See Publication: *Molecular Neurodegeneration*

Untargeted serum metabolomics reveals novel metabolite associations and disruptions in amino acid and lipid metabolism in Parkinson's disease

Kimberly C Paul[1], PhD, Keren Zhang[2], BS, Douglas I Walker[3], PhD, Janet Sinsheimer[4,5], PhD, Yu Yu[6], PhD, Cynthia Kusters[4], PhD, Irish Del Rosario[2], BS, Aline Duarte Folle[2], PhD, Adrienne M Keener[1,7], MD, Jeff Bronstein[1], MD, PhD, Dean P Jones[8], PhD, Beate Ritz[1,2] MD, PhD

[1]Department of Neurology, UCLA David Geffen School of Medicine, Los Angeles, California, USA
[2]Department of Epidemiology, UCLA Fielding School of Public Health, Los Angeles, California, USA
[3]Gangarosa Department of Environmental Health, Rollins School of Public Health, Emory University, Atlanta, Georgia, USA
[4]Department of Human Genetics, UCLA David Geffen School of Medicine, Los Angeles, California, USA
[5]Department of Biostatistics, UCLA Fielding School of Public Health, Los Angeles, California, USA
[6]Center for Health Policy Research, UCLA Fielding School of Public Health, Los Angeles, California, USA
[7]Parkinson's Disease Research, Education, and Clinical Center, Greater Los Angeles Veterans Affairs Medical Center, Los Angeles, California, USA
[8]Department of Medicine, School of Medicine, Emory University, Atlanta, Georgia, USA

Study Population

We provide untargeted serum metabolomic profiles from 642 Parkinson's disease (PD) patients and 277 controls recruited as part of a community-based study of Parkinson's disease (Parkinson's Environment and Genes study, PEG). PEG is a population-based PD case-control study conducted in three Central California counties[1]. Participants were recruited in two, independent study waves: PEG1, 2000-2007 and PEG2, 2011-2018. All those with serum for metabolomics were included (PEG1: n=282 PD patients, n=185 controls; PEG2: n=360 PD patients, n=90 controls). Patients were early in disease course at enrollment (3.0 years [SD=2.6] on average from diagnosis) and all were seen by UCLA movement disorder specialists for in-person neurologic exams and confirmed as having idiopathic PD based on clinical characteristics[2]. Characteristics of the PEG study subjects are shown in Supplemental Table 1. The patients were on average slightly older than the controls and a higher proportion of the patients were men, Hispanic, and never smokers compared to the controls.

Sample Collection

Blood samples were drawn from participants during field visits. Samples were centrifuged, kept on dry ice, and then stored in a −80 °C freezer at UCLA. Serum samples were shipped frozen to Emory University on dry ice for metabolomics analyses, where they were stored at −80 °C until analyses.

High-Resolution Metabolomics (HRM)

HRM profiling was conducted according to established methods. Detailed methods are provided in the accompanying file EmoryUniversity_SOP_DataAnalysis_092017_v1.pdf. Briefly, serum samples were randomly sorted into batches of 40. Each sample was thoroughly mixed with ice-cold acetonitrile (2:1 acetonitrile to serum), placed on ice for 30 minutes, precipitated protein was removed by centrifugation, and the resulting supernatant was transferred to an autosampler vial containing a low volume insert. We analyzed all sample

extracts in triplicate with a dual-column, dual-polarity approach, including hydrophilic interaction (HILIC) chromatography with positive electrospray ionization (ESI) and C18 chromatography with negative ESI, and used two types of quality control samples. We included two methods of performance quality control. First, a NIST 1950 QC sample was analyzed at the beginning and end of the entire analytical run. A second QC sample (Q-Std), which is commercially purchased plasma pooled from an unknown number of men and women, was analyzed at the beginning, middle, and end of each batch of 40 samples for normalization and batch effect evaluation (n=180 Q-Std samples total included).

The Emory metabolomics lab uses a quality control procedure based on XCMS and a set of confirmed metabolites and internal standards to evaluate the data quality of each batch: number of features detected, missing values, mass accuracy (threshold <5 ppm), Pearson correlation within technical replicates (threshold: 0.9), and average coefficient of variation (CV) of feature intensities within replicates (threshold: <30%). Samples were re-analyzed if the data did not meet the defined criteria.

Our samples were processed across two LC-HRMS runs conducted approximately 6-months apart, to pool the metabolite data across runs, we used the *apLCMS* R package to perform retention time adjustment and feature alignment for both HILIC and C18 feature tables, using the adjust.time and feature.align functions[3]. For feature alignment, the *m*/*z* tolerance was 1e-05 and retention time tolerance was 37.016 (C18) and 38.246 (HILIC) seconds. Overall, 2226 features aligned for C18 and 2919 for HILIC across the two LCMS runs. For analyses, we included metabolomic features with median CV among technical replicates <30% and Pearson correlation >0.9 and features detected in >50% of all study samples, leaving 2046 C18 features and 2716 HILIC features for analysis.

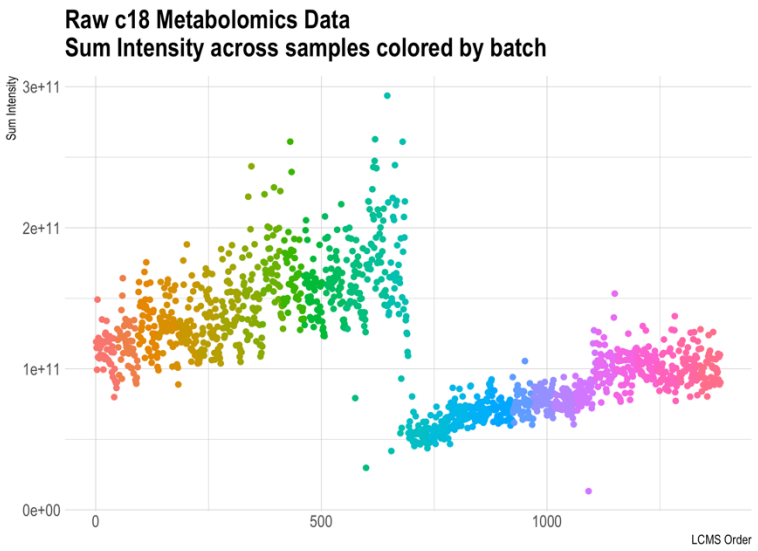
We log 2 transformed the metabolite data, quantile normalized, and batch corrected with ComBat after replacing zeroes with the lowest detected value which has been recommended for metabolomics data. Data pre-processing visualization is shown in Supplemental Figures 1-4. From principal component (PC) analysis with the HILIC features, we discovered two clusters of samples seemingly separating based on technical, non-biologic factors. As a result, we performed an additional correction to remove variation between the PCs (Supplemental Figures 5-7). This was done with ComBat, using an indicator for whether the sample was part of the outlying cluster as the correction term.

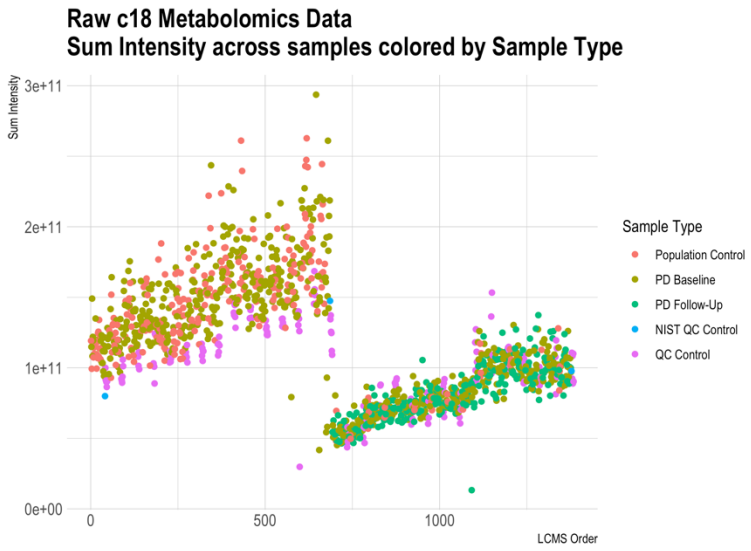
Within the Q-Std samples across both runs and all batches (n=180), the mean CV across all C18 metabolite features before the data processing steps was 157.1% (median=75.2%, IQR=127.1%) but after the processing steps it reduced to 7.2% (median=6.3%, IQR=5.5%). For HILIC features, the mean CV before processing was 148.0% (median=69.3%, IQR=128.0%) and after the processing steps 8.7% (median=8.0%, IQR=8.3%).

References:

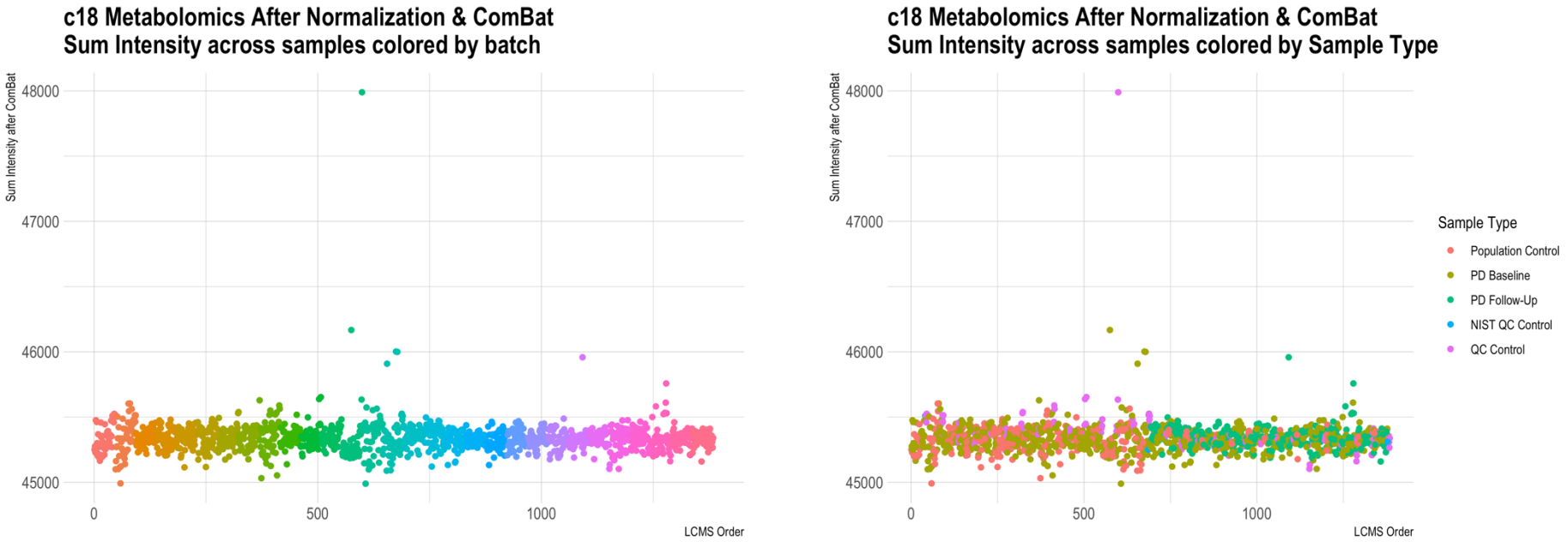
1. Ritz BR, Paul KC, Bronstein JM. Of Pesticides and Men: a California Story of Genes and Environment in Parkinson's Disease. *Curr Environ Health Rep*. 2016;**3**(1).
2. Hughes AJ, Ben-Shlomo Y, Daniel SE, Lees a J. What features improve the accuracy of clinical diagnosis in Parkinson's disease: a clinicopathologic study. *Neurology*. 1992;**42**(6):1142–1146.
3. Yu T, Park Y, Johnson JM, Jones DP. apLCMS-adaptive processing of high-resolution LC/MS data. *Bioinformatics*. 2009;**25**(15).

Supplemental Figures

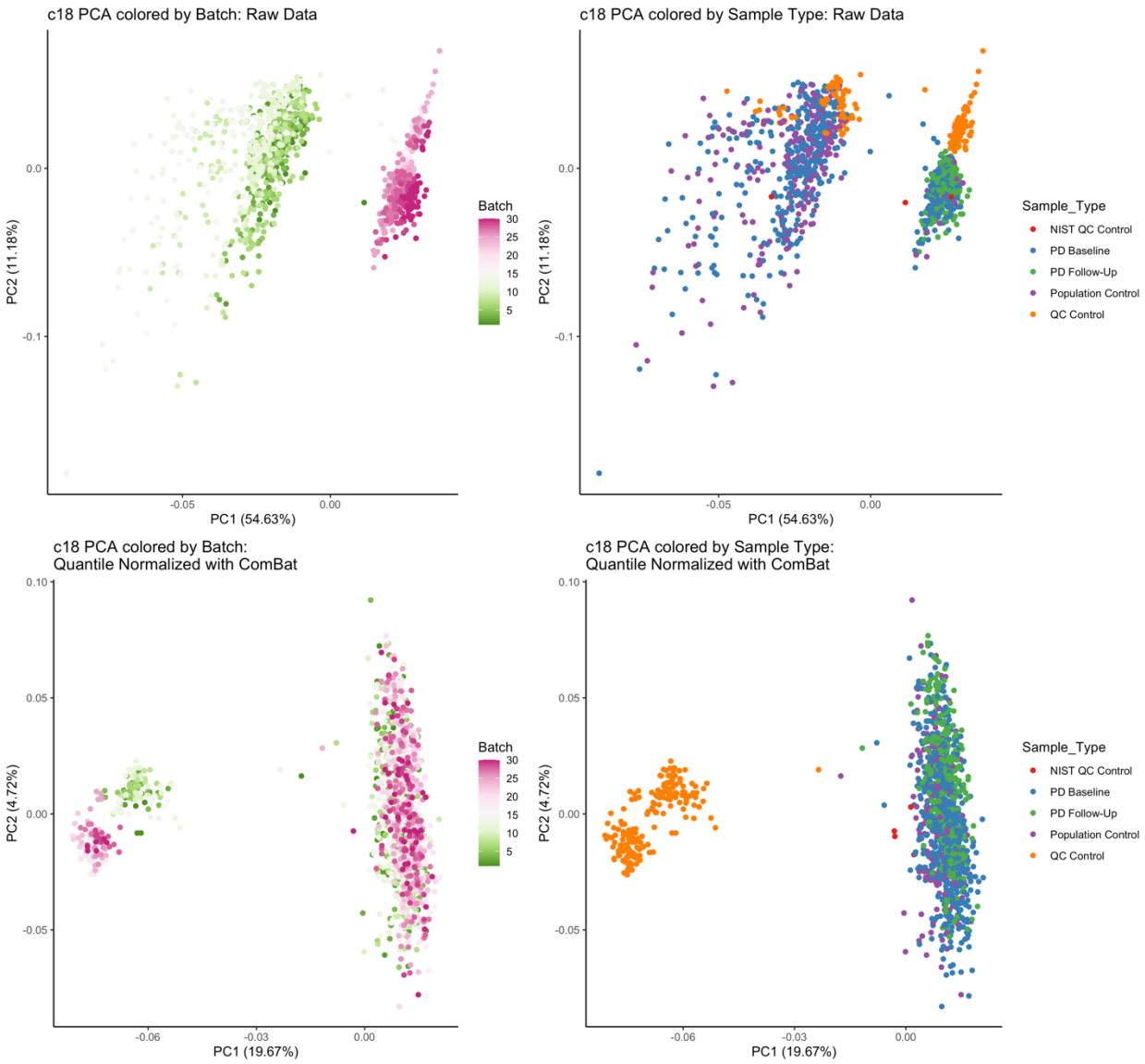




Supplemental Figure 1. C18 negative column metabolomics processing: Sum of metabolite intensities across samples colored by batch & sample type, before pre-processing (log transformation, quantile normalization, ComBat batch correction). LCMS was run across 30 batches (n=46); machine was reset after 694 samples (i.e., samples ran in two larger groups of n=694 samples, each with 15 smaller batches within run). Run, batch, and drift effects are apparent in raw data.



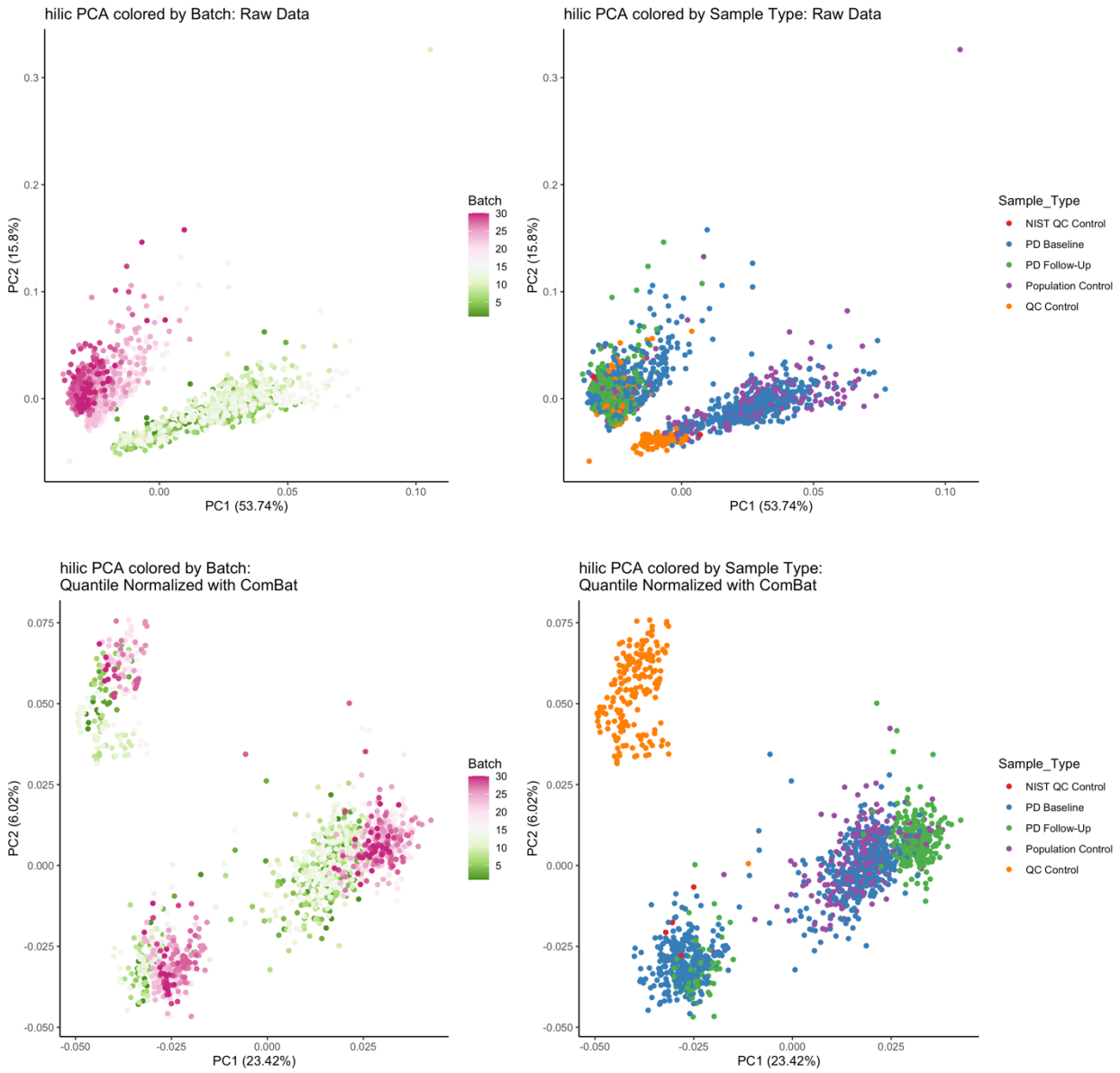
Supplemental Figure 2. C18 negative column after metabolomics processing. Raw c18 data was log transformation, quantile normalized, followed by ComBat for batch correction. LCMS was run across 30 batches (n=46); machine was reset after 694 samples (i.e., samples ran in two larger groups of n=694 samples, each with 15 smaller batches within run. While there are several apparent outliers, after processing, technical variation has been removed.



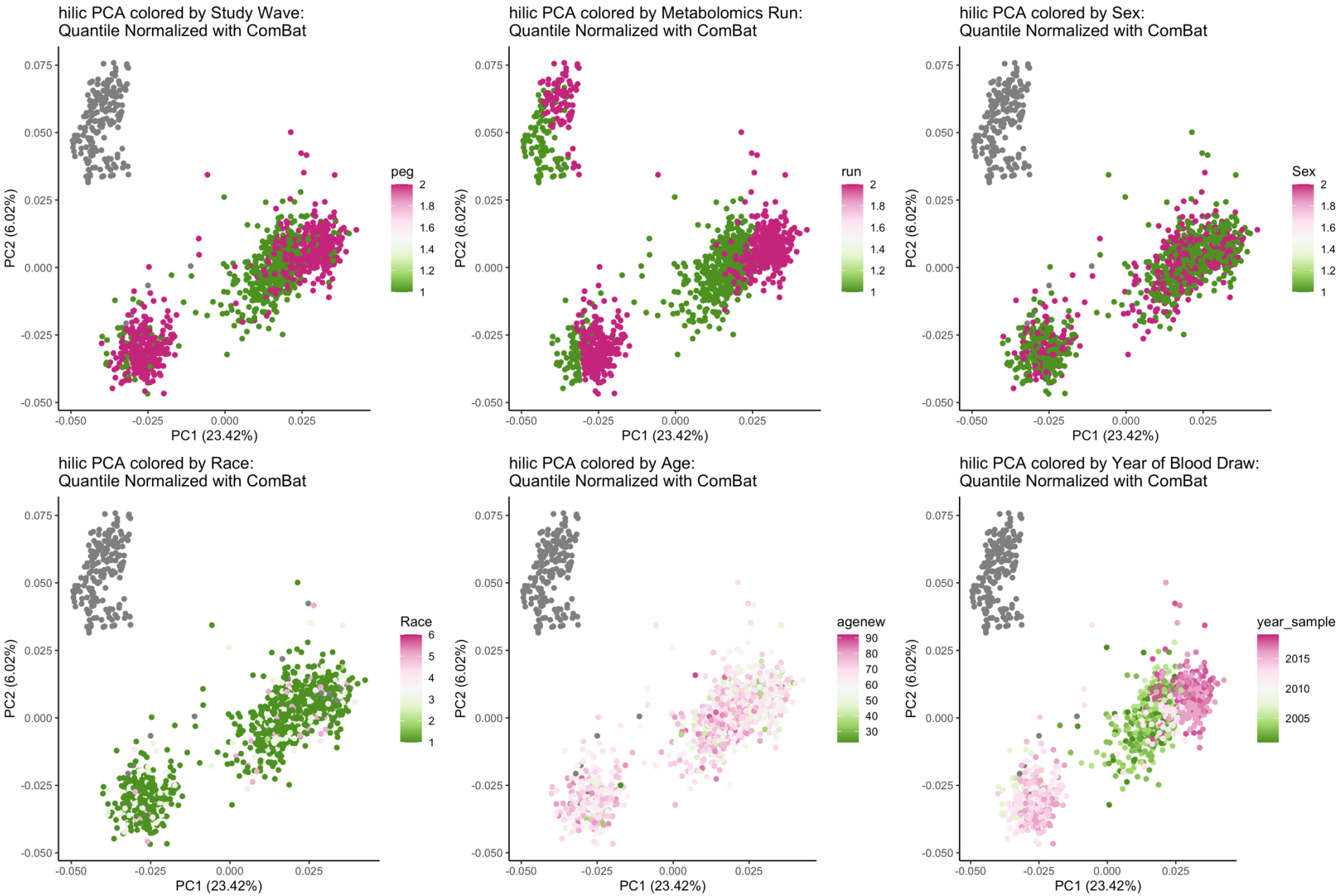
Supplemental Figure 3. C18 negative column metabolomics processing: Principal component analysis of raw and processed metabolomics data. PC variation primarily explained by LCMS run in raw data. After correction, sample type (quality control sample versus the study serum samples) primarily explains variation.

Supplemental Figure 4. HILIC positive column metabolomics processing: Sum of metabolite intensities across samples colored by batch & sample type before and after pre-processing (log transformation, quantile normalization, ComBat batch correction). LCMS ran in across 30 batches (n=46); machine was reset after 694 samples (i.e., samples ran in two larger groups of n=694 samples, each with 15 smaller batches within run). Run,

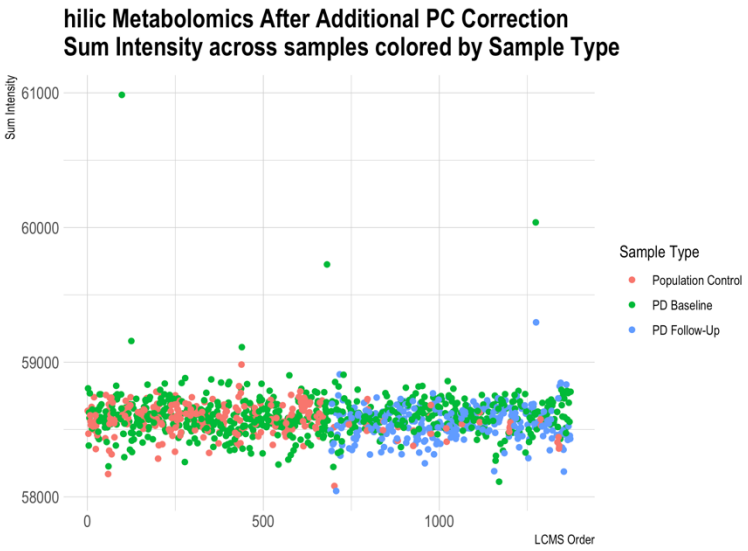
batch, and drift effects are apparent in raw data. While there are several apparent outliers, after processing, the technical variation has been removed.

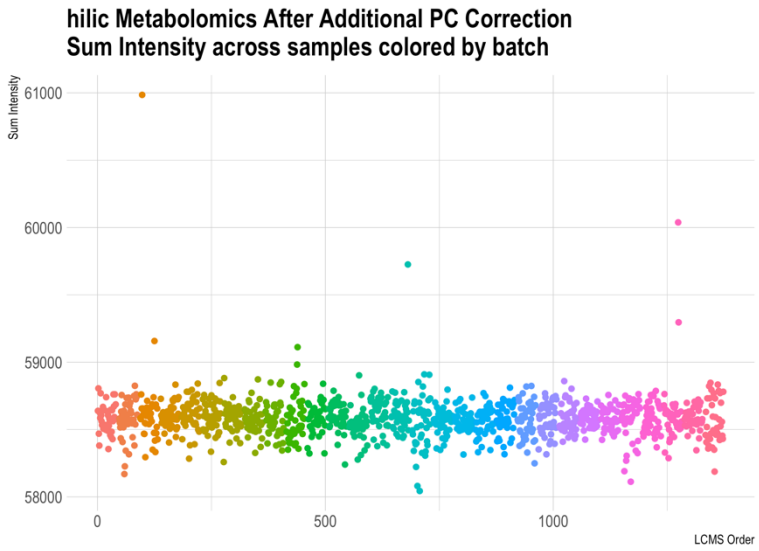


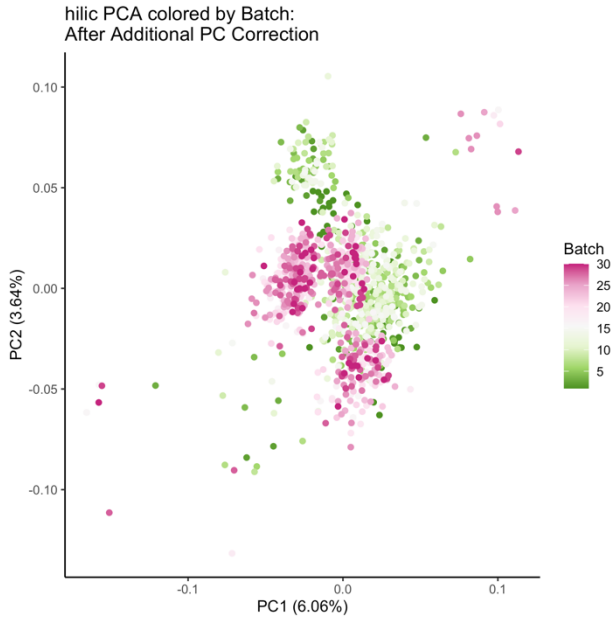
Supplemental Figure 5. HILIC positive column metabolomics processing: Principal component analysis of metabolomics data after median normalization and ComBat correction for batch effects. PC variation primarily explained by batch in raw data, after correction sample type (quality control sample versus the population-based serum samples) primarily explains variation. However, there are two apparent clusters of population-based serum samples, potentially explained by non-biologic (PD) technical variation (see **Supplemental Figure 6**).

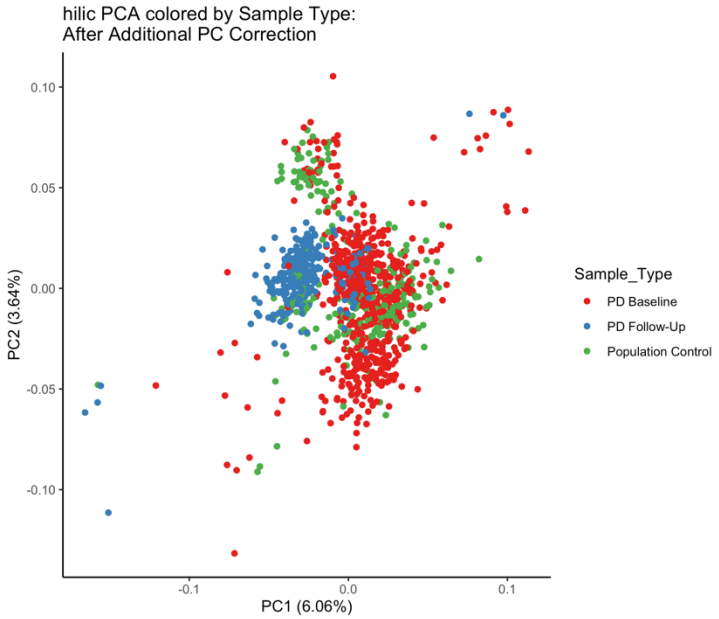


Supplemental Figure 6. HILIC positive PCA of processed data, colored by different covariates. No distinguishing variables to describe the different clusters of study samples, though there is some separation by year of sample. Note gray indicates the QC samples. Therefore, we additionally corrected for inclusion in this cluster, as variation appears technical and is very influential in MWAS (Supplemental Figure 7).









Supplemental Figure 7. HILIC positive metabolomics data after processing: Log transformation, quantile normalization, ComBat batch correction, and additional adjustment for unexplained PC.