Comprehensive Analysis of Volatile Biomarkers for Female Fertility

by

Stephanie Marie Ong

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved April 2018 by the
Graduate Supervisory Committee:

Barbara Smith, Chair
Heather Bean
Christopher Plaisier

ARIZONA STATE UNIVERSITY

May 2018

ABSTRACT

One out of ten women has a difficult time getting or staying pregnant in the United States. Recent studies have identified aging as one of the key factors attributed to a decline in female reproductive health. Existing fertility diagnostic methods do not allow for the non-invasive monitoring of hormone levels across time. In recent years, olfactory sensing has emerged as a promising diagnostic tool for its potential for real-time, non-invasive monitoring. This technology has been proven promising in the areas of oncology, diabetes, and neurological disorders. Little work, however, has addressed the use of olfactory sensing with respect to female fertility. In this work, we perform a study on ten healthy female subjects to determine the volatile signature in biological samples across 28 days, correlating to fertility hormones. Volatile organic compounds (VOCs) present in the air above the biological sample, or headspace, were collected by solid phase microextraction (SPME), using a 50/30 µm divinylbenzene/carboxen/ polydimethylsiloxane (DVB/CAR/PDMS) coated fiber. Samples were analyzed, using comprehensive two-dimensional gas chromatography-time-of-flight mass spectrometry (GC×GC-TOFMS). A regression model was used to identify key analytes, corresponding to the fertility hormones estrogen and progesterone. Results indicate shifts in volatile signatures in biological samples across the 28 days, relevant to hormonal changes. Further work includes evaluating metabolic changes in volatile hormone expression as an early indicator of declining fertility, so women may one day be able to monitor their reproductive health in real-time as they age.

DEDICATION

I dedicate this work to my family, especially my parents, Henry and Lina Ong, for their unwavering support.

ACKNOWLEDGMENTS

I would like to acknowledge Vi Nguyen for her support in gas chromatography-mass spectrometry (GC-MS) testing, and Devika Krishnamurthy for the collection of biological samples in this study. I would also like to acknowledge Dr. John Stufken and Abigael Nachtsheim for statistical support. A special thanks to Christopher Miranda for programming support and Trenton Davis for GC-MS and statistical support. I would also like to thank Jarrett Eshima for his help in running biological samples on the GC x GC-TOFMS. Lastly, I would like to recognize my graduate committee, Dr. Christopher Plaisier, Dr. Heather Bean, and especially my advisor, Dr. Barbara Smith, for their support and mentorship throughout my graduate studies.

TABLE OF CONTENTS

APPENDIX Page

LIST OF TABLES

Table Page

LIST OF FIGURES

CHAPTER 1

INTRODUCTION/BACKGROUND LITERATURE

In the United States, one out of ten women 1.6 million women (ages 15-44) has complications with fertility, affecting approximately 1.6 million individuals annually [1]. In 2012, the estimated market for individuals seeking fertility services reached $3.5 billion [2]. Fertility, or the natural ability to reproduce, directly correlates to hormone production. The reproductive health of a woman is indicated by hormones such as estrogen, progesterone, anti-mullerian hormone (AMH), follicle stimulating hormone (FSH), and luteinizing hormone (LH) [3-5]. During a woman's 28-day menstrual cycle, hormone levels naturally change. These hormone-related changes affect the release of an egg from an ovary, known as ovulation. During this process, fluctuations in hormone levels, which are indicative of ovulation, are present in metabolic shifts within biological fluids [6]. Current methods of testing ovulation include clinical and at-home diagnostics such as: blood tests, basal body temperature monitoring, and fertility kits [7-9], evaluated at a single time point. A woman's reproductive health, however, changes with time [10]. These available methods lack the ability to quantify metabolic alterations in hormones across years. Understanding the long-term ovulatory and physiological health of a women can give insight into trends in reproductive health over time. Thus, a critical need exists to measure reproductive hormones and their corresponding metabolites in real-time across a woman's reproductive years.

Metabolic profiling of biological samples has recently shown promise of developing into an accurate and real-time monitoring technique to diagnose hormone-related diseases, such as diabetes, cancer, and neurological disorders [11-15]. Hormones

are part of the reproductive metabolic pathway and are directly correlated to metabolites [16]. Metabolite evaluation in biological samples has therefore emerged as a promising way to indicate hormone levels in real-time due to being readily available in biological samples [16]. An example of this can be found in the hormone, estrogen, where metabolic by-products result when estrogen is broken down from the parent estrogens of estrone and estradiol. During this reaction, oxidation of estrone and estradiol occur at the C-2 or C-4 positions to produce catechol estrogen metabolites, which include 2-hydroxyestrone, 2-hydroxyestradiol, and 4-hydroxyestrone [16]. Likewise, progesterone can be reduced to 5α-pregnan-3, 20-dione and 3α-hydroxy-5α-pregnan-20-one metabolites [17]. Specific volatiles released from these metabolites can be traced back to the parent hormones. Volatiles consist of low molecular weight compounds, which are usually less than 400 g/mol [18] and readily become vapors or gases at room temperature. These vapors, or volatile organic compounds (VOCs), can be collected in the headspace of, or air above, biological samples [19]. Current studies show VOCs eluded from within different biological samples, including blood, saliva, and urine [20-23]. In addition, VOCs have been studied across a variety of fields, including toxicology, oncology, and neurology [13-15], for applications in personalized diagnostics. Research utilizing VOCs to monitor reproductive hormones in real-time, however, is limited. Investigations have shown gas chromatography-mass spectrometry (GC-MS) to be a reliable method of detecting VOCs at physiological levels. With high specificity at lower limits of detection [24], GC-MS can distinguish different VOCs within complex biological samples. To gain even better resolution, two-dimensional gas chromatography-time-of-flight mass spectrometry (GC x GC-TOFMS) differentiates structurally similar compounds by

preventing peaks from co-eluding [25]. Furthermore, solid-phase microextraction (SPME) reduces interference from higher molecular weight compounds, therefore, enabling volatile binding between the molecular weights of 40-275, promoting a wider range of volatile collection [25]. Thus, GC x GC-TOFMS combined with SPME provides a simple and rapid extraction technique to collect VOCs, and delivers better reproducibility, detection, and separation of VOCs from complex biological sample matrices [25].

In this study, we employed GC x GC-TOFMS with SPME to analyze the VOCs in the urine samples of ten healthy women across a 28-day cycle. The aim of this study was to evaluate metabolic shifts of VOCs as they relate to fertility hormones across a healthy woman's menstrual cycle. Results indicate a significant shift in VOCs correlating to the estrogen hormone during ovulation. Furthermore, VOCs from the urine samples consist of different functional groups, including alcohols, aldehydes, amides, amines, aromatics, carboxylic acids, ethers, hydrocarbons, ketones, and thiols. Further studies will need to investigate the potential of metabolic changes in volatile hormone expression for use as an early indicator of declining fertility. Through this work, women may one day be able to monitor their reproductive health in real-time as they age.

CHAPTER 2

METHODOLOGY

2.1 Sample Collection and Preparation

The sample collection protocol was approved by the Institutional Review Board at Arizona State University. Urine samples were obtained from ten healthy women of Asian, African, Caucasian, and Hispanic origin in the age group of 18-28 years (Table 1). Women included in this study were not taking any medication known to affect hormonal balance, including birth control. The urine samples were acquired daily in 20 mL glass vials, for 30 consecutive days. All sample collection was completed with a 1-month time frame. All samples were collected in the morning between 8 am and 12 pm. The subjects did not eat for 2 hours, prior to collection. A standard sterile collection procedure was followed [26]. Prior to collection, subjects used alcohol wipes to clean the area of collection. The samples were collected mid-stream during urination. Urine samples were capped and remained at 4°C for up to an hour. Within one hour of sample collection, the samples were aliquoted into 1.5mL cryogenic vials (VWR International, Radnor, PA) for storage in a -80°C freezer for one year, prior to testing.

In preparation for sample testing, the 10-mL VOC-free vials and PTFE/silicone caps were baked at 100°C for 12 hours in an oven to reduce contamination and variation across the vials and caps (Supelco/Sigma-Aldrich, St. Louis, MO). One hour before testing, samples, vials, and caps were brought to room temperature. The samples were inverted to mix, and 1 mL of sample was transferred to vials and securely closed with a cap. The sample vials were placed into a chilled tray held at 4°C until tested. A Gerstel MultiPurpose Sampler (MPS; Gerstel, Mülheim an der Ruhr, Germany) was used for

sample preparation, transport, and extraction. Samples were incubated at 60°C with agitation at 250 rpm for 5 minutes. The volatile metabolites of urine were sampled from the headspace by SPME, using a 50/30 µm divinylbenzene/carboxen/ polydimethylsiloxane (DVB/CAR/PDMS; Supelco/Sigma-Aldrich) coated fiber. Prior to each sample extraction, fiber bake-out was performed for 10 min at 270°C. Sample extraction was performed at 60°C with agitation at 250 rpm for 60 min. Subsequently, the fiber was injected into the GC inlet for 5 min at 250°C.

Table 1. Demographics of Subjects. The table shows the age, ethnicity, height, weight, and ovulation confirmation by ovulation kit of the 10 subjects.

Subject No	Age (years)	Ethnicity	Height	Weight (lbs)	Ovulation Confirmation
1	24	Hispanic	5'4"	122	Yes
2	21	Hispanic	5'3"	110	Yes
3	19	Asian	5'6"	150	Yes
4	21	African American	5'1"	105	Yes
5	20	Caucasian	5'5"	109	Yes
6	24	Asian	5'9"	121	Yes
7	19	Hispanic	5'0"	100	Yes
8	18	Caucasian	5'8"	127	No
9	18	Asian	5'4"	130	No
10	27	Caucasian	5'5"	150	No

2.2 Method Optimization

Test conditions were optimized prior to running the study. Selection criteria was based on the maximum number of VOCs that could be collected. The optimal number of analytes collected by SPME were evaluated through sample volume and extraction time measurements. Sample volume was tested at 500 µL, 1 mL, 1.5 mL, and 2 mL and analyzed by GC x GC-TOFMS three times at each sample volume. The average number

of analytes, with less than 15% deviation, collected in the headspace across the three runs was 320 at 500 µL, 394 at 1 mL, 438 at 1.5 mL, and 470 at 2 mL. The sample volume of 1 mL was selected to maximize the number of volatiles collected while keeping the sample volume consistent across runs, with a limited sample amount.

Extraction time was tested at 30, 45, and 60 min, using 1mL urine samples. Preliminary studies were tested using a 65 µm CAR/PDMS fiber within a 10-mL headspace vial. Extraction was performed by SPME in the headspace and analyzed by GC x GC-TOFMS three times consecutively. The average number of analytes, with less than 10% deviation, collected across the three runs was 626 for 30 min, 689 for 45 min, and 819 for 60 min. The extraction time of 60 min was therefore selected to maximize the number of volatiles collected, and used in the final testing conditions, as indicated above. Furthermore, a 50/30 µm DVB/CAR/PDMS coated fiber was chosen due to the wide range (molecular weight of 40-275) of volatiles and semi-volatiles collected.

Degradation of samples was tested by running 10 samples consecutively from one urine collection in a 4°C chilled tray. The signal-to-noise ratio of the compound 2-heptanone across the 10 samples is shown in Figure 1 as an example compound that did not degrade over the 12.6 hours the 10 samples ran. In addition, no significant degradation was found in all the other compounds in the samples over time.

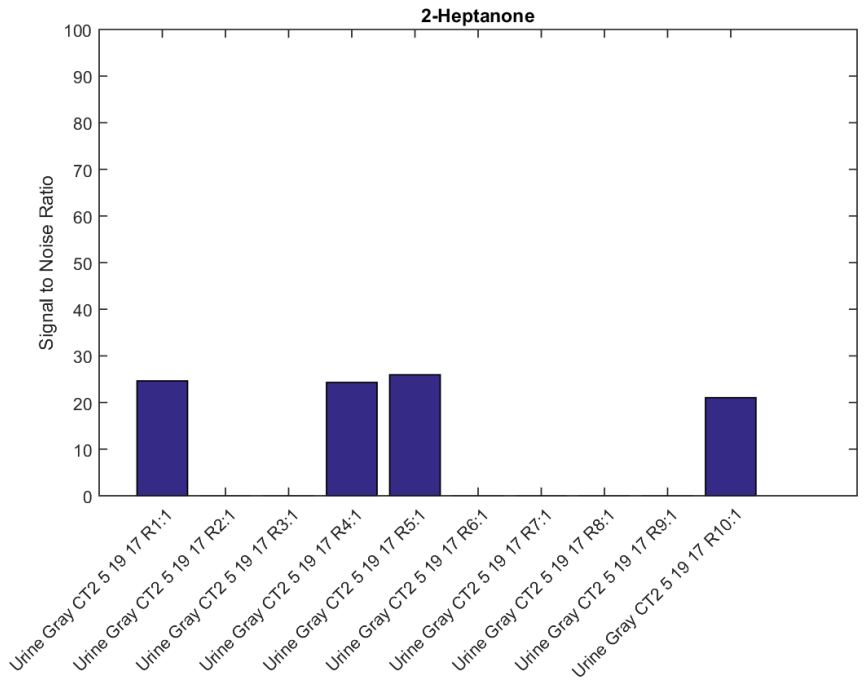


Figure 1. Urine Degradation Study. The analyte 2-heptanone is shown as an example compound, which did not degrade over 10 samples, as a result of the samples being in a 4°C chilled tray.

2.3 Instrumentation

Calibration of the mass spectrometer was executed daily to maintain instrument performance, which included ion optic and source focusing, acquisition system adjustments, mass calibration, tune checks, and leak checks. All analyses of urine samples were performed by GC x GC-TOFMS [27]. The instrument was fitted with a two-dimensional column set, joined together by a press-fit connection. The first column consisted of an Rxi-624Sil MS (60 m x 250 µm x 1.4 µm [length x internal diameter x film thickness]; Restek, Bellefonte, PA). The second column consisted of a Stabilwax (1 m x 250 µm x 0.5 µm; Restek). Each of the two columns were heated independently. The first column in the primary oven was heated with an initial temperature of 50°C, held for 2 min. The oven was ramped at 5°C/min to 225°C and held for 2 min for temperature

stability. Subsequently, the oven was ramped at 30°C/min to 230°C and held for 30 min for post-column bake-out. The second column was heated with a +5°C offset relative to the primary oven. A quad-jet modulator was used with 2 s modulation periods (0.5 s hot, 0.5 s cold pulses) and a +15°C offset relative to the secondary oven. The helium carrier gas flow rate was 2 mL/min. Mass spectra were acquired at 100 Hz over a range of *m/z* = 35-550. Data acquisition was captured by ChromaTOF software, Version 4.60.8.0 (Leco Corp., St. Joseph, MI) [27]. A resulting chromatogram of one urine sample of a single subject on a single day is shown in Figure 2.

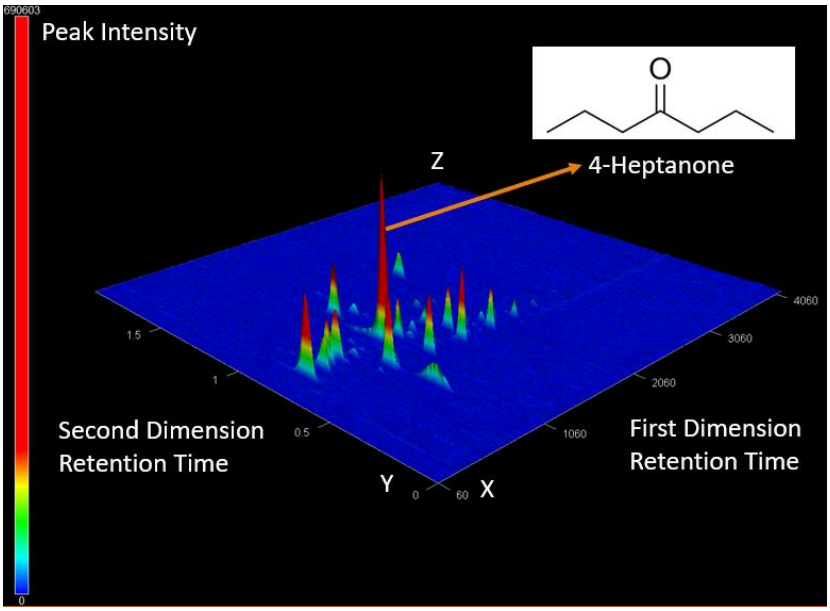


Figure 2. Chromatogram of Urine Sample. The resulting chromatogram of a urine sample of one of the subjects is shown. The compound 4-heptanone has been identified as one of the compounds in the headspace of the urine sample.

2.4 Data Alignment and Normalization

A Kovats index (KI) standard mix of alkanes was used to identify retention times at the beginning, middle, and end of the GC x GC-TOFMS runs. KI is used to confirm compounds eluded at retention times within known values and can be used as a reference

from instrument to instrument as it is independent of the system [28]. Retention indexes were also compared to ensure that the equipment did not drift over time. Data processing and chromatographic alignment were completed using the Statistical Compare package of ChromaTOF software, Version 4.60.8.0. For a peak to be identified as the same compounds across chromatograms, both the retention times and the mass spectra had to meet minimum match criteria. The first-dimension retention time could not vary more than 2 s from chromatogram-to-chromatogram, and the second-dimension retention time could not vary more than 0.2 s from chromatogram-to-chromatogram. The mass spectrum for aligned peaks had to meet a minimum match threshold of 600. The baseline was fixed to be through the middle of the noise, and the signal-to-noise (S/N) cutoff for peak finding was set to 50 for a minimum of two apexing masses. Peaks were putatively identified using the National Institute of Standards and Technology (NIST) Mass Spectral Library and published retention time data. All peaks eluded during blank runs and known contaminants were excluded from data analysis.

Multiple normalization and alignment analysis were performed on the raw data prior to statistical analysis. Relative abundance of the analytes was normalized across chromatograms using the Probabilistic Quotient Normalization (PQN) method [29]. Data normalization was performed in R, Version 1.0.136 [27] prior to any data analysis. Ovulation across all subjects was tested using commercially available ovulation kits (Clearblue). Ovulation is a well-known indicator of healthy fertility in women [30]. The first day of menses was reported by each subject. The ovulation day, signified by an increase in the luteinizing hormone (LH), was confirmed in 7 out of the 10 subjects. All 7 subjects were aligned, so ovulation day corresponded with day 14 of the menstrual cycle.

The data obtained from all 10 subjects were used in the functional group classification. The data acquired from 7 out of the 10 subjects were used for all data analysis, where ovulation was validated.

2.5 Data Analysis: Random Forest and PCA

After the data was processed, filtered, and normalized, the remaining compounds were imported into R for further analysis. Unsupervised learning principal component analysis (PCA) of the compounds showed there was no pattern in the data between four baseline estrogen days and four peak days, across the 7 subjects. Thus, the supervised learning technique of random forest was employed to discover the discriminatory analytes that differentiated between estrogen peak days and estrogen baseline days for the 7 subjects, where ovulation was confirmed [31]. The code used the function "randomForest" from the "Random Forest" package in R. The random forest was run 100 times and outputted the top discriminatory analytes with the least mean decrease in accuracy from the model that drove the classification between baseline estrogen days and peak days. After supervised learning by random forest, PCA was performed on the top discriminatory variables [32] in R with the function "prcomp". The function "ggplot" was used to plot the first principal component to account for the largest possible variance in the variables, and the second principal component to account for the second largest possible variance in the variables. Data analysis was performed in R, Version 1.0.136.

2.6 Data Analysis: Regression Model and Lasso Technique

The data was imported into JMP for regression model analysis. A total of 935 compounds were included in the regression model. The "Fit Model" menu was utilized to calculate the forward-stepwise regression model with Akaike Information Criterion

Corrected (AICc) [33, 34]. The model was employed to identify potential biomarkers from the biological samples, as per the estrogen and progesterone literature curves. To prevent the model from overfitting and reduce the number of compounds, the adaptive Lasso technique with 5-fold cross-validation was performed in JMP with the "Fit Model" menu [35]. The data set was split into 5 different sets, with the Lasso technique using 80% of the data to train the model and 20% of the data to test the model. Significant compounds ($p<0.001$) were retained in the final models. Data analysis was performed in JMP Pro, Version 13.1.0. Regression model calculations used 7 out of the 10 subjects, where ovulation was verified.

2.7 Data Analysis: t-test and Heat Map

Data used in this analysis was evaluated and determined to be accurate for compound significance. The data set was imported into R for analysis. R was used to filter the data set, so the core compounds that appeared in at least half the subjects (3 out of 7 subjects) in at least half the days (14 out of 28 days) were utilized in the data analysis (Appendix C). The missing data values were converted from 0 to NA (not applicable) in the data set and were not used in the analysis. The pairwise Student's t-test was performed on the median of 5 baseline estrogen days vs. the median of 5 peak estrogen days [36]. The function "t.test" was employed in R for the pairwise Student's t-test. The resulting significant compounds ($p<0.05$) were plotted in a heat map to look at the signals of each compound across the menstrual cycle. The "heatmap" function was used to plot the significant compounds in a heatmap. Data analysis was performed in R, Version 1.0.136.

2.8 Compound Validation and Quality of Data

Compounds, selected using the t-test, were validated through a confirmation of the retention indices as shown in data produced from the GC x GC-TOFMS. The presence of retention indices confirms retention times as they relate the identified compounds to the known compounds. The NIST database clarifies the retention time in the GC x GC-TOFMS data. Unconfirmed compounds result in a lower quality analysis. The quality of the compound identifications was assigned (levels 1-4) following published guidelines [28]. Level 2 compounds were categorized based on greater than or equal to 60% mass spectral match, utilizing forward searches of the NIST mass spectral library. In addition, level 2 compounds had validated retention time data with experimentally-determined retention indices that are consistent with the mid-polar Rxi-624Sil stationary phase. Level 2 was the highest classification in this study. Level 3 compounds were classified on greater than or equal to 60% mass spectral match to the NIST library. Level 4 compounds have mass spectral matches less than or equal to 60%, but can still be determined from mass spectral data.

CHAPTER 3

DATA ANALYSIS AND RESULTS

3.1 Functional Groups

With the optimized parameters, the functional groups resulting from the volatiles expressed in all the urine samples of the 10 subjects were analyzed (Figure 3). Volatile analytes were classified as core, rare, or accessory metabolites according to the following criteria:

Core: Identified in all 10 subjects;

Rare: Identified in only 1 subject;

Accessory: Identified in 2-9 subjects.

Functional group analysis shows there are more core analytes (498) associated between all the samples than rare analytes (30). The total number of accessory analytes was 451.

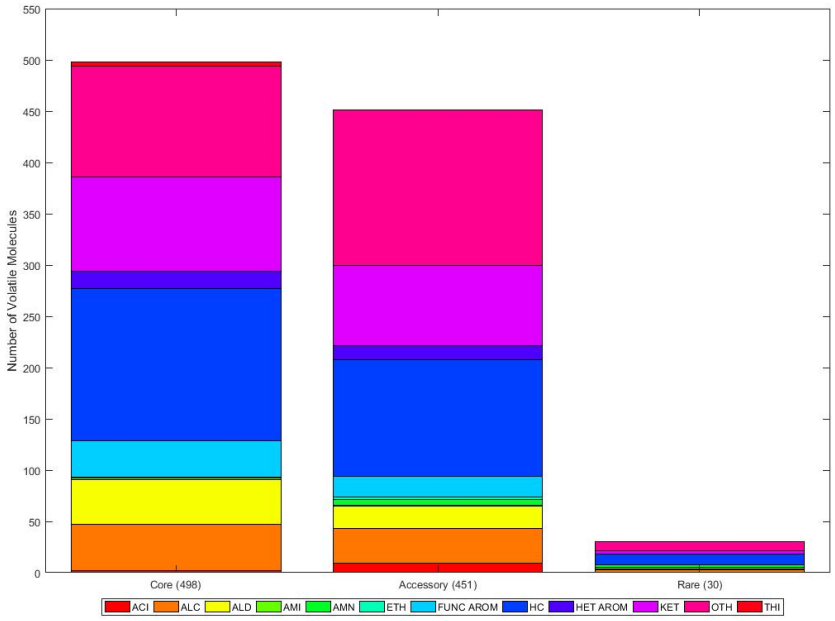


Figure 3. Volatile Analyte Classification. Analytes were classified as core, accessory, or rare, and each one was assigned to a functional group (ACI = Acids; ALC = Alcohols; ALD = Aldehydes; AMI = Amides; AMN = Amines; ETH = Ethers; FUNC AROM = Functionalized Aromatics; HC = Saturated/Unsaturated Hydrocarbons; HET AROM = Heteroaromatics; KET = Ketones; OTH = OTHER; THI = Thiols).

3.2 Random Forest and PCA

This model was not utilized in the final selection of compounds due to the fact that continued analysis of data proved this method to be biased. It was discovered that the model had data values of 0 for missing data, which significantly biased the model. Future studies simplified the model in terms of high and low abundance and corrected the values from 0 to "NA". To find particular analytes correlated with estrogen baseline days and estrogen peak days per the literature curve, random forest was employed to discover the top discriminatory analytes from the 7 subjects, where ovulation was confirmed. Using balanced data of 4 baseline estrogen days and 4 peak estrogen days across all 7 subjects, the top 10 discriminatory analytes from random forest were 1,3,5-Cycloheptatriene, 3,7,7-trimethyl, 1-Butene, 4-isothiocyanato-1-(methylthio), 2,3-Pentanedione, 4-methyl, 2-Butenal, 2-Butenal, (E), 2-Hexanone, 4-Acetyl-1-methylcyclohexene, Allyl Isothiocyanate, Benzyl 4-nitrophenyl carbonate, and Cyclobutylamine. PCA was performed on the 10 resulting analytes to determine the first principal component and second principal component (Figure 4). The first principal component accounted for 37.4% of the variance, and the second principal component accounted for 21.4% of the variance. From the PCA, the baseline estrogen days cluster together, and the peak estrogen days cluster together. Peak estrogen days correspond with ovulation, so a distinct difference can be seen between the peak estrogen days and the baseline estrogen days. It was discovered the data set had values of 0 for missing data, which significantly skewed the model. Thus, random forest and PCA was not used in the final selection of compounds. Future studies simplified the model in terms of high and low abundance, thus, removing any time dependence.

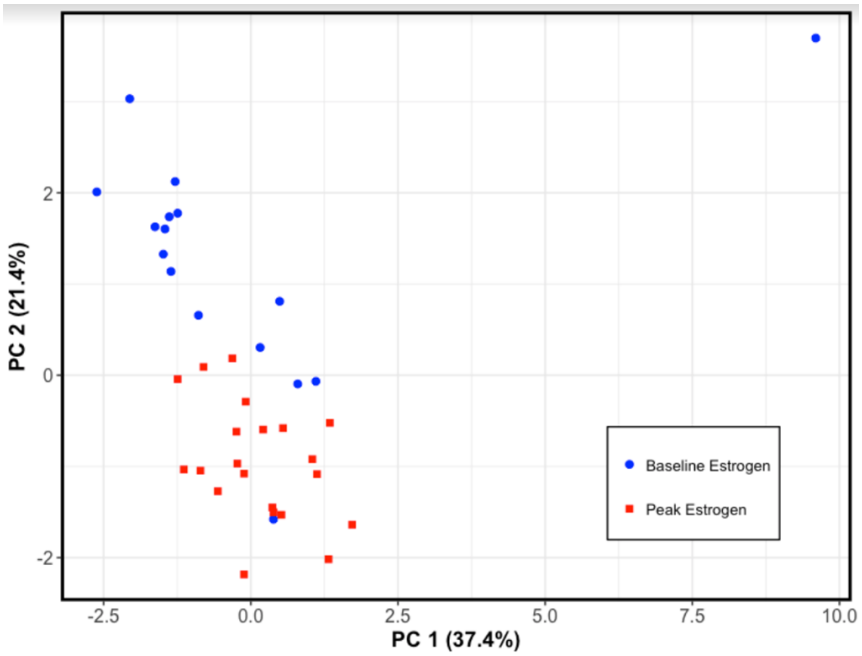


Figure 4. Balanced PCA of 7 Subjects. The PCA of the 7 subjects shows a distinct separation between baseline estrogen days and peak estrogen days. The first principal component accounts for 37.4% of the variance, and the second principal component accounts for 21.4% of the variance. This data was later refuted and proven to be invalid.

3.3 Regression Modeling and Lasso Technique

This model was not utilized in the final selection of compounds due to the fact that continued analysis of the time dependency proved this method to be an invalid approach. Future studies simplified the model in terms of high and low abundance, removing any time dependence and correcting the values from 0 to "NA". To further investigate the particular analytes correlated with the estrogen and progesterone literature curves across a woman's menstrual cycle, forward step-wise regression with AICc was used. A total of 19 of 935 analytes were found to be significant ($p<0.001$) in the estrogen regression model, and a total of 18 of 935 analytes were found to be significant ($p<0.001$) in the progesterone regression model. The R^2 value of both models were 0.999. In order

to prevent the data from being overfit, adaptive Lasso technique was performed. Further analysis using the adaptive Lasso technique with 5-fold cross-validation reduced the significant analytes down to 8 analytes for the estrogen model and down to 9 analytes for the progesterone model. For each hormone analysis, the model with the least out-of-sample mean squared error of the 5 different train/test sets was selected as the predictive model. Model regression showed the predictive model of estrogen to be a linear combination of 8 analytes (Table 2), which strongly fits the estrogen curve as found in literature (Figure 5). Similarly, the model regression showed the predictive model of progesterone to be a linear combination of 9 analytes (Table 3), which strongly fits the progesterone curve as found in literature (Figure 6). The retention indices (Table 2, Table 3) were only able to be calculated for the analytes that had a retention time between the retention times of the KI standard mix. The analyte 2-furanmethanol, 5-ethenyltetrahydro-α,α,5-trimethyl-,cis- was common across estrogen and progesterone models. The predicted expression of the estrogen model is shown in Figure 7, and the predicted expression of the progesterone model is shown in Figure 8. Checking the assumptions of the model, the data was normal (Figure 9, Figure 10). Furthermore, the residual estrogen (Figure 11) and residual progesterone analysis (Figure 12) vs. menstrual cycle days show that no time dependence exists in the models. However, this was disputed and later removed in the study. It was also discovered that the model had data values of 0 for missing data, which significantly biased the model.

Table 2. Top Analytes from Estrogen Regression Model. Key analytes identified by the forward stepwise estrogen regression model with AICc and adaptive Lasso with 5-fold cross-validation. The first and second dimension retention times of each analyte from the GC x GC-TOFMS are shown. The retention indices shown were calculated for the analytes that had a retention time between the retentions time of the KI standard mix. Later refuted and proven to be invalid.

No.	Analyte	First Dimension Retention Time (s)	Second Dimension Retention Time (s)	Retention Index
1	1-Tetrazol-2-ylethanone	360	0.67	N/A
2	2-Furanmethanol, 5-ethenyltetrahydro-α,α,5-trimethyl-,cis-	1580	0.92	1099
3	3-Buten-2-ol, 3-methyl-	704	0.83	N/A
4	3-Hexen-1-ol, acetate,(E)-	1570	0.89	1094
5	4-Octanone	1214	0.77	923
6	5-Ethyl-1-nonene	2108	0.68	N/A
7	Acetone	346	0.79	N/A
8	Decane,2,6,8-trimethyl-	1434	0.65	1029

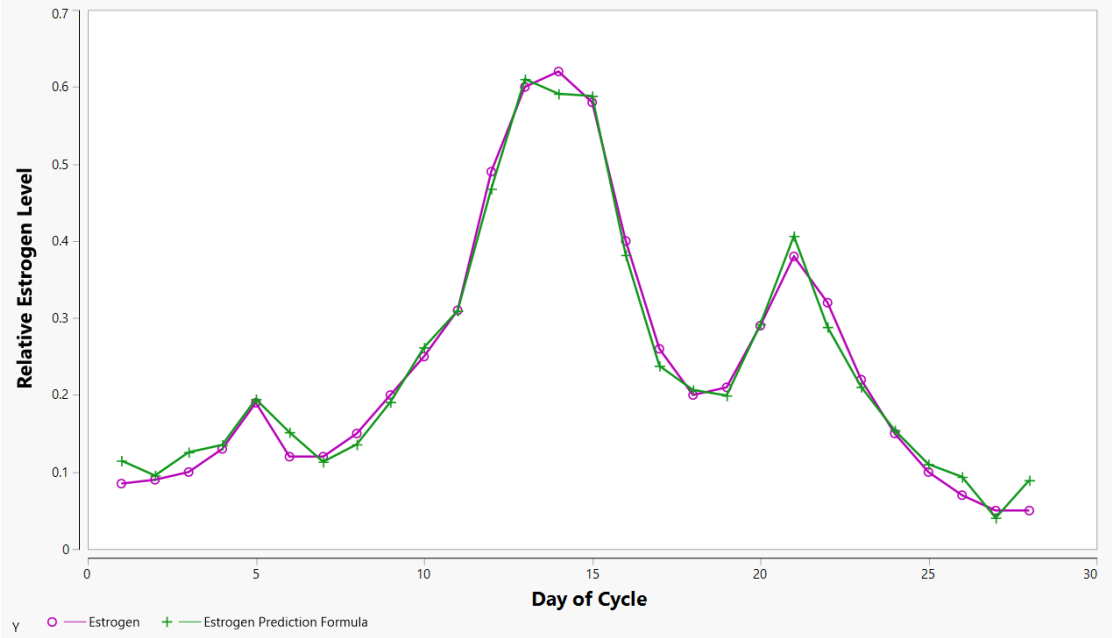


Figure 5. Estrogen Predicted Model vs. Estrogen Literature Curve. The estrogen predicted model vs. the estrogen literature curve is shown across a healthy woman's menstrual cycle. Later refuted and proven to be invalid.

Table 3. Top Analytes from Progesterone Regression Model. Key analytes identified by the forward stepwise progesterone regression model with AICc and adaptive Lasso with 5-fold validation. The first and second dimension retention times of each analyte from the GC x GC-TOFMS are shown. Later refuted and proven to be invalid.

No.	Analyte	First Dimension Retention Time (s)	Second Dimension Retention Time (s)	Retention Index
1	2-Furanmethanol, 5-ethenyltetrahydro-α,α,5-trimethyl-,cis-	1580	0.92	1099
2	2-Heptanone, 6-methyl-6-[3-methyl-3-(1-methylethenyl)-1-cyclopropen-1-yl]-	2472	0.94	N/A
3	2H-Pyran-2-one, tetrahydro-4-(2-methyl-1-propen-3-yl)	1962	1.1	N/A
4	4,4-Dimethyl-1-hexene	1304	0.87	966
5	Cis-Verbenol-	1712	1.08	1163
6	Cyclohexene, 3-(1,5-dimethyl-4-hexenyl)-6-methylene-, [S-R*, S*)]-	2314	0.86	N/A
7	Isopropylcyclobutane	2342	1.2	N/A
8	Pentanal, 2,4-dimethyl-	1416	0.79	1020
9	Undecanal, 2-methyl-	1874	0.78	N/A

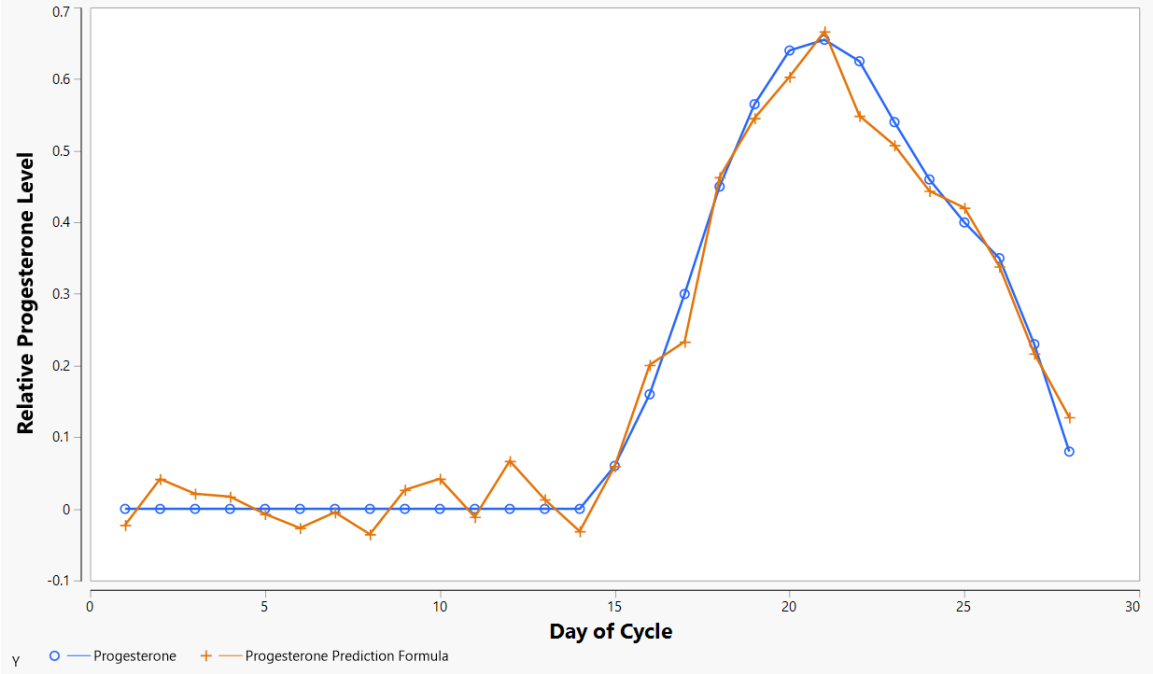


Figure 6. Progesterone Predicted Model vs. Progesterone Literature Curve. The progesterone predicted model vs. the progesterone literature curve is shown across a healthy woman's menstrual cycle. Later refuted and proven to be invalid.

$$y = 0.933 + 0.098*[1] - 0.295*[2] - 0.123*[3] + 0.035*[4] + 0.063*[5] - 0.211*[6] + 0.354*[7] - 0.052*[8]$$

Figure 7. Estrogen Model Predictive Expression. Predictive model expression of the estrogen model is shown with the corresponding analytes in Table 2 ([1] in the equation refers to Analyte No.1 in Table 2). Later refuted and proven to be invalid.

$$y = 0.333 - 0.040*[1] - 0.139*[2] + 0.212*[3] + 0.212*[4] + 0.051*[5] - 0.259*[6] + 0.326*[7] - 0.062*[8] - 0.117*[9]$$

Figure 8. Progesterone Model Predictive Expression. Predictive model expression of the progesterone model with the corresponding analytes in Table 3 ([1] in the equation refers to Analyte No.1 in Table 3). Later refuted and proven to be invalid.

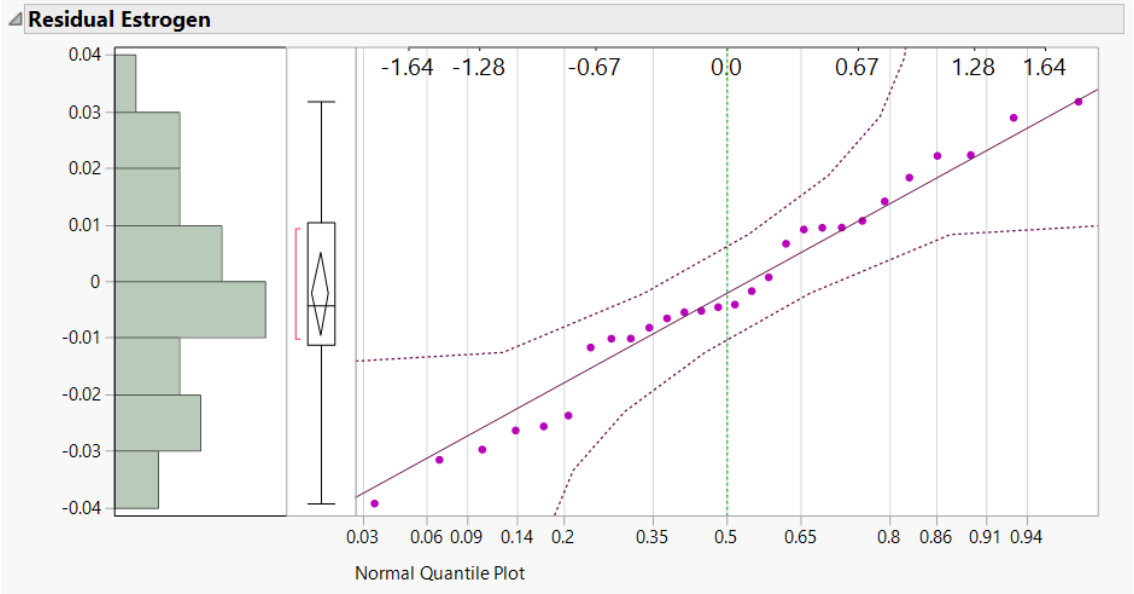


Figure 9. Estrogen Residuals Normal Quantile Plot. Normal quantile plot of the estrogen residuals shows the data to be normal.

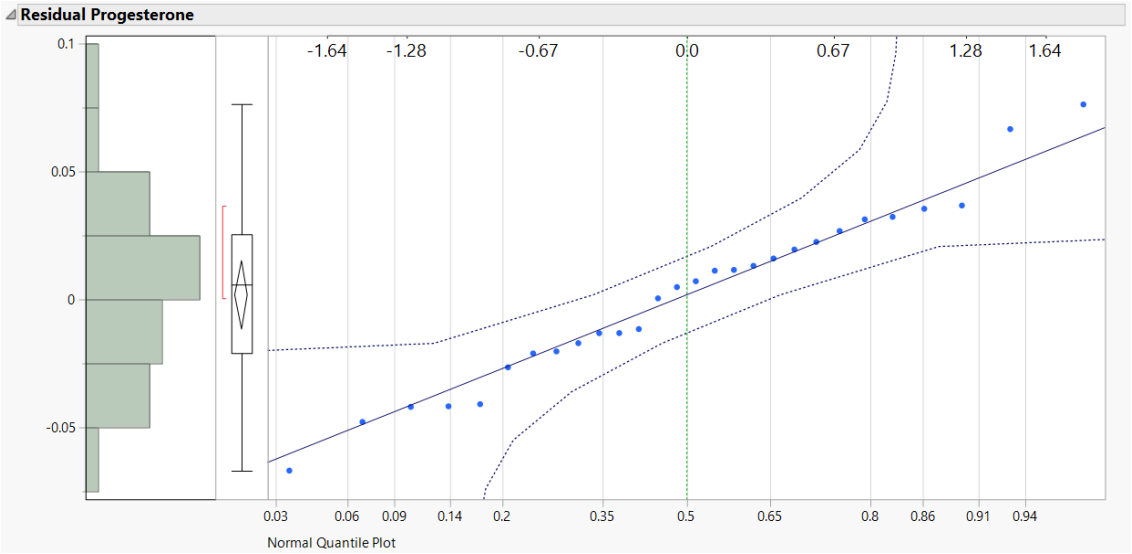


Figure 10. Progesterone Residuals Normal Quantile Plot. Normal quantile plot of the progesterone residuals shows the data to be normal.

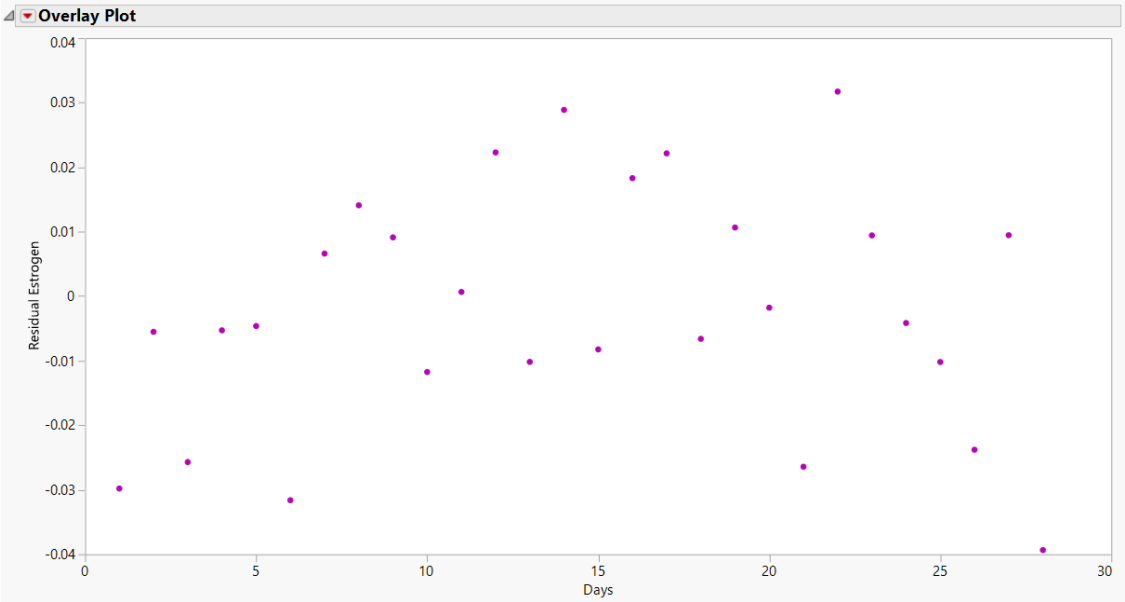


Figure 11. Estrogen Residuals vs. Days. The estrogen residuals plotted vs. days shows there is no time trend in the data. Later refuted and proven to be invalid.

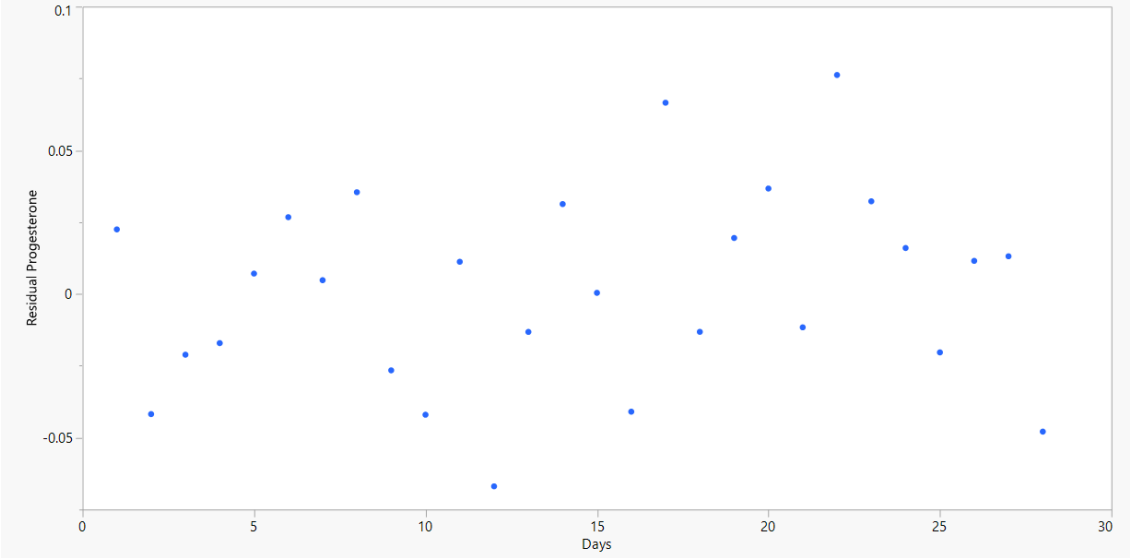


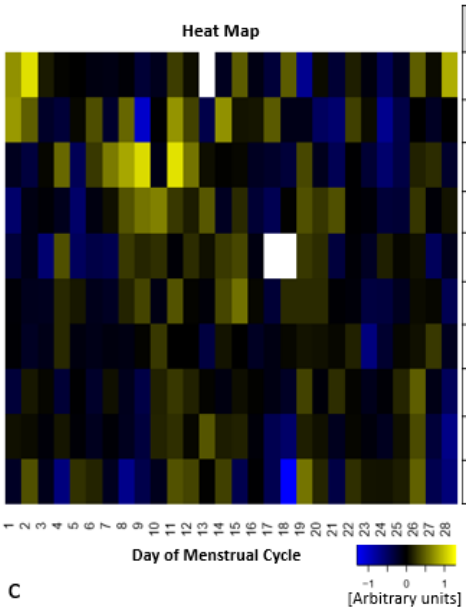
Figure 12. Progesterone Residuals vs. Days. The progesterone residuals plotted vs. days shows there is no time trend in the data. Later refuted and proven to be invalid.

3.4 Data Analysis: t-test and Heat Map

From the data filtering, using the core compounds in half of the subjects in half of the days, 935 compounds were reduced to 347 compounds. The Student's t-test was performed on the median of 5 baseline estrogen days vs. the median of 5 peak estrogen days resulted in 10 significant compounds (Figure 13). These 10 compounds indicated a significant difference between the baseline estrogen days and peak estrogen days, which could help in the predication of ovulation. A heat map of the 10 compounds shows the signal of the compounds across the menstrual cycle (Figure 13). The compound, 3-Octen-2-one, (E), seemed to show a signal around the time of ovulation, and it was validated by the retention index. Ketones were the chemical class that appeared the most number of times (3) from the 10 compounds.

From the Student's t-test, key compounds were discovered, which correlate to the fertility hormone of estrogen a woman's menstrual cycle. Future tests involve validating the 10 significant compounds with independent data sets from other subjects. The independent data sets will seek to test the robustness of the compounds. Further investigations will seek to confirm if the analytes found are caused by the hormones of estrogen. In the future, quantification of the fertility hormones could be predicted, using this model, which could be used to identify declining fertility in women.

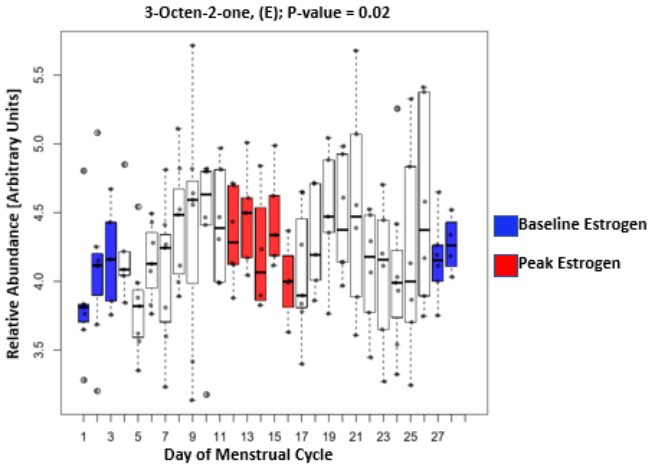
a



b

Compound Name	Mass	ID Level	Chemical Class	^1t_R (s)	^2t_R (s)	RI
o-Cymene	119	2	FUNC AROM	1414	0.82	1019
Benzofuran,4,7-dimethyl	145	3	OTH	1846	1.08	N/A
Bicyclo[7.2.0]undec-4-ene, 4,11,11-trimethyl -8-methylene-,[1R-(1R*,4Z,9S*)]-	41	3	HC	2182	0.79	N/A
3-Octen-2-one, (E)-	55	2	KET	1494	0.92	1058
2(3H)-Furanone, dihydro-5-pentyl-	85	3	KET	2192	1.18	N/A
Cyclobutane, butyl	43	2	HC	1540	1.01	1080
Butanal, 2,2-dimethyl-	43	2	ALD	1546	1.03	1083
3-Ethylcyclopentanone	83	2	KET	1346	0.96	987
Furan, 2-hexyl	81	2	OTH	1320	0.81	974
(3E,5Z)-1,3,5-Undecatriene	79	2	HC	1712	0.79	1163

c



d

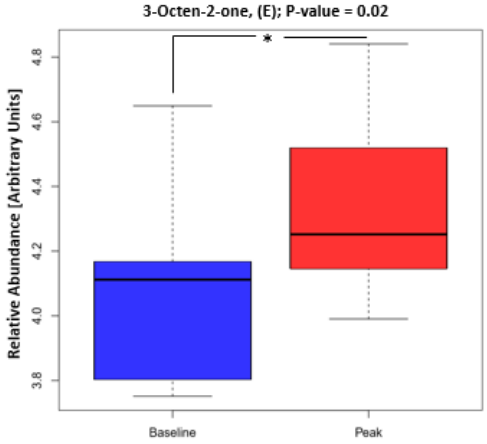


Figure 13. Significant Compounds from t-test Analysis. (a) The heat map of the 10 significant compounds from the t-test analysis is shown. (b) The 10 significant compounds and their mass, ID level, chemical class, first dimension retention time (^1t_R), second dimension retention time (^2t_R), and retention index (RI) are displayed (For chemical class: ALD = Aldehydes; FUNC AROM = Functionalized Aromatics; HC = Saturated/Unsaturated Hydrocarbons; KET = Ketones; OTH = OTHER). (c) One of the 10 significant compounds, 3-Octen-2-one, (E), and the plot of the relative abundance of the 7 subjects across the menstrual cycle are shown. Baseline estrogen days are marked in blue and peak estrogen days are marked in red. (d) Plot of baseline estrogen days vs peak days for 3-Octen-2-one, (E) shows a significant difference between the two.

CHAPTER 4

CONCLUSION

An analytical method was established to discover a more complete representation of the VOCs present in the urine samples of 10 healthy women. A total of 935 different analytes were identified in the urine samples. The t-test and heat map show there are 10 key analytes within a woman's menstrual cycle that can track the fertility hormone estrogen. From the functional group comparison, core common analytes were found to be shared across all the subjects. Through this study, we have identified potential volatile biomarkers that show statistical significance in correlation to hormonal changes in healthy women. Future research will aim to validate these biomarkers as an early expression of declining fertility in women.

REFERENCES

1. Chandra A, Copen CE, Stephen, EH. Infertility and Impaired Fecundity in the United States, 1982-2010. In: National Health Statistics Reports. Centers for Disease Control and Prevention. 2013. https://www.cdc.gov/nchs/data/nhsr/ nhsr067.pdf. Accessed 4 Oct 2017.

2. Marketdata Enterprises, Inc. U.S. fertility clinics & infertility services: an industry analysis. 2013. https://www.prlog.org/12236385-us-fertility-clinics-infertility-services-market-worth-35-billion.html. Accessed 4 Aug 2017.

3. Wunder DM, Bersinger NA, Yared M, Kretschmer R, Birkhäuser MH. Statistically significant changes of antimüllerian hormone and inhibin levels during the physiologic menstrual cycle in reproductive age women. Fertility and sterility. 2008 Apr 1;89(4):927-33.

4. Schally AV, Kastin AJ, Arimura A. Hypothalamic follicle-stimulating hormone (FSH) and luteinizing hormone (LH)-regulating hormone: structure, physiology, and clinical studies. Fertility and sterility. 1971 Nov 1;22(11):703-21.

5. Ho SM. Estrogen, progesterone and epithelial ovarian cancer. Reproductive Biology and Endocrinology. 2003 Dec;1(1):73.

6. Ziegler RG, Rossi SC, Fears TR, Bradlow HL, Adlercreutz H, Sepkovic D, Kiuru P, Wahala K, Vaught JB, Donaldson JL, Falk RT. Quantifying estrogen metabolism: an evaluation of the reproducibility and validity of enzyme immunoassays for 2-hydroxyestrone and 16alpha-hydroxyestrone in urine. Environmental health perspectives. 1997 Apr;105(Suppl 3):607.

7. Jaslow CR, Carney JL, Kutteh WH. Diagnostic factors identified in 1020 women with two versus three or more recurrent pregnancy losses. Fertility and sterility. 2010 Mar 1;93(4):1234-43.

8. Moghissi KS. Accuracy of basal body temperature for ovulation detection. Fertility and sterility. 1976 Dec 1;27(12):1415-21.

9. Guermandi E, Vegetti W, Bianchi MM, Uglietti A, Ragni G, Crosignani P. Reliability of ovulation tests in infertile women. Obstetrics & Gynecology. 2001 Jan 1;97(1):92-6.

10. van Rooij IA, Broekmans FJ, Scheffer GJ, Looman CW, Habbema JD, de Jong FH, Fauser BJ, Themmen AP, te Velde ER. Serum antimüllerian hormone levels best reflect the reproductive decline with age in normal women with proven fertility: a longitudinal study. Fertility and sterility. 2005 Apr 1;83(4):979-87.

11. Liebich HM, Al-Babbili O. Gas chromatographic-mass spectrometric study of volatile organic metabolites in urines of patients with diabetes mellitus. Journal of Chromatography A. 1975 Jan 1;112:539-50.

12. Matsuo K, Opper NR, Ciccone MA, Garcia J, Tierney KE, Baba T, Muderspach LI, Roman LD. Time interval between endometrial biopsy and surgical staging for type I endometrial cancer: association between tumor characteristics and survival outcome. Obstetrics & Gynecology. 2015 Feb 1;125(2):424-33.

13. Yan J, Kuzhiumparambil U, Bandodkar S, Solowij N, Fu S. Development and validation of a simple, rapid and sensitive LC-MS/MS method for the measurement of urinary neurotransmitters and their metabolites. Analytical and bioanalytical chemistry. 2017 Dec 1;409(30):7191-9.

14. Fuertig R, Ceci A, Camus SM, Bezard E, Luippold AH, Hengerer B. LC–MS/MS-based quantification of kynurenine metabolites, tryptophan, monoamines and neopterin in plasma, cerebrospinal fluid and brain. Bioanalysis. 2016 Sep;8(18):1903-17.

15. Albrecht J, Zielińska M. Mechanisms of excessive extracellular glutamate accumulation in temporal lobe epilepsy. Neurochemical research. 2017 Jun 1;42(6):1724-34.

16. Eliassen AH, Spiegelman D, Xu X, Keefer LK, Veenstra TD, Barbieri RL, Willett WC, Hankinson SE, Ziegler RG. Urinary estrogens and estrogen metabolites and subsequent risk of breast cancer among premenopausal women. Cancer research. 2012 Feb 1;72(3):696-706.

17. Armstrong DT, King ER. Uterine progesterone metabolism and progestational response: effects of estrogens and prolactin. Endocrinology. 1971 Jul 1;89(1):191-7.

18. Khan FI, Ghoshal AK. Removal of volatile organic compounds from polluted air. Journal of loss prevention in the process industries. 2000 Nov 1;13(6):527-45.

19. Wahl HG, Hoffmann A, Luft D, Liebich HM. Analysis of volatile organic compounds in human urine by headspace gas chromatography–mass spectrometry with a multipurpose sampler. Journal of Chromatography A. 1999 Jun 25;847(1-2):117-25.

20. Li Y, Song X, Zhao X, Zou L, Xu G. Serum metabolic profiling study of lung cancer using ultra high performance liquid chromatography/quadrupole time-of-flight mass spectrometry. Journal of Chromatography B. 2014 Sep 1;966:147-53.

21. Sugimoto M, Wong DT, Hirayama A, Soga T, Tomita M. Capillary electrophoresis mass spectrometry-based saliva metabolomics identified oral, breast and pancreatic cancer-specific profiles. Metabolomics. 2010 Mar 1;6(1):78-95.

22. Arasaradnam RP, Westenbrink E, McFarlane MJ, Harbord R, Chambers S, O'Connell N, Bailey C, Nwokolo CU, Bardhan KD, Savage R, Covington JA. Differentiating Coeliac disease from irritable bowel syndrome by urinary volatile organic compound analysis–a pilot study. PloS one. 2014 Oct 16;9(10):e107312.

23. James-Todd T, Stahlhut R, Meeker JD, Powell SG, Hauser R, Huang T, Rich-Edwards J. Urinary phthalate metabolite concentrations and diabetes among women in the National Health and Nutrition Examination Survey (NHANES) 2001–2008. Environmental health perspectives. 2012 Sep;120(9):1307.

24. Franke AA, Custer LJ, Morimoto Y, Nordt FJ, Maskarinec G. Analysis of urinary estrogens, their oxidized metabolites, and other endogenous steroids by benchtop orbitrap LCMS versus traditional quadrupole GCMS. Analytical and bioanalytical chemistry. 2011 Sep 1;401(4):1319.

25. Bean HD, Dimandja JM, Hill JE. Bacterial volatile discovery using solid phase microextraction and comprehensive two-dimensional gas chromatography–time-of-flight mass spectrometry. Journal of Chromatography B. 2012 Jul 15;901:41-6.

26. Rabinovitch A, Sarewitz SJ, Woodcock SM, Allinger DB, Azar M, et al. Urinalysis and Collection, Transportation, and Preservation of Urine Specimens. 2nd ed. Wayne: The National Committee for Clinical Laboratory Standards; 2001.

27. Bean HD, Rees CA, Hill JE. Comparative analysis of the volatile metabolomes of Pseudomonas aeruginosa clinical isolates. Journal of breath research. 2016 Nov 21;10(4):047102.

28. Sumner LW et al 2007. Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI) Metabolomics 3 211–21.

29. Dieterle F, Ross A, Schlotterbeck G, Senn H. Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics. Analytical chemistry. 2006 Jul 1;78(13):4281-90.

30. Richards JS, Pangas SA. The ovary: basic biology and clinical implications. The Journal of clinical investigation. 2010 Apr 1;120(4):963-72.

31. Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP. Random forest: a classification and regression tool for compound classification and QSAR modeling. Journal of chemical information and computer sciences. 2003 Nov 24;43(6):1947-58.

32. Jolliffe IT. Principal component analysis and factor analysis. In Principal component analysis 1986 (pp. 115-128). Springer, New York, NY.

33. Park KY, Qiu P. Model selection and diagnostics for joint modeling of survival and longitudinal data with crossing hazard rate functions. Statistics in medicine. 2014 Nov 20;33(26):4532-46.

34. Weisberg S. Applied linear regression. John Wiley & Sons; 2005 Apr 1.

35. Ma S, Huang J. Penalized feature selection and classification in bioinformatics. Briefings in bioinformatics. 2008 Jun 18;9(5):392-403.

36. Wiklund S, Johansson E, Sjöström L, Mellerowicz EJ, Edlund U, Shockcor JP, Gottfries J, Moritz T, Trygg J. Visualization of GC/TOF-MS-based metabolomics data for identification of biochemically interesting compounds using OPLS class models. Analytical chemistry. 2008 Jan 1;80(1):115-22.

APPENDIX A

MATLAB CODE FOR DATA FILTERING

```
%% Preliminary
% Start with a clean slate
close all
clear
clc
format compact

%% Variables % 0 = no, 1 = yes
plotBySubject = 0;
plotByVOC = 0;
plotByTotalArea = 0;
writeGroups = 0 ;
maxAbsence = 10000;
maxArea = 0.000001;
stepSize = 20; % VOCs per graph when plotBySubject == 1

%% Functional Groups
path = [pwd, '\Data\']; % set path to excel sheet
files = dir([path, '*.xlsx']);
funcGroups = {files.name};

%% Read data
page = 1; % page of excel sheet to read
```

```
matrix = zeros(3,length(funcGroups));
ii = 1;
close all
[~, ~, raw]  = xlsread([path, funcGroups{ii}], page); % Read data from excel sheet
raw = raw';
% Compartmentalize
[rows, cols] = size(raw);                % Size of excel sheet                % Get names of
VOCs
subjectInfo = raw(1,2:end)';             % Get information on subject
subject = zeros(length(subjectInfo)-100,3); % Allocate memory
for i = 1 : length(subjectInfo)          % Run through subject info
   str = subjectInfo{i,1};               % Set info to string
   C = strsplit(str,'_');                % Split string based on _
   if strcmp(C{1}, 'Blank') == 1         % If Blank, do nothing
   else
      subject(i,1) = str2double(C{1});   % Set Subject ID to col 1
      subject(i,2) = str2double(C{2});   % Set day to col 2
      subject(i,3) = i;                  % Set index to col 3
   end
end

%% Remove based on absences and max area
[m, n] = size(raw);
```

```
for j = 2:m
   for i = 2:n
      if raw{j,i} == 0
         raw{j,i} = NaN;
      end
   end
end
for j = m : -1 : 2
   numOfNans = sum(isnan(cell2mat(raw(j,2:end))));
   tmp = max(cell2mat(raw(j,2:end)));
   if numOfNans > maxAbsence || tmp < maxArea
      raw(j,:) = [];
   end
end

%%
uni = unique(subject(:,1));
cnt = 1;
superCell = cell(7,1);
xAxis = cell(7,1);
tmpMatrix = zeros(5,3);
for i = 1 : length(uni)
   for j = 1 : length(subject)
```

```
        if uni(i) == subject(j,1)
            tmpMatrix(cnt,:) = subject(j,:);
            cnt = cnt + 1;
        end
    end
    [values, order] = sort(tmpMatrix(:,2));
    sortedMatrix = tmpMatrix(order,:);
    sortedCell = raw(:,1);
    cnt2 = 2;
    for k = 1 : length(sortedMatrix)
        index = sortedMatrix(k,3);
        sortedCell(:,cnt2) = raw(:,index+1);
        %       sortedCell(:,cnt2) = raw(:,3+5*index+3);
        cnt2 = cnt2 + 1;
    end
    superCell{i} = sortedCell;
    xAxis{i} = sortedMatrix(:,2);
    clear tmpMatrix
    clear sortedCell
    clear sortedMatrix
    cnt = 1;
end
```

```
%% Filter for Core
nameCell = superCell{i,1}(:,1);
numSubjects = length(uni);
barGraphMat = zeros(m,numSubjects+1);
tempVect = 1:m;
barGraphMat = [barGraphMat, tempVect'];
for i = 1 : numSubjects
   barGraphMat(1,i) = uni(i);
   tmpMatrix = superCell{i,1};
   [m, n] = size(tmpMatrix);
   for j = 2 : m
      numOfNans = sum(isnan(cell2mat(tmpMatrix(j,2:end))));
      superCell{i,1}(j,n+1) = {n-1-numOfNans};
      barGraphMat(j,i) = n-1-numOfNans;
   end
end

core = 0;
accessory = 0;
rare = 0;
for j = m : -1 :2
   idx = barGraphMat(j,1:numSubjects) == 0;
   tempValue = numSubjects - sum(idx(:));
```

```
    barGraphMat(j,numSubjects+1) = tempValue;
    if tempValue == 7
        core = core + 1;
    elseif tempValue == 1
        rare = rare + 1;
        barGraphMat(j,:) = [];
        nameCell(j,:) = [];
    elseif tempValue == 0
        display('empty???')
        barGraphMat(j,:) = [];
        nameCell(j,:) = [];
    else
        accessory = accessory + 1;
        barGraphMat(j,:) = [];
        nameCell(j,:) = [];
    end
end

%% Analyze

core = 0;
accessory = 0;
rare = 0;
```

```
[m2, ~] = size(barGraphMat);
for j = m2 : -1 : 2
    idx = barGraphMat(j,1:numSubjects) < 14;
    tempValue = numSubjects - sum(idx(:));
    barGraphMat(j,numSubjects+1) = tempValue;
    if tempValue > 3
        core = core + 1;
    elseif tempValue == 1
        rare = rare + 1;
        barGraphMat(j,:) = [];
        nameCell(j,:) = [];
    elseif tempValue == 0
        display('empty???')
        barGraphMat(j,:) = [];
        nameCell(j,:) = [];
    else
        accessory = accessory + 1;
        barGraphMat(j,:) = [];
        nameCell(j,:) = [];
    end
end

nameCell = [nameCell, num2cell(barGraphMat)];
```

```
nameCell{1,9} = 'Over 14';
nameCell{1,10} = 'Row in Raw';
matrix(1,ii) = core;
matrix(2,ii) = accessory;
matrix(3,ii) = rare;
xlswrite([path, funcGroups{ii}(1:end-5),'_updated1.xlsx'], nameCell)

%% Add data to matrix from raw
concatCell = superCell{1};
for i = 2 : numSubjects
    concatCell = [concatCell, superCell{i}(:,2:end)];
end

nameCell2(1,:) = [nameCell(1,:), concatCell(1,:)];
[f, g] = size(nameCell);
for i = 2 : f
    nameCell2(i,:) = [nameCell(i,:), concatCell(nameCell{i,g},:)];
end
xlswrite([path, funcGroups{ii}(1:end-5),'_updated2.xlsx'], nameCell2)

%% Graph
legendNames = cell(1, numSubjects);
for i = 1 : length(funcGroups)
```

```
    legendNames{i} = funcGroups{i}(1:end-5);
end
% figure('units','normalized','outerposition',[0 0 1 1]);
bar(matrix,'stacked')
xt = get(gca, 'XTick');
set(gca, 'XTick', xt, 'XTickLabel', {'Core' 'Accessory' 'Rare'})
% legend(legendNames, 'Location','northwest','Orientation','horizontal')
% legend(legendNames)
ylabel('Number of Volatile Molecules')
% ylim([0 550])

%%
```

APPENDIX B

R CODE FOR PROBABIBILISTIC QUOTIENT NORMALIZATION (PQN)

```
## Function: Probabilistic Quotient Normalization ##
PQN <- function (X) {

  obs <- dim(X)[1]     #Define number of observations.
  dimm <- dim(X)[2]     #Define number of variables (dimensions).

  X[0==X] <- 1E-08     #Set zeroes to an arbitrarily small value.
  normRef <- apply(X,2,function(x){median(x[x>1E-08])})     #Define reference
spectrum as median for all analytes.

  M <- matrix(rep(normRef, each=obs), ncol=length(normRef))     #Convert reference
spectrum in matrix equivalent in size to data matrix.
  Q <- X/M     #Divide the concentration of the analyte in each sample by the median
value for each analyte.
  Q[0.001 >= Q] <- NA      #Set very small values of "Q" equal to NA for elimination in a
subsequent step.
  for (i in 1:obs) {
    X[i,] <- X[i,]/median(Q[i,], na.rm=TRUE)}     #Divide each analyte in a given sample
by the median quotient in that sample.

  X[1 >= X] <- 0     #Convert very small normalized values to 0.
  return(matrix(X, nrow=obs, ncol=dimm))
}
```

```
PREP <- function (X) {
  dimm <- dim(X)[1] #Define the number of variables (dimensions).
  obs <- dim(X)[2] #Define the number of observations.
  analyte.names <- X[,1] #Determine the analyte names.
  X <- t(X[,-1]) #Transpose the data matrix without the first column.
  colnames(X) <- analyte.names #Make the analyte names the names of matrix columns.
  X[is.na(X)] <- 0
  return(as.matrix(X))
}

## Function: Data Matrix Final Transposition ##
REMATRIX <- function (X) {
  X <- t(X) #Transpose the data matrix without the first column.
  X[0==X] <- NA
  return(X)}
```

APPENDIX C

R CODE FOR T-TEST AND HEAT MAP

```
# Load data
d1 =
read.csv('urine_data_normalized_7subjects_orderedformenstrualcycle_rev1_NA.csv',
header=T, row.names=1)

# Half the subjects across half the days
compounds = c(3:length(colnames(d1)))
f1 = sapply(colnames(d1)[compounds], function(c1) { tmp = sapply(1:28, function(x) {
sum(is.na(d1[which(d1[,'Day.of.Menstrual.Cycle']==x),c1]))<=4 } );
return(sum(tmp)/length(tmp)) } )
f2 = names(which(f1>=0.5))

# Get sample information
s1 = unique(d1[,'Day.of.Menstrual.Cycle'])
n1 =  unique(d1[,'Subject.ID'])

# Samples to median for T-test
baseline = c(27,28,1,2,3)
peak = c(12,13,14,15,16)
#baseline = c(28,1,2)
#peak = c(13,14,15)
#baseline = c(1)
#peak = c(14)
```

```
# For each compound comapre using T-test
pdf('boxplot_p_m_2days_split.pdf')
m1 = matrix(ncol=2,nrow=length(f2))
colnames(m1) = c('FC','P.value')
rownames(m1) = f2
for(c1 in f2) {
   bsl1 = sapply(n1, function(x) {
median(d1[intersect(which(d1[,'Day.of.Menstrual.Cycle'] %in%
baseline),which(d1[,'Subject.ID']==x)),c1],na.rm=T) } )
   pk1 = sapply(n1, function(x) {
median(d1[intersect(which(d1[,'Day.of.Menstrual.Cycle'] %in%
peak),which(d1[,'Subject.ID']==x)),c1],na.rm=T) } )
   t1 = try(t.test(pk1,bsl1,paired=T))
   print(t1)
   if(!class(t1)=='try-error') {
      m1[c1,'FC'] = median(pk1,na.rm=T)/median(bsl1,na.rm=T)
      m1[c1,'P.value'] = t1$p.value
      if(t1$p.value<=0.05) {
         boxplot(c(bsl1,pk1) ~
c(rep('Baseline',length(bsl1)),rep('Peak',length(pk1))),col=c(rgb(0,0,1,0.8),rgb(1,0,0,0.8)),
main=paste(c1,'; P-value =',signif(m1[c1,'P.value'],2),sep=''))
      }
```

```
      #stripchart(c(bsl1,pk1) ~ c(rep('Baseline',length(bsl1)),rep('Peak',length(pk1))),
vertical = TRUE, method = "jitter", add = TRUE, pch = 20, col = rgb(0,0,0,0.5))
   }
}
m2 = na.omit(m1)
m2 = cbind(m2,Adj.P.value=p.adjust(m2[,'P.value'],method='BH'))
write.csv(m2,'p_m_2days.csv')
dev.off()

# Volcano plot
plot(log2(m2[,'FC']),-log10(m2[,'P.value']),col=rgb(1,0,0,0.5),pch=20)
abline(h=-log10(0.05),lty=2)
abline(v=c(-log2(2),log2(2)),lty=2)

# Plot major players
up = rownames(m2)[intersect(which(m2[,'FC']>1),which(m2[,'P.value']<=0.05))]
down = rownames(m2)[intersect(which(m2[,'FC']<1),which(m2[,'P.value']<=0.05))]
pdf('up_and_down_p_m_2days.pdf')
for(i in c(up,down)) {
   boxplot(as.numeric(d1[,i]) ~
addNA(as.numeric(d1[,'Day.of.Menstrual.Cycle'])),col=c(rep(rgb(0,0,1,0.8),3),rep('white'
,8),rep(rgb(1,0,0,0.8),5),rep('white',10),rep(rgb(0,0,1,0.8),3)),main=paste(i,'; P-value
=',signif(m2[i,'P.value'],2),sep=''))
```

```
   stripchart(as.numeric(d1[,i]) ~ addNA(as.numeric(d1[,'Day.of.Menstrual.Cycle'])),
vertical = TRUE, method = "jitter", add = TRUE, pch = 20, col = rgb(0,0,0,0.5))
}
dev.off()

# Matrix of median expression across participants for each compound
tmp = sapply(f2, function(c1) { sapply(1:28, function(x) {
median(d1[which(d1[,'Day.of.Menstrual.Cycle']==x),c1],na.rm=T) }) } )
tmp2 = sapply(colnames(tmp), function(x) { (tmp[,x] - median(tmp[,x],
na.rm=T))/mad(tmp, na.rm=T) } )

library(gplots)
pdf('heatmap.pdf')
heatmap.2(t(as.matrix(tmp2[,c(up,down)])),trace='none',col=colorpanel(256,'blue','black',
'yellow'),Colv=F,Rowv=T,dendrogram='row',density.info='none')
dev.off()



library(gplots)
m3 = t(sapply(1:length(unique(cut1)), function(x) {
apply(tmp[,names(which(cut1==x))],1,mean,na.rm=T) }))
```

```
heatmap.2(as.matrix(m3),trace='none',col=colorpanel(256,'blue','black','yellow'),Colv=F,
Rowv=T,dendrogram='none',density.info='none')
```

BIOGRAPHICAL SKETCH

Stephanie Marie Ong is a native of Phoenix, Arizona, and earned her Bachelor of Science in Mechanical Engineering from the University of California, Berkeley. Before returning to graduate school, she interned at Apple and worked full-time at Intel Corporation as a process engineer. Upon entering graduate school, she joined Dr. Barbara Smith's lab as a graduate research assistant. Under the mentorship of Dr. Barbara Smith, she has worked on discovering biomarkers in biological samples relating to fertility hormones in women through two-dimensional gas chromatography-time-of-flight mass spectrometry (GC×GC-TOFMS) and regression models. Her work has resulted in poster presentations at the Molecular, Cellular, and Tissue Bioengineering (MCTB) Symposium and Biomedical Engineering Society (BMES) Conference. She currently has one journal article in process. She received the Merit Award Stipend from the School of Biological and Health Systems Engineering (SBHSE) Department for academic achievements and was inducted into ASU's chapter of the national biomedical engineering honor society, Alpha Eta Mu Beta (AEMB). In her free time, she enjoys playing sand volleyball and basketball and is an avid fan of the Arizona Cardinals.