Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Lifan Zhang Date

Identification of Significant Metabolites Associated with Multi-drug Resistant Tuberculosis through Metabolome-wide Association Study

By

Lifan Zhang

MSPH

Emory University

Rollins School of Public Health

Department of Biostatistics

_____ [Chair's Signature]

Dr. Tianwei Yu

_____ [Member's Signature]

Dr. Thomas Ziegler

Identification of Significant Metabolites Associated with Multi-drug Resistant Tuberculosis through Metabolome-wide Association Study

By

Lifan Zhang

MSPH

Emory University

Rollins School of Public Health

Department of Biostatistics

Advisor: Tianwei Yu, PhD

An abstract of

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of

Master of Science in Public Health

in Biostatistics

2015

Identification of Significant Metabolites Associated with Multi-drug Resistant Tuberculosis through Metabolome-wide Association Study

By: Lifan Zhang

Abstract:

Tuberculosis is an infectious disease that still causes a huge disease burden worldwide, and the treatment of multi-drug resistant tuberculosis is particularly difficult. The object of thesis is to show that tuberculosis drug susceptibility can be identified using metabolome-wide association study techniques. We collected plasma samples from drug-sensitive and multi-drug resistant tuberculosis patients in Georgia, and performed high throughput mass spectrometry to identify possible metabolites. We then built statistical models to identify metabolites significantly correlated with drug susceptibility. In addition, we highlighted the different behaviors of those significant metabolites by visualizing them on heatmaps. Some possible pathways identified through the significant metabolites are also presented.

Keywords:

Multiple-drug resistant tuberculosis, Georgia, high throughput mass spectrometry, metabolome-wide association study, metabolomic pathway

Identification of Significant Metabolites Associated with Multi-drug Resistant Tuberculosis through Metabolome-wide Association Study

By

Lifan Zhang

MSPH

Emory University

Rollins School of Public Health

Department of Biostatistics

Advisor: Tianwei Yu, PhD

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of

Master of Science in Public Health

in Biostatistics

2015

Acknowledgements

I would first like to thank Dr. Tianwei Yu and Dr. Elizabeth Chong. They have given me tremendous advice and support in helping me through data analysis and thesis writing.

I would also like to express my gratitude to Dr. Thomas Ziegler, Dr. Jennifer Frediani, Dr. Jessica Alvarez, and their colleagues in Georgia. It is with their efforts that the data for this study is collected and ready for analysis. Especially, I appreciate Dr. Jennifer Frediani for guiding me through the pathway analysis, and Dr. Thomas Ziegler for taking the time to read my thesis.

Lastly I want to give my special thanks to the faculty and staff of the Biostatistics Department for making the past two years of my study so rewarding.

Chapter I Introduction

Tuberculosis

Tuberculosis is an infectious disease primarily affecting the lungs. Caused most commonly by the bacteria *Mycobacterium tuberculosis*, tuberculosis may induce such symptoms as chronic cough, bloody sputum, fever, and weight loss. Most people infected by the bacteria undergo a latent phase and do not show any symptoms; only 10% of latent infection cases develop into active cases. Although tuberculosis has been recorded in human history since ancient times, the global disease burden of tuberculosis still remains high. 8.6 million cases of tuberculosis are estimated to have occurred in 2012 worldwide, including 940,000 estimated deaths (Zumla, et al., 2013).

An especially alarming issue that severely impedes the treatment of tuberculosis is the occurrence of multidrug-resistant tuberculosis. The issue arose after early 20th century, when the disease began to be treated with antibiotics. If the patients are treated with only one anti-tuberculosis drug, the tuberculosis bacteria strains will be under selection pressure, and that could result in some strains evolving into drug-resistant mutants. The situation is even more complicated by the fact that many tuberculosis patients also have impaired immune system due to AIDS. Over 310,000 cases of multidrug-resistant tuberculosis have been observed worldwide in 2011, with the highest per capita incidence rate in Sub-Saharan Africa, where the incidence rate of AIDS is also the highest (Zumla, et al., 2013). Multidrug-resistant tuberculosis is best prevented by using more than one anti-tuberculosis drug, and not adding another anti-tuberculosis drug to a patient not responding to another; although in the developing countries, the above measures are not always practical. Drug-resistant tuberculosis cases are usually confirmed

by culturing the tuberculosis strains from the patients' sputum, followed by assessing the strain growth under presence of anti-tuberculosis drugs (Frediani, et al., 2014, Schaaf, et al., 2003). This process usually takes some time to complete, and a delay in the diagnosis of the drug-resistant strain may prevent the patient from switching to second-line drugs earlier.

Annual tuberculosis incidence rate in Georgia, a former Soviet republic, exceeds 100 cases per 100,000. The World Health Organization has hence declared Georgia as a country with high disease burden for tuberculosis. Tuberculosis patients in this country are usually of lower social class and often suffer from malnutrition. A pilot study (Frediani, et al., 2013) has previously been conducted in this country to assess the effects of macronutrients on tuberculosis treatment. The behavior of metabolites, including those in the glutamate metabolism pathway, has been identified to be significantly different between drug-resistant and drug-sensitive tuberculosis patients. Vitamin D3 has been identified as a macronutrient that may potentially affect treatment; the intake of Vitamin D3 is thus controlled in this study.

Metabolomics

Metabolomics is the systematic study of metabolites and their interactions. A metabolic profile of a patient's bodily fluids can give a snapshot of the biochemical reactions going on in the body. The patient plasma sample can be analyzed using untargeted high throughput mass spectrometry, a technique that will identify numerous metabolites in the sample (Frediani, et al., 2014; Jones, et al., 2012). The relative concentrations of metabolites, identified through careful biostatistics and bioinformatics analysis, could indicate the activation status of diseases-related biological pathways. However, the identification of metabolites (i.e. metabolite annotation) is not a trivial task; the mass spectrometry experiment only reports the mass/charge (m/z) ratio,

retention time and intensity of an ionized molecule in chromatography, and there could be multiple molecules that share these characteristics. In addition, given the numerous ways the metabolites interact in the human body, the identification of pathways from these metabolites also require considerable effort (Li, et al., 2013).

In this thesis I present a regression method to identify metabolites associated with multiple-drug-resistant tuberculosis using high throughput metabolomics analysis of plasma samples. The study also finds pathways through these identified metabolites.

Chapter II Material and Methods

Data Collection

The metabolite data was collected in a double blind, randomized, controlled trial. 23 multi-drug resistant pulmonary tuberculosis patients were recruited from the Georgia National Center for Tuberculosis and Lung Diseases (NCTBLD) and an affiliated outpatient TB clinic in Tbilisi, Georgia. Each patient is then matched by two drug-sensitive tuberculosis patients (by sex and age ± 15 years) in a case-control manner. One of the matched patients had to be later dropped due to the fact that his tuberculosis cleared at baseline, resulting in a total of 23 multi-drug resistant patients and 45 matched controls. Each patient is then randomly assigned to the Vitamin D_3 (cholecalciferol; 1.4. million IU given in divided doses over 16 weeks) or identical placebo group. The descriptive statistics of patient demographics information is summarized in table 1.

Each patient has their peripheral blood samples taken at 4 time points: week 0 (baseline), week 4, week 8 and week 16. Each blood sample is then centrifuged, and plasma is isolated from the samples. The plasma samples are then frozen to -80 °C and delivered to Emory University, where they are analyzed, in triplicate, with high-resolution liquid chromatography – mass spectrometry (LC-MS), using anion exchange and C18 chromatography (Higgins Analytical, Targa, Mountain View, CA, USA, 2.1×10 cm) combined with the Thermo Orbitrap-Velos (Thermo Fisher, San Diego, CA) mass spectrometer. The mass spectrometry analysis is set to capture all ions with *m/z* ratios between 85 and 850.

Data Pre-processing

The data was pre-processed using xMSAnalyzer (Uppal, et al., 2013) in combination with apLCMS (Yu and Jones, 2014; Yu, et al., 2009). The metabolite concentrations were each

measured in triplicates. I computed the average concentration for each feature, which is computed by averaging the non-zero concentration readings; if all three readings are zero then an average of zero is recorded. This average concentration is the combined metabolite data we used for subsequent analysis. Also, metabolites with excessive (more than 20%) zeros in the averaged readings are removed; 5,715 metabolic features were selected in this way for downstream analysis.

Statistical Analysis

Identifying significant metabolites with cross-sectional analysis

Metabolome-wide association study (MWAS) was used to select significant metabolic features distinguishing multi-drug resistant and drug sensitive patients. We fitted one logistic regression model for each metabolite, and the general modeling formula is as follows:

$$logit(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 log(x_i + 1) + \sum_{k=1}^{K} \gamma_k z_k + \varepsilon$$

Where π is the probability of being a multi-drug resistant tuberculosis patient, x_i is the level of the i^{th} metabolite, and the $z_k's$ are other confounding predictors. The log intensity was used lieu of raw metabolite intensity values to address potential scaling and homoscedasticity issues; the add one part is to address the situation where the intensity is zero.

Other variables are also accounted for in our analysis. As samples analyzed in different batches of the mass spectrometry process may be subjected to different systematic errors, in our analysis, batch effect was added in linearly as a confounder. We also accounted for other demographic confounders such as vitamin D level in plasma, history of diabetes, income level, body mass index, sex age, and whether the patient was randomized to the vitamin D or placebo group.

Raw p-values were used to determine which features were considered significant, using 0.05 as the cutoff threshold.

Identifying significant metabolites using differences between metabolite concentrations

We also attempted to identify significant metabolites using changes in metabolite concentration. This would potentially make the identified metabolites more pronounced, as we are essentially encouraging picking metabolites whose concentrations have changed the most. The modeling statement is as follows:

$$logit(\pi) = \beta_0 + \beta_1(log(x_{i,t1} + 1) - log(x_{i,t0} + 1)) + \sum_{k=1}^{K} \gamma_k z_k + \varepsilon$$

Where π is the probability of being a multi-drug resistant tuberculosis patient, $x_{i,t1}$ and $x_{i,t0}$ are the levels of the i^{th} metabolite at time point 1 and 0, respectively, and the $z_k's$ are other confounding predictors, same as the ones used in cross-sectional analysis. Models were fitted using metabolite levels at week 4 and baseline, and again using metabolite levels at week 8 and baseline.

The metabolites identified from the analyses above are then put through Mummichog (Li, et al., 2013) for metabolite annotation and pathway analysis.

All analysis was carried out using R version 3.1.1, Anaconda version 2.1.0, Python version 2.7.8, and Mummichog version 0.10.3.

Chapter III Results and Discussion

Metabolites Identified

The analysis identified 246, 216, and 229 metabolites as significantly different between multi-drug resistant and drug-sensitive tuberculosis patients, using baseline, week 4, and week 8 samples, respectively. These identified metabolites do not overlap heavily (12 overlaps between baseline and week 4 samples; 7 overlaps between week 4 and week 8 samples; no metabolite is identified in all three groups).

The analysis using concentration differences between week 4 and baseline samples yielded 249 significant metabolites, while the analysis using concentration differences between week 8 and baseline samples yielded 281 significant metabolites. Again, the two sets of significant metabolites do not heavily overlap (only 12); nor do they overlap heavily with metabolites identified through cross-sectional analysis (varying between 9 and 19).

The lack of overlapping metabolites may indicate that the use of *m/z* ratios alone in identifying metabolites is inadequate, and further analysis should be conducted using more advanced metabolite annotation techniques. However, it is possible that lack of overlap is due to underlying metabolic changes as a function of tuberculosis drug resistance.

We then produced Manhattan plots for each of the datasets (Figure 1). We observed apparent clustering of significant metabolites at a retention time between 200 and 400 seconds. One of the most significant metabolites in the Manhattan plots, 1-Pyrroline-2-carboxylate (*m/z* 114.054; retention time 251), has been linked to Mycobacterium tuberculosis as a respiration intermediary metabolite (Yang 2006). We also plotted heatmaps (Figure 2) with hierarchical clustering on the metabolites, showing that the identified metabolites have significant differences between the drug resistant and drug sensitive patient groups.

Metabolite Annotation and Pathway Analysis

The pathway analysis revealed several significant pathways that were altered in subjects with multi-drug resistance tuberculosis at different time points. These included Fatty Acid Metabolism (identified in 4 datasets), Glutamate metabolism (3 datasets), Pyrimidine metabolism (3 datasets) and Tryptophan metabolism (3 datasets). The characteristics of these pathway findings are summarized in Table 2. After further investigation into the top pathways, we found some interesting links to tuberculosis disease. There were several *m/z* matches within the leukotriene system; we found a *m/z* match to 5(S)-HPETE, LTB4 and 6-trans LTB4 to be lower in multi-drug resistant vs. drug-sensitive patients and 10,11 dihydro-12, oxo-LTB4 and 6E-12 epi-LTB4 to be upregulated. This could suggest the multi-drug resistant subjects may have suppressed immune systems and poor initiation of cell mediated immunity. (Tobin 2012) Furthermore, excess LXA4 and LTB4 can promote extracellular bacterial growth. (Tobin 2013) Metabolites related to glutamate metabolism were all increased in multi-drug resistant tuberculosis patients, indicating that tuberculosis drug susceptibility may alter glutamate metabolism. Metabolites related to retinol metabolism were higher in MDR-TB subjects. The theory behind these trials is related to vitamin A's antioxidant and anti-inflammatory properties (Wheelwright 2014).

Constraints and future directions

This study has certain constraints. Due to the difficulty in recruiting patients, this study has a relatively small sample size, which could limit the power of statistical testing. In fact, no metabolites were found to be significant using false discovery rates, which could be potentially

attributed to the small sample size. Also, not all patients have blood samples taken at all intervals; several patients only had metabolite measurements at baseline, and this may impede analyses using concentration differences. Some of the patient demographics data is also missing. Future work should be focused on getting a larger cohort of patients and thus enabling a more statistically rigorous interpretation of the metabolomics results.

Chapter IV Conclusion

This study shows that tuberculosis drug susceptibility can be identified using metabolome-wide association study techniques from high throughput metabolite profiling of plasma, and presents metabolites correlated with drug susceptibility. This study also identifies several metabolic pathways that are differently regulated between the two groups, but given the small sample size, more mechanistic studies of these pathways are needed.

Appendix

Tables

	All patients (n=68)	Drug resistant (n=23)	Drug sensitive (n=45)
Age (yr)	34 (10.81)	34 (9.44)	34 (11.56)
Sex			
Male	37 (54.41)	13 (56.52)	24 (53.33)
Female	29 (42.65)	10 (43.48)	19 (42.22)
Unknown	2 (2.94)	0 (0)	2(4.44)
Income (1000 lari ≈ 600USD)			
<1000 lari	27 (39.71)	5 (21.74)	22 (48.89)
1000-5000 lari	28 (41.18)	14 (60.87)	14 (31.11)
5001-10,000 lari	11 (16.18)	4 (17.39)	7 (15.56)
Unknown	2 (2.94)	0 (0)	2 (4.44)
BMI (kg/m^2)	20.53 (3.69)	20.26 (2.54)	20.67 (4.19)
Baseline Vitamin D (ng/mL)	14.13 (7.73)	13.43 (6.82)	14.51 (8.23)
Diabetes			
Yes	4 (5.88)	1 (4.35)	3 (4.67)
No	62 (91.17)	22 (95.65)	40 (88.89)
Unknown	2 (2.94)	0 (0)	2(4.44)

Table 1: summary of patient demographics, grouped by drug susceptibility. For continuous variables, the mean is followed by standard deviation in parenthesis; for categorical variables, the count is followed by percentage in parenthesis.

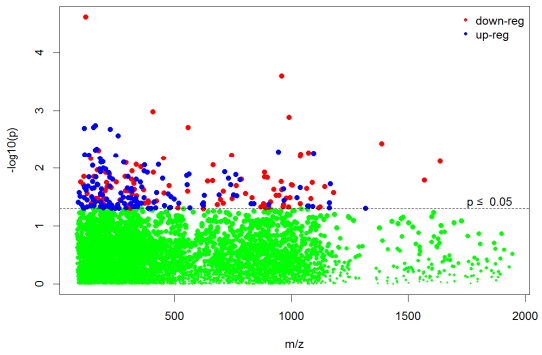
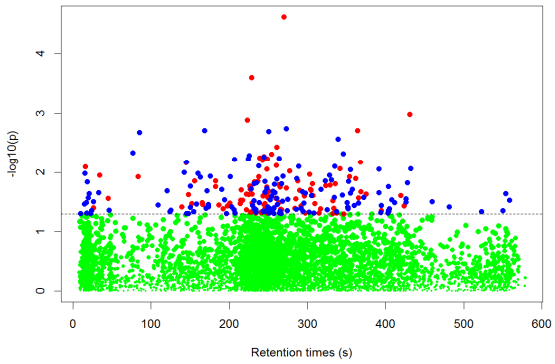
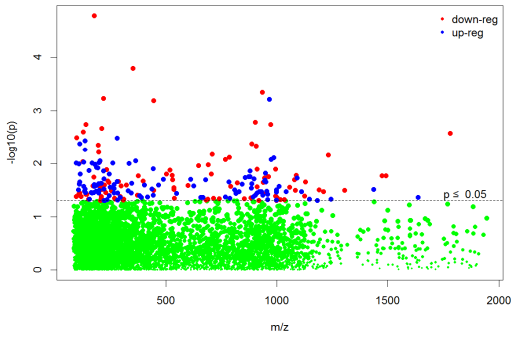
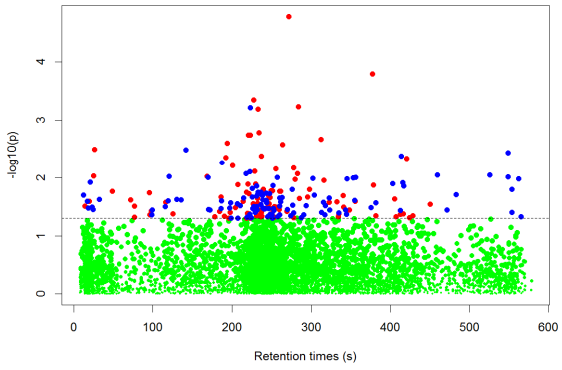
Dataset	**Pathways**	**Significant Metabolites /Total Metabolites**	**Adjusted p-value**
Baseline	Leukotriene metabolism	5 / 40	0.04494
	Glutamate metabolism	2 / 8	0.05194
	Vitamin A (retinol) metabolism	4 / 32	0.06481
	Drug metabolism - cytochrome P450	4 / 33	0.07314
	Pyrimidine metabolism	3 / 25	0.11963
Week 4	Saturated fatty acids beta-oxidation	4 / 26	0.00181
	Fatty acid activation	4 / 29	0.00215
	Histidine metabolism	3 / 17	0.00261
	Urea cycle/amino group metabolism	4 / 38	0.00389
	Fatty Acid Metabolism	3 / 23	0.00461
	Beta-Alanine metabolism	2 / 11	0.00879
	Tryptophan metabolism	4 / 50	0.00922

	Fatty acid oxidation	2 / 14	0.01375
	Methionine and cysteine metabolism	3 / 35	0.01414
	Lysine metabolism	2 / 24	0.04383
	Pyrimidine metabolism	2 / 25	0.04811
Week 8	Nitrogen metabolism	2 / 3	0.00335
	Vitamin B6 (pyridoxine) metabolism	2 / 4	0.00528
	Glycosphingolipid biosynthesis - globoseries	2 / 4	0.00528
	Keratan sulfate degradation	2 / 4	0.00528
	De novo fatty acid biosynthesis	6 / 40	0.0058
	Glycosphingolipid metabolism	4 / 23	0.00757
	Fatty Acid Metabolism	4 / 23	0.00757
	N-Glycan Degradation	2 / 5	0.008
	Pyrimidine metabolism	4 / 25	0.01077
	Glycosphingolipid biosynthesis - ganglioseries	2 / 6	0.01167
	Xenobiotics metabolism	6 / 47	0.01443
	Phosphatidylinositol phosphate metabolism	3 / 18	0.01936
	Tryptophan metabolism	6 / 50	0.02106
	Glutamate metabolism	2 / 8	0.0226
	Polyunsaturated fatty acid biosynthesis	2 / 8	0.0226
	Aminosugars metabolism	3 / 19	0.02364
	N-Glycan biosynthesis	2 / 9	0.03014
	Vitamin B9 (folate) metabolism	2 / 10	0.03921
	Sialic acid metabolism	3 / 22	0.04104
	Tyrosine metabolism	7 / 68	0.04711
	Butanoate metabolism	3 / 23	0.04854
	Omega-6 fatty acid metabolism	2 / 11	0.04989
Week_4 - Baseline	Drug metabolism - cytochrome P450	7 / 33	0.00092
	Glutamate metabolism	3 / 8	0.00145
	Glycosphingolipid metabolism	4 / 23	0.00468
	Fatty Acid Metabolism	4 / 23	0.00468
	Methionine and cysteine metabolism	5 / 35	0.00652
	Vitamin B3 (nicotinate and nicotinamide) metabolism	3 / 16	0.00738
	Histidine metabolism	3 / 17	0.00893
	Urea cycle/amino group metabolism	5 / 38	0.00958
	Tryptophan metabolism	6 / 50	0.01158
	Vitamin B9 (folate) metabolism	2 / 10	0.0203
	Beta-Alanine metabolism	2 / 11	0.02531
	Selenoamino acid metabolism	2 / 13	0.03746

	Alanine and Aspartate Metabolism	2 / 13	0.03746
	Saturated fatty acids beta-oxidation	3 / 26	0.03768
	Purine metabolism	3 / 28	0.04849
Week_8 - Baseline	Vitamin H (biotin) metabolism	2 / 3	0.00606
	Aminosugars metabolism	4 / 19	0.01241
	Fatty Acid Metabolism	4 / 23	0.02905
	Lysine metabolism	4 / 24	0.03532
	Polyunsaturated fatty acid biosynthesis	2 / 8	0.04917

Table 2: pathways associated with MDR.

Figures

Dataset	*m/z* manhattan plot	Retention time plot
Baseline		
Week 4		

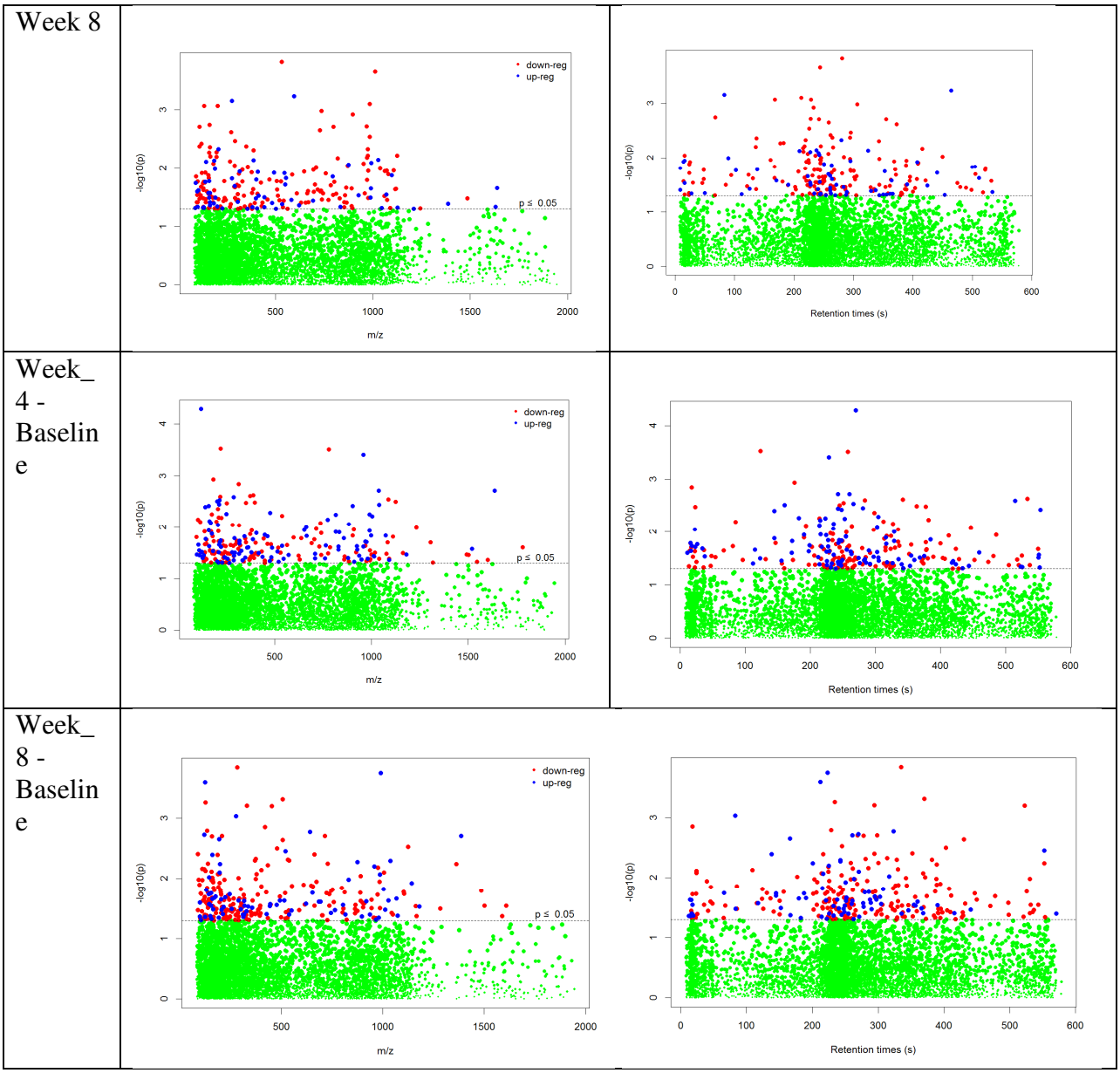
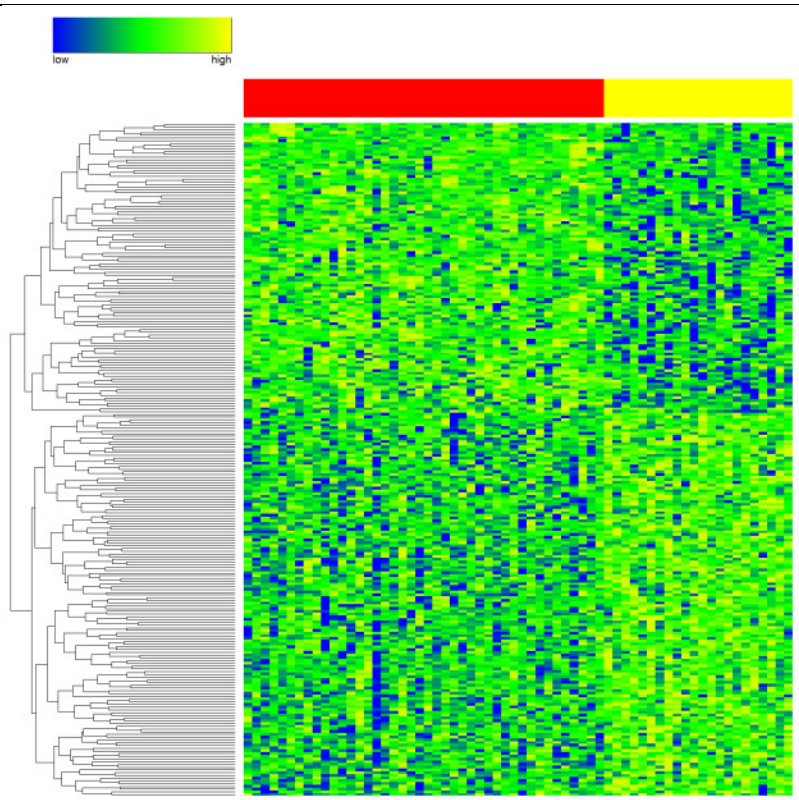
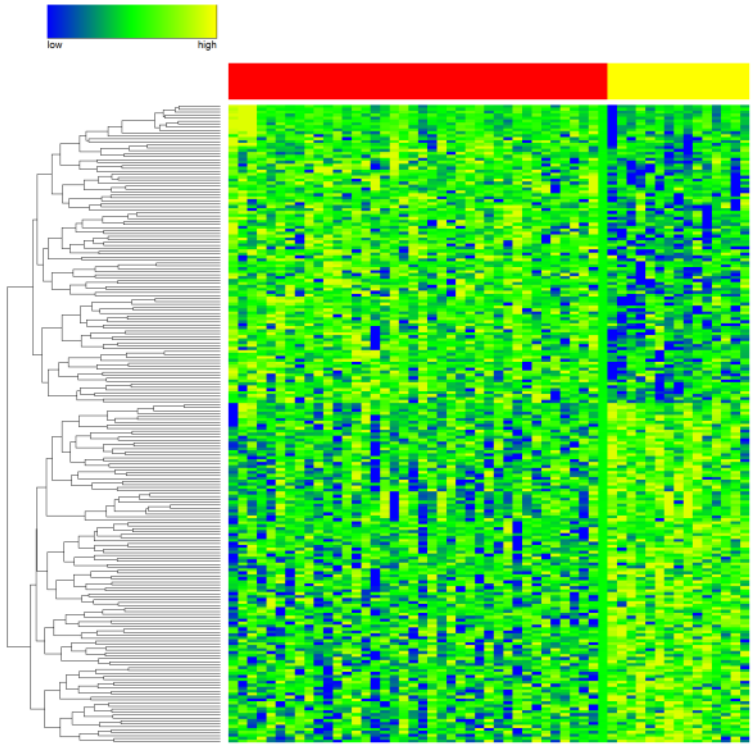
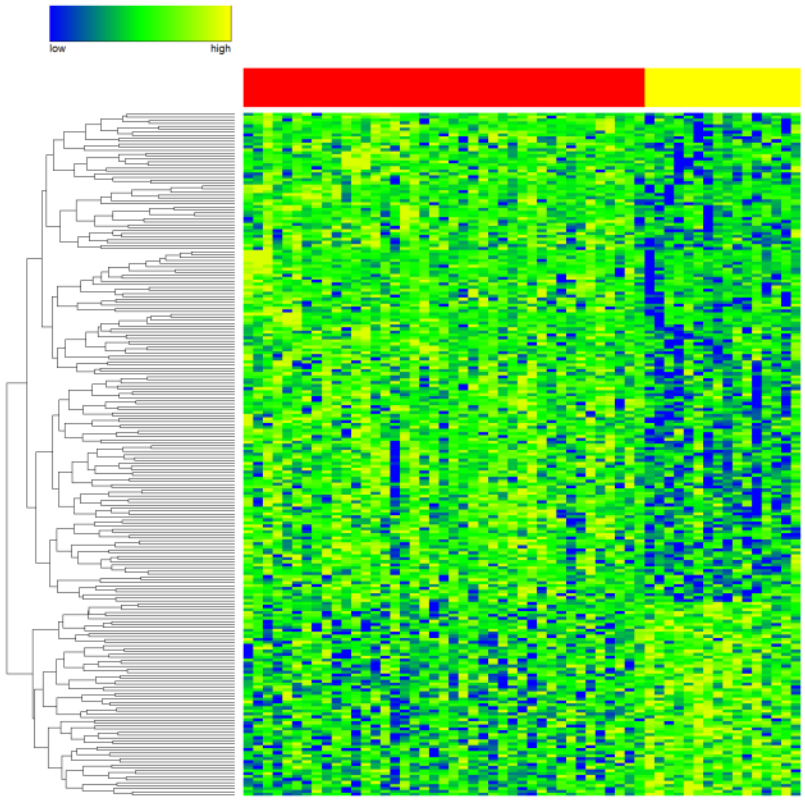
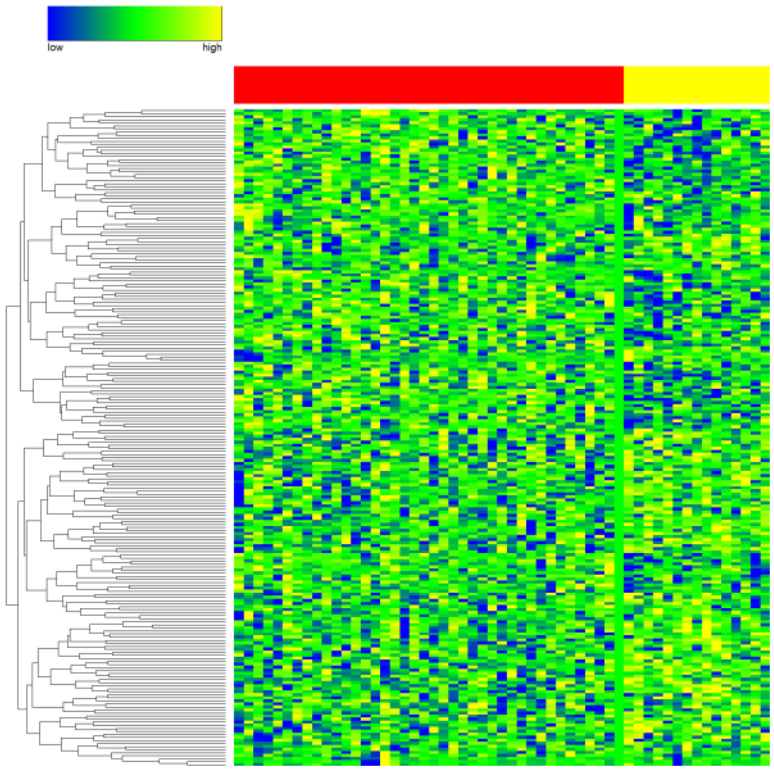


Figure 1: Manhattan plots for all 5 datasets

Baseline	
Week 4	

Week 8	
Week_4 – Baseline	

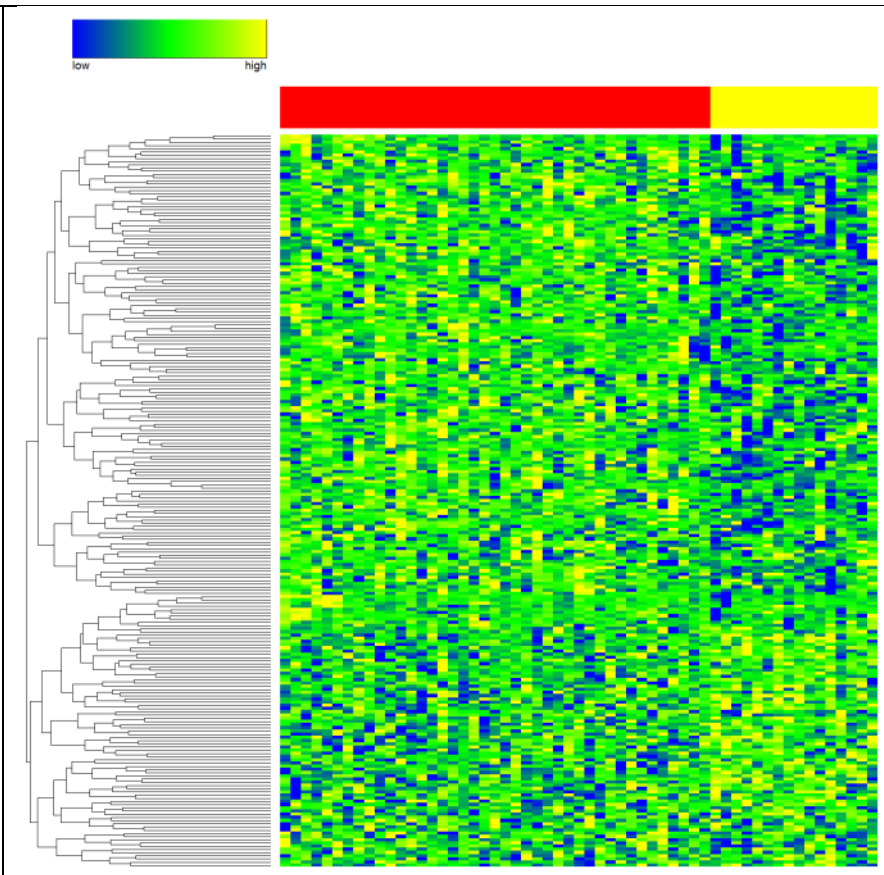
Week 8 – Baseline	

Figure 2: Heatmap for all 5 datasets. The red/yellow bar on top of the heatmaps indicates patient drug sensitivity, with red for drug sensitive and yellow for multi-drug resistant.

References

Frediani JK, Jones DP, et al. Plasma Metabolomics in Human Pulmonary Tuberculosis Disease: A Pilot Study. PlosOne. October 15, 2014. DOI: 10.1371/journal.pone.0108854

Frediani JK, Sanikidze E, et al. Macronutrient intake and body composition changes during anti-tuberculosis therapy in adults, Clinical Nutrition, Available online 26 February 2015

Frediani JK, Tukvadze N, et al. A culture-specific nutrient intake assessment instrument in patients with pulmonary tuberculosis, Clinical Nutrition, Volume 32, Issue 6, December 2013, Pages 1023-1028

Jones DP, Park Y, Ziegler TR. Nutritional metabolomics: Progress in addressing complexity in diet and health. Annual review of nutrition. 2012;32:183-202. doi:10.1146/annurev-nutr-072610-145159.

Li S, Park Y, Duraisingham S, et al. Predicting Network Activity from High Throughput Metabolomics. Ouzounis CA, ed. PLoS Computational Biology. 2013;9(7):e1003123. doi:10.1371/journal.pcbi.1003123.

Schaaf H, Shean K, Donald P. Culture confirmed multidrug resistant tuberculosis: diagnostic delay, clinical features, and outcome. Archives of Disease in Childhood. 2003;88(12):1106-1111.

Tobin DM, Roca FJ, Oh SF, McFarland R, Vickery TW, Ray JF, Ko DC, Zou Y, Bang ND, Chau TTH, Vary JC, Hawn TR, Dunstan SJ, Farrar JJ, Thwaites GE, King MC, Serhan CN, Ramakrishnan L. Host genotype-specific therapies can optimize the inflammatory response to mycobacterial infections. Cell. 2012; 148(3): 434-446.

Tobin DM, Roca FJ, Ray JF, Ko DC, Ramakrishnan L. An enzyme that inactivates the inflammatory mediator leukotriene B4 restricts mycobacterial infection. PLOS One. 2013; 8(7): e67828.

Uppal K, Soltow QA, et al. xMSanalyzer: automated pipeline for improved feature detection and downstream analysis of large-scale, non-targeted metabolomics data. BMC Bioinformatics 2013, 14:15

Wheelwright M, Kim EW, Inkeles MS, De Leon A, Pellegrini M, Krutzik SR, Liu PT. All-trans retinoic acid-triggered antimicrobial activity against Mycobacterium tuberculosis is dependent on NPC2. J Immunol. 2014; 192: 2280-2290.

Yang Y, Xu S, Zhang M, Jin R, Zhang L, Bao J, Wang H. Purification and characterization of a functionally active Mycobacterium tuberculosis pyrroline-5 carboxylate reductase. Protein Expression and Purification. 2006; 45: 241-248.

Yu T, Jones DP. Improving peak detection in high-resolution LC/MS metabolomics data using preexisting knowledge and machine learning approach. Bioinformatics.2014 30(20): 2941-2948.

Yu T, Park Y, apLCMS--adaptive processing of high-resolution LC/MS data. Johnson JM, Jones DP. Bioinformatics. 2009 Aug 1;25(15):1930-6.

Zumla A, Raviglione M, et al. Tuberculosis. N Engl J Med 2013; 368:745-755