some models quickly recover towards a non-pregnant status (below the first trimester) while others remain stable after delivery

3.3 Stacked generalization
A stacked generalization strategy was used to combine the predictive powers of the different omics datasets as described in Wolpert (1992). As illustrated in Figure 3A, an EN model was first trained on each dataset. Then, the estimations of gestational age produced by the seven independent models were merged using an additional EN model. Cross-validation was synchronized across all layers to ensure predictions were made on samples that had not been used for optimizing model coefficients. The free parameters of the models, as calculated using the inner CV procedure (see Section 2), are visualized in Supplementary Figure S1.

Ablation analysis, a procedure for investigating the path of dataset weights by iteratively retraining the stacked generalization model, was used to measure the relative contribution of each dataset

to the final predictions (Fawcett and Hoos, 2016). This procedure was performed by iteratively removing the most important dataset from the mix (Fig. 4A). Importantly, for each iteration, the algorithm was able to recalculate new weights for the remaining datasets to partially compensate for any lost information. For example, after removal of the proteomic and metabolomic datasets, the algorithm significantly increased the weight of the predictions based on the immune system to compensate for the two removed datasets. Similar analysis in reverse order (Fig. 4B) revealed a minimal decrease in the predictive power when the most important dataset was preserved.

To enable biological exploration, the top hits from each model were extracted using a bootstrapping strategy for sensitivity analysis (see Section 2 for details) and visualized using a minimum spanning tree of Spearman correlations between the selected features on a Fruchterman-Reingold layout (Fruchterman and Reingold, 1991), in Figure 3B and C, respectively. This resulted in a set of 226 interrelated features (Supplementary Table S1), revealing statistically robust interactions within and between each omics dataset. A Minimum Spanning Tree (MST) representation organized these interactions into a branched structure in which the distance between two features is proportional to the strength of the correlation between them. Metabolomics, transcriptomics and proteomics features primarily segregated into three clusters (Fig. 3C). Cell-based features from the immune system were distributed across the MST graph, forming a link between other omics datasets rather than being confined to a single cluster. The MST graph highlighted the connectivity between biological processes measured in the plasma (metabolomic, transcriptomic and proteomic measurements) or local compartments (microbiomic data) and cell-specific immune responses measured in the peripheral blood compartment.

3.4 Biological hypothesis generation
Several biologically plausible and hypothesis generating correlations between omics datasets emerged. Here, we highlight three of these data-driven hypotheses. In one instance, we illustrate how the integrative dataset can inform additional experiments that allow further exploration of the nature of observed interaction between different omics features.

With respect to the microbiomic data, a strong correlation was observed between changes in the composition of Neisseria bacterial species localized in the oral cavity as well as Bacteroides species in the gut and $TCR\gamma\delta^+$ T cells. This finding is consistent with the unique role of $TCR\gamma\delta^+$ T cells in mucosal immunity, particularly in the control of oral pathogens (Chien *et al.*, 2014; Moutsopoulos and Konkel, 2018; Wu *et al.*, 2014). Given increasing epidemiological evidence linking oral cavity dysbiosis and pregnancy-related complications, such as preterm labor and preeclampsia (Bassani *et al.*, 2007; Boggess *et al.*, 2003; Bošnjak *et al.*, 2006; Hajishengallis, 2015; Herrera *et al.*, 2007; Nabet *et al.*, 2010), our results raise the hypothesis that the correlation between the changes in oral bacterial species and $TCR\gamma\delta^+$ T cell frequencies may be disrupted in pathological pregnancies, such as preterm pregnancies.

With respect to the metabolomics dataset, the model revealed strong correlations between the plasma factor pregnanolone sulfate and the NF-κB signaling in myeloid dendritic cells (mDCs) and regulatory T cells (Tregs). Pregnanolone sulfate, or $3\alpha, 5\beta$-tetrahydroprogesterone($3\alpha, 5\beta$-THP), is an endogenous steroid biosynthesized from progesterone. Modulation of immune cell function by progesterone and its derivative is well established (Druckmann and Druckmann, 2005). However, their roles in regulating the function of specific immune cell subsets during pregnancy are not fully
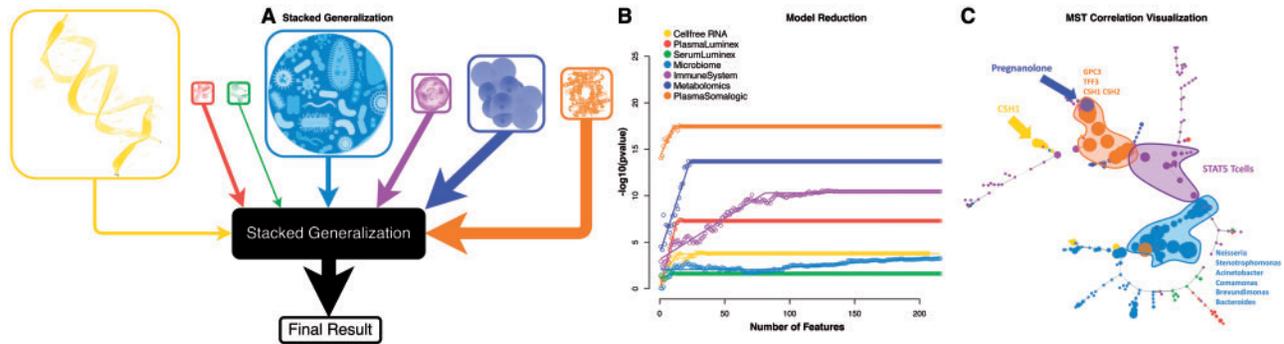
Fig. 3. (**A**) Stacked generalization analysis. The size of the boxes is proportional to the $\log_{10}$ of the number of measurements in each dataset. The thickness of the arrow is proportional to the $-\log_{10}$ of *P*-value of a correlation test for gestational age; (**B**) The number of model components (x-axis) versus the *P*-value of the Spearman correlation between each model and gestational age (y-axis). Lines represent the piece-wise regression fit for calculation of the number of features. (**C**) Visualization of the most predictive features in a correlation network. The size of each node is proportional to the univariate correlation between that feature and gestational age. Color represents the corresponding dataset
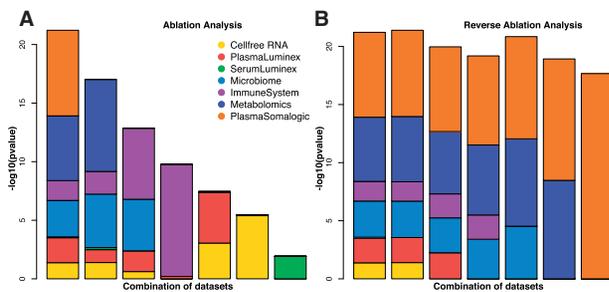
Fig. 4. Ablation analysis to measure the collective predictive power of the model after removal of each dataset. At each iteration, the most (**A**) or least (**B**) important datasets were removed from stacked generalization. Color is proportional to the coefficients of the stacked generalization model. At each iteration, the algorithm was able to readjust the coefficients. This demonstrated that the algorithm could effectively use the remaining datasets to compensate for the latest removals
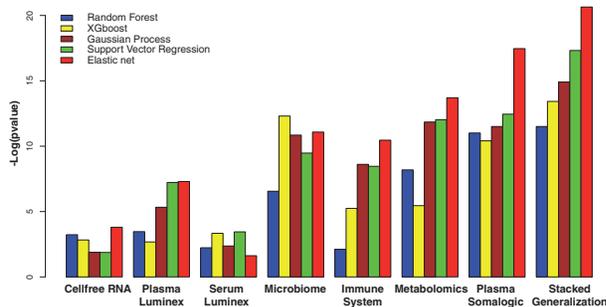
Fig. 5. Empirical evaluation of elastic-net, random forest, XGboost, Gaussian Process and Support Vector Regression on each dataset, and the combination of all datasets. The hyper parameters of each method were tuned by the same two-layer leave-one-patient-out CV procedure for the prediction of gestational age on the test set. EN predominantly outperformed the other methods on most datasets, followed by support vector regression. XGboost outperformed the other algorithms on the microbiome dataset

understood. The results thus generated a novel hypothesis that pregnanolone sulfate may regulate important aspects of mDC and Treg functions during pregnancy.

With respect to the proteomic dataset, a three-way interaction between the transcriptomic, proteomic and cytomic datasets was particularly interesting, as it highlighted a novel connection between previously reported models of molecular clocks of pregnancy. This

interaction contained the Chorionic Somatomammotropin Hormone-1 (CSH-1), represented at the transcript (cell-free RNA dataset) and protein (Somalogic dataset) levels, and the endogenous activity of the transcription factor STAT5 measured at the single-cell level in CD4+ and CD8+ T cell subsets. CSH-1 is known to bind to the prolactin receptor (Walsh and Kossiakoff, 2006), which signals through the JAK2/STAT5 signaling pathway (Gouilleux *et al.*, 1994). As such, results from the integrative analysis informs a novel hypothesis that CSH-1 may directly activate the JAK2/STAT5 signaling pathway in $CD4^+$ and $CD8^+$ T cell subsets during pregnancy.

The strong correlation observed between CSH-1 RNA and protein levels, and STAT5 activity in T cells ($R = 0.59$, $P = 4.40 \times 10^{-06}$) prompted further examination of this hypothesis in an in vitro model to determine whether CSH-1 can directly activate the JAK2/STAT5 signaling pathway in T cells. However, incubation of whole blood samples from non-pregnant or pregnant (Supplementary Fig. S3) women with CSH-1 did not induce the phosphorylation of STAT5 in CD4+ or CD8+ T cell subsets. On further inspection of the proteomic dataset, CSH-1 was found to belong to a community of tightly correlated plasma factors known to regulate the JAK/STAT signaling pathway. This community included the inflammatory cytokine Interleukin-2. Supplementary Figure S3 shows that, in contrast to CSH-1 or prolactin, incubation of whole blood samples with IL-2 induced a robust STAT5 phosphorylation signal in all major T cell subsets. These results suggested that in the context of pregnancy, the progressive increase in intracellular STAT5 activity in T cell subsets is likely driven by changes in IL-2 rather than CSH-1.

4 Discussion

We have described an analysis of seven high-throughput biological modalities during term pregnancy. An agnostic machine learning approach was used to evaluate the predictive power of each dataset for estimation of gestational age using biological signals. An additional machine learning layer was used to combine these estimations to further increase predictive power. Importantly, these datasets differed in both size and modularity. By taking this two layer approach, we prevented higher-dimensional datasets from overwhelming the final model. This both increased predictive power and facilitated biological interpretation.

Using this approach, we estimated the gestational age of the fetus at the time of each sampling. The stacked generalization algorithm produced models more accurate than models derived from any individual dataset. Ablation analysis (Fawcett and Hoos, 2016) was used to study the impact of each dataset on the final predictions. Importantly, this analysis showed that by retraining the stacked generalization model, other datasets could partially compensate for the removal of a given dataset. Using sensitivity analysis and piece-wise regression and sequential feature-reduction, each model was reduced to a limited number of required measurements. These were then used for correlation analysis, visualization and biological interpretation. These two complementary model reduction procedures lay the foundation for objective analysis to strike a balance between predictive-power and assay/sampling costs in resource-poor settings (e.g. a more expensive assay which requires a larger sample size from a complex biopsy may be replaceable by two cheaper and more feasible assays).

The study provided an integrated biological model of maternal changes during pregnancy, highlighting the interconnectivity of multiple biological systems. Notably, strong correlations between metabolomic, proteomic, transcriptomic features and specific immune cell signaling responses pointed at biologically plausible interactions. For example, the model identified a strong relationship between the steroid hormone pregnanolone sulfate and the signaling behavior of mDCs and Tregs. mDCs and Tregs play a critical role in feto-maternal tolerance and the maintenance of pregnancy (Aluvihare *et al.*, 2004; Erlebacher, 2013). Our data provide the basis for a novel hypothesis that pregnanolone sulfate plays a role in regulation of the function of these two cell types during pregnancy. Alternatively, recent evidence indicating that T cells can produce pregnenolone, the precursor of pregnanolone sulfate (Mahata *et al.*, 2014), suggests that immune cells may be a cellular source of pregnanolone sulfate production, providing another hypothesis for the observed correlations.

The study also shows that the biological interpretation of observed interactions between two model components benefits from exploring the communities of features that strongly correlate with these model components. As such, the integrative model revealed a strong interaction between the protein factor CSH-1 and STAT5 activity in CD4+ T cells. However, a community of protein factors correlating with CSH-1 contained the cytokine IL-2, a canonical activator of the JAK/STAT5 signaling pathway in CD4+ T cells (Mahmud *et al.*, 2013). Together with our in vitro data showing that stimulation with IL-2, but not with CSH-1, results in STAT5 phosphorylation in CD4+ T cells, these findings suggest that the interaction between CSH-1 and STAT5 activity in CD4+ T cells is likely indirectly mediated by IL-2. For example, activation of the PRL/CSH-1 receptor in cells other than T lymphocytes has been shown to promote the transcription of IL-2 (Sun *et al.*, 2004). CSH-1 may thus be implicated in the paracrine regulation of T cell function through positive regulation of IL-2 gene expression in other immune or non-immune cell types. When applied to postpartum samples collected 6 weeks after delivery, these models demonstrated that different biological modalities return to a non-pregnant state at different rates, reflecting synchronized pacemakers (Diemert and Arck, 2018). This finding motivates detailed biological analysis of the role of the inter-pregnancy interval (Girsen *et al.*, 2018) and history of preterm birth in adverse outcomes (Gaudillière *et al.*, 2015).

Selecting the hyperparameters of an EN model is largely a balancing act between sparsity and accuracy. In complex biological datasets, this is often confounded by the intrinsic characteristics of data including size and modularity (Waldmann *et al.*, 2013). To address this, a two-step CV procedure was used in this analysis. The inner layer enables optimization for the free-parameters of the EN model using an exhaustive grid search (Supplementary Fig. S1). The outer layer ensures the generalizability of the results to previously unseen samples. To increase sample size, each sample extracted at a trimester from a single subject was treated as an independent data point. To ensure the models were not biased by the dependency between samples donated by the same subject, all three trimesters of a given subject were excluded together in the same CV fold. Therefore, reported results are based on models that had access to no samples from a subject in the test-set. The samples used for testing purposes in all CV steps were synchronized across all models. Therefore, all test-set results (including those of the stacked generalization models) are reported only on samples that were blinded in all previous analyses.

This study has several limitations that have inspired our future plans. First, the number of subjects in this 'proof-of-concept' cohort was small relative to the number of measurements. In addition, recruitment from a single-care center limited the diversity of the dataset. Despite this, we were able to capture the chronology of biological changes during pregnancy. This correlation was not driven by age, BMI, or parity (partial correlation test $P > 0.05$). However, given the racial disparities in pregnancy outcomes, replicating this analysis in more diverse cohorts is crucial. The March of Dimes Prematurity Center at Stanford University has already engaged in several international collaborations to directly address this. Similarly, the number of measurements was significantly larger than the cohort size, which increased the possibility of false positives. In addition to carefully designed cross-validation, feature reduction and clustering (e.g. Bien and Tibshirani, 2011) can be used to improve the predictive power of multivariate models in high-dimensional settings and enable exploration of more interactions between different datasets. These various approaches should be tested in an unbiased and collaborative setting (e.g. Aghaeepour *et al.*, 2016; Stolovitzky *et al.*, 2007) as large multiomics datasets become available. Finally, the current dataset included only one sample per trimester, and these samples were treated as independent datapoints. In the future, high-resolution sampling together with mixed effect models (Gałecki and Burzykowski, 2013) will combine the information content of different timepoints to produce increasingly more accurate prediction of pregnancy related events using serial sampling throughout pregnancy.

In summary, our study revealed a chronology of biologically-diverse events over the course of pregnancy. Our findings were enabled using seven high-throughput longitudinal biological assays of the same patient cohort. The computational pipeline introduced in this article can increase predictive power by combining datasets of various sizes and modularities in a balanced way. We expect this pipeline to be applicable to a wide range of studies beyond the field of pregnancy. Similarly, the dataset produced here provides a unique resource for future biological investigations. Particularly, this study can be used as a resource to identify correlates of any other features from one of the seven datasets that may be identified in future studies. Finally, by characterizing the biological chronology of normal pregnancy, this study provides the conceptual and analytical framework to analyze the complex interplays between various biological modalities that govern preterm birth and other pregnancy-related pathologies.

Acknowledgements

Murray, Trevor Hastie, Ryan R. Brinkman and Holger H. Hoos for their feedback and inspiration.

References

Aghaeepour,N. and Hoos,H.H. (2013) Ensemble-based prediction of RNA secondary structures. *BMC Bioinformatics*, 14, 139.

Aghaeepour,N. *et al.* (2016) A benchmark for evaluation of algorithms for identification of cellular correlates of clinical outcomes. *Cytometry Part A*, 89, 16–21.

Aghaeepour,N. *et al.* (2017) An immune clock of human pregnancy. *Sci. Immunol.*, 2, eaan2946.

Aghaeepour,N. *et al.* (2018) A proteomic clock of human pregnancy. *Am. J. Obstetr. Gynecol.*, 218, 347.e1–347.e14.

Akavia,U.D. *et al.* (2010) An integrated approach to uncover drivers of cancer. *Cell*, 143, 1005–1017.

Aluvihare,V.R. *et al.* (2004) Regulatory t cells mediate maternal tolerance to the fetus. *Nat. Immunol.*, 5, 266–271.

Bassani,D. *et al.* (2007) Periodontal disease and perinatal outcomes: a case-control study. *J. Clin. Periodontol.*, 34, 31–39.

Bien,J. and Tibshirani,R. (2011) Hierarchical clustering with prototypes via minimax linkage. *J. Am. Stat. Assoc.*, 106, 1075–1084.

Boggess,K.A. *et al.* (2003) Maternal periodontal disease is associated with an increased risk for preeclampsia. *Obstetr. Gynecol.*, 101, 227–231.

Borgwardt,K.M. *et al.* (2005) Protein function prediction via graph kernels. *Bioinformatics*, 21, i47–i56.

Bošnjak,A. *et al.* (2006) Pre-term delivery and periodontal disease: a case–control study from croatia. *J. Clin. Periodontol.*, 33, 710–716.

Breiman,L. (1996) Stacked regressions. *Mach. Learn.*, 24, 49–64.

Breiman,L. (2001) Random Forests. *Mach. Learn.*, 45, 5–32.

Callahan,B.J. *et al.* (2017) Replication and refinement of a vaginal microbial signature of preterm birth in two racially distinct cohorts of us women. *Proc. Natl. Acad. Sci. USA*, 114, 9966–9971.

Chang,C.-C. and Lin,C.-J. (2011) Libsvm: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2, 1–27: 27. 27: 1–May ISSN 2157-6904.

Chen,T. and Guestrin,C. (2016) Xgboost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, pp. 785–794.

Chen,T. and He,T. (2015) Xgboost: extreme gradient boosting. *R Package Version 0.4-2*.

Chien,Y-h. *et al.* (2014) $\gamma\delta$ t cells: first line of defense and beyond. *Annu. Rev. Immunol.*, 32, 121–155.

Dethlefsen,L. *et al.* (2007) An ecological and evolutionary perspective on human–microbe mutualism and disease. *Nature*, 449, 811–818.

Diemert,A. and Arck,P.C. (2018) Pregnancy around the clock. *Trends Mol. Med.*, 24, 1–3.

DiGiulio,D.B. *et al.* (2015) Temporal and spatial variation of the human microbiota during pregnancy. *Proc. Natl. Acad. Sci. USA*, 112, 11060–11065.

Druckmann,R. and Druckmann,M.-A. (2005) Progesterone and the immunology of pregnancy. *J. Steroid Biochem. Mol. Biol.*, 97, 389–396.

Dunn,O.J. (1961) Multiple comparisons among means. *J. Am. Stat. Assoc.*, 56, 52–64.

Emilsson,V. *et al.* (2008) Genetics of gene expression and its effect on disease. *Nature*, 452, 423–428.

Erlebacher,A. (2013) Immunology of the maternal-fetal interface. *Annu. Rev. Immunol.*, 31, 387–411.

Fawcett,C. and Hoos,H.H. (2016) Analysing differences between algorithm configurations through ablation. *J. Heuristics*, 22, 431–458.

Ferrero,D.M. *et al.* (2016) 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.

Fridley,B.L. *et al.* (2012) A bayesian integrative genomic model for pathway analysis of complex traits. *Genet. Epidemiol.*, 36, 352–359.

Fruchterman,T.M. and Reingold,E.M. (1991) Graph drawing by force-directed placement. *Softw. Pract. Exp.*, 21, 1129–1164.

Gałecki,A. and Burzykowski,T. (2013) *Linear Mixed-Effects Model*. Springer, New York, pp. 245–273.

Gaudillière,B. *et al.* (2015) Implementing mass cytometry at the bedside to study the immunological basis of human diseases: distinctive immune features in patients with a history of term or preterm birth. *Cytometry A*, 87, 817–829.

Ge,G. and Wong,G.W. (2008) Classification of premalignant pancreatic cancer mass-spectrometry data using decision tree ensembles. *BMC Bioinformatics*, 9, 275.

Girsen,A.I. *et al.* (2018) What factors are related to recurrent preterm birth among underweight women? *J. Maternal-Fetal Neonatal Med.*, 31, 560–566.

Gouilleux,F. *et al.* (1994) Prolactin induces phosphorylation of tyr694 of stat5 (mgf), a prerequisite for dna binding and induction of transcription. *EMBO J.*, 13, 4361–4369.

Hajishengallis,G. (2015) Periodontitis: from microbial immune subversion to systemic inflammation. *Nat. Rev. Immunol.*, 15, 30–44.

Halilaj,E. *et al.* (2018) Physical activity is associated with changes in knee cartilage microstructure. *Osteoarthritis Cartilage* 26, 770–774.

He,L. *et al.* (2013) Extracting drug-drug interaction from the biomedical literature using a stacked generalization-based approach. *Plos One*, 8, e65814–e65812.

Herrera,J.A. *et al.* (2007) Periodontal disease severity is related to high levels of c-reactive protein in pre-eclampsia. *J. Hypertension*, 25, 1459–1464.

Hershey,D.W. (2014) Fetal imaging: executive summary of a joint eunice kennedy shriver national institute of child health and human development, society for maternal-fetal medicine, american institute of ultrasound in medicine, american college of obstetricians and gynecologists, american college of radiology, society for pediatric radiology, and society of radiologists in ultrasound fetal imaging workshop. *J. Ultrasound Med.*, 124, 836.

Holzinger,E.R. *et al.* (2014) Athena: the analysis tool for heritable and environmental network associations. *Bioinformatics*, 30, 698–705.

Hsu,C.-W. and Lin,C.-J. (2002) A comparison of methods for multiclass support vector machines. *IEEE Trans. Neural Netw.*, 13, 415–425. Mar

Karatzoglou,A. *et al.* (2004) S4 package for kernel methods in R. *J. Stat. Softw.*, 11, 1–20.

Kim,D. *et al.* (2012) Synergistic effect of different levels of genomic data for cancer clinical outcome prediction. *J. Biomed. Inf.*, 45, 1191–1198.

Lackritz,E.M. *et al.* (2013) A solution pathway for preterm birth: accelerating a priority research agenda. *Lancet Global Health*, 1, e328–e330.

Larranaga,P. *et al.* (2006) Machine learning in bioinformatics. *Brief. Bioinf.*, 7, 86–112.

Liaw,A. and Wiener,M. (2002) Classification and regression by randomforest. *R. News*, 2, 18–22.

Liu,L. *et al.* (2012) Global, regional, and national causes of child mortality: an updated systematic analysis for 2010 with time trends since 2000. *Lancet*, 379, 98322151–98322161.

Mahata,B. *et al.* (2014) Single-cell rna sequencing reveals t helper cells synthesizing steroids de novo to contribute to immune homeostasis. *Cell Rep.*, 7, 1130–1142.

Mahmud,S.A. *et al.* (2013) Interleukin-2 and stat5 in regulatory t cell development and function. *JAK-STAT*, 2, e23154–e23156.

Mankoo,P.K. *et al.* (2011) Time to recurrence and survival in serous ovarian tumors predicted from integrated genomic profiles. *Plos One*, 6, e24709.

Maynard,N.D. *et al.* (2008) Genome-wide mapping of allele-specific protein-dna interactions in human cells. *Nat. Methods*, 5, 307–309.

Menon,R. *et al.* (2016) Novel concepts on pregnancy clocks and alarms: redundancy and synergy in human parturition. *Hum. Reprod. Update*, 22, 535–560.

Moutsopoulos,N.M. and Konkel,J.E. (2018) Tissue-specific immunity at the oral mucosal barrier. *Trends Immunol.*, 39, 276–287.

Nabet,C. *et al.* (2010) Maternal periodontitis and the causes of preterm birth: the case–control epipap study. *J. Clin. Periodontol.*, 37, 37–45.

Oosterbaan,R.J. (1994) Frequency and regression analysis. *Drainage Principles Appl.*, 16, 175–224.

Pan,W. *et al.* (2017) Simultaneously monitoring immune response and microbial infections during pregnancy through plasma cfrna sequencing. *Clin. Chem.*, 63, 1695–1704.

Piening,B. *et al.* (2018) Integrative personal omics profiles during periods of weight gain and loss. *Cell Syst.*, 6, 157–170.e8.

Ritchie,M.D. *et al.* (2015) Methods of integrating data to uncover genotype-phenotype interactions. *Nat. Rev. Genet.*, 16, 85–97.

Rohart,F. *et al.* (2017) mixOmics: an R package for 'omics feature selection and multiple data integration. *PLOS Comput. Biol.*, 13, 1–19.

Schadt,E.E. *et al.* (2005) An integrative genomics approach to infer causal associations between gene expression and disease. *Nat. Genet.*, 37, 710–717.

Shabalin,A.A. (2012) Matrix eqtl: ultra fast eqtl analysis via large matrix operations. *Bioinformatics*, 28, 1353–1358.

Sharkey,A.J.C. (1996) On combining artificial neural nets. *Connect. Sci.*, 8, 299–314.

Shaw,G.M. *et al.* (2018) Residential agricultural pesticide exposures and risks of preeclampsia. *Environ. Res.*, 164, 546–555.

Shen,R. *et al.* (2009) Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, 25, 2906–2912.

Singh,A. *et al.* (2016) Diablo-an integrative, multi-omics, multivariate method for multi-group classification. *bioRxiv*, 10.1101/067611.

Stekhoven,D.J. and Bühlmann,P. (2012) Missforest–non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28, 112–118.

Stevenson,D., M. of Dimes Prematurity Research Center at Stanford University School of Medicine. *et al.* (2013) Transdisciplinary translational science and the case of preterm birth. *J. Perinatol.*, 33, 251–258.

Stolovitzky,G. *et al.* (2007) Dialogue on reverse-engineering assessment and methods. *Ann. N.Y. Acad. Sci.*, 1115, 1–22.

Sun,R. *et al.* (2004) Expression of prolactin receptor and response to prolactin stimulation of human nk cell lines. *Cell Res.*, 14, 67–73.

Tibshirani,R. (1996) Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B (Methodological)*, 58, 267–288.

Waldmann,P. *et al.* (2013) Evaluation of the lasso and the elastic net in genome-wide association studies. *Front. Genet.*, 4, 270–211.

Walsh,S.T. and Kossiakoff,A.A. (2006) Crystal structure and site 1 binding energetics of human placental lactogen. *J. Mol. Biol.*, 358, 773–784.

Wang,S.-Q. *et al.* (2006) Using stacked generalization to predict membrane protein types based on pseudo-amino acid composition. *J. Theor. Biol.*, 242, 941–946.

Williams,C.K.I. and Barber,D. (1998) Bayesian classification with gaussian processes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20, 1342–1351. Dec

Wise,P.H. *et al.* (2017) Risky business: meeting the structural needs of transdisciplinary science. *J. Pediatrics*, 191, 255–258.

Wolpert,D.H. (1992) Stacked generalization. *Neural Netw.*, 5, 241–259.

Woodward,L.J. *et al.* (2006) Neonatal mri to predict neurodevelopmental outcomes in preterm infants. *N. Engl. J. Med.*, 355, 685–694.

Wu,R.-Q. *et al.* (2014) The mucosal immune system in the oral cavity-an orchestra of t cell diversity. *Int. J. Oral Sci.*, 6, 125–132.

Yang,P. *et al.* (2010) A review of ensemble methods in bioinformatics. *Current Bioinf.*, 5, 296–308.

Zhu,J. *et al.* (2008) Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nat. Genet.*, 40, 854–861.

Zhu,J. *et al.* (2012) Stitching together multiple data dimensions reveals interacting metabolomic and transcriptomic networks that modulate cell regulation. *PLOS Biol.*, 10, e1001301–e1001319.

Zou,H. and Hastie,T. (2005) Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)*, 67, 301–320.