The 2012/2013 ABRF Proteomic Research Group Study: Assessing Longitudinal Intralaboratory Variability in Routine Peptide Liquid Chromatography Tandem Mass Spectrometry Analyses*[S]

Keiryn L. Bennett‡§§§, Xia Wang§, Cory E. Bystrom¶, Matthew C. Chambers‖, Tracy M. Andacht, Larry J. Dangott‡‡, Félix Elortza§§, John Leszyk¶¶, Henrik Molina‖‖, Robert L. Moritz[a], Brett S. Phinney[b], J. Will Thompson[c], Maureen K. Bunger§§§[d,e], and David L. Tabb‖§§§**

Questions concerning longitudinal data quality and reproducibility of proteomic laboratories spurred the Protein Research Group of the Association of Biomolecular Resource Facilities (ABRF-PRG) to design a study to systematically assess the reproducibility of proteomic laboratories over an extended period of time. Developed as an open study, initially 64 participants were recruited from the broader mass spectrometry community to analyze provided aliquots of a six bovine protein tryptic digest mixture every month for a period of nine months. Data were uploaded to a central repository, and the operators answered an accompanying survey. Ultimately, 45 laboratories submitted a minimum of eight LC-MSMS raw data files collected in data-dependent acquisition (DDA) mode. No standard operating procedures were enforced; rather the participants were encouraged to analyze the samples according to usual practices in the laboratory. Unlike previous studies, this investigation was not designed to compare laboratories or instrument configuration, but rather to assess the temporal intralaboratory reproducibility. The outcome of the study was reassuring with 80% of the participating laboratories performing analyses at a medium to high level of reproducibility and quality over the 9-month period. For the groups that had one or more outlying experiments, the major contributing factor that correlated to the survey data was the performance of preventative maintenance prior to the LC-MSMS analyses. Thus, the Protein Research Group of the Association of Biomolecular Resource Facilities recommends that laboratories closely scrutinize the quality control data following such events. Additionally, improved quality control recording is imperative. This longitudinal study provides evidence that mass spectrometry-based proteomics is reproducible. When quality control measures are strictly adhered to, such reproducibility is comparable among many disparate groups. Data from the study are available via ProteomeXchange under the accession code PXD002114. ***Molecular & Cellular Proteomics 14: 10.1074/mcp.O115.051888, 3299–3309, 2015.***

From the ‡CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences, Vienna, Austria; §University of Cincinnati, Department of Mathematical Sciences, University of Cincinnati, Cincinnati, Ohio, 45221-0025; ¶Cleveland HeartLab, Inc., Research and Development, Cleveland HeartLab, Inc., Cleveland, Ohio, 44103; ‖Vanderbilt University, Department of Biomedical Informatics, Vanderbilt University, Nashville, Tennessee, 37232; **Centers for Disease Control and Prevention, Emergency Response Branch, Division of Laboratory Sciences, National Center for Environmental Health, Centers for Disease Control and Prevention, Atlanta, Georgia, 30341; ‡‡Texas A&M University, Department of Biochemistry & Biophysics, Texas A&M University, College Station, Texas, 77843; §§CIC bioGUNE, Centro de Investigacion Cooperativa en Biociencias, ProteoRed-ISCIII, Bilbao, Spain; ¶¶University of Massachusetts, Department of Biochemistry and Molecular Pharmacology Proteomics and Mass Spectrometry Facility, University of Massachusetts Medical School, Shrewsbury, Massachusetts, 01545; ‖‖The Rockefeller University, Proteomics Resource Center, The Rockefeller University, New York, New York, 10065; [a]Institute for Systems Biology, Seattle, Washington, 98109; [b]University of California, Davis, Proteomics Core, University of California-Davis Genome Center, Davis, California, 95616; [c]Duke University, Proteomics and Metabolomics Core Facility, Duke University Medical Center, Durham, North Carolina, 27708; [d]Proteovations, LLC, RTP, North Carolina 27709

Received June 1, 2015, and in revised form, September 25, 2015
Published, MCP Papers in Press, October 4, 2015, DOI 10.1074/mcp.O115.051888

Author contributions: K.L.B., C.E.B., T.M.A., L.J.D., F.E., J.L., H.M., R.L.M., B.S.P., J.W.T., and M.K.B. designed research; K.L.B., X.W., M.C.C., H.M., B.S.P., and D.L.T. performed research; X.W., M.C.C., and D.L.T. contributed new reagents or analytic tools; K.L.B., X.W., C.E.B., M.C.C., J.W.T., and D.L.T. analyzed data; K.L.B., X.W., C.E.B., and D.L.T. wrote the paper; K.L.B. coordinated the manuscript with all other authors; T.M.A. contributed to accompanying survey questions; L.J.D. ombudsman/coordinator of the study; F.E. proof-read and contributed to manuscript; H.M. created and coordinated survey questions; R.L.M. contributed text to the manuscript; M.K.B. coordinator/chair of the abrf-prg.

The broad-reaching use and application of mass spectrometry-based proteomics in the international research commu-

nity continues to exponentially grow and expand. As the technology has developed and practitioners have become skilled in performing complex workflows, the community has not only gained interest in assessing data across laboratories but also in maintaining consistent quality control within a laboratory. Koecher *et al.* raised the issue of quality control measures and how this aspect of mass spectrometry-based proteomics is generally neglected in scientific publications (1). Fortunately, studies characterizing the stability of liquid chromatography-tandem MS (LC-MSMS)[1] quality control performance among numerous laboratories are emerging. The relationship between sample preparation schemes, data acquisition and reduction strategies, and bioinformatic analyses have been comprehensively reviewed by Tabb (2).

Several studies exist where intra- and interlaboratory reproducibility between multiple sites has been assessed under different settings. Perhaps the most systematic and detailed of these investigations are from the Human Proteome Organization (HuPO) test sample working group (3); the National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (NCI CPTAC) (4); and the ProteoRed Consortium (5, 6). The HuPO group utilized an equimolar mixture of 20 highly purified recombinant human proteins (5 pmol per protein) distributed to 27 different laboratories and analyzed without constraint according to optimized LCMS and database search protocols from each of the laboratories (3). The study was not an assessment of instrument performance for highly sensitive detection of proteins, as all participating laboratories had acquired raw data of sufficient quality to identify all 20 proteins (and a specific subset of tryptic peptides). The study revealed, however, that discrepancies in peptide identification and protein assignment were the result of differences in data analysis strategies rather than data collection.

The NCI CPTAC group used a standardized *Saccharomyces cerevisiae* proteome digest that was analyzed on ion-trap-based LCMS platforms in five independent laboratories according to both an established standard operating procedure (SOP) and with no SOP constraint (4). All data analysis was centralized, and thus, any observed variations were entirely because of the LCMS platform. By applying the performance metrics developed by Rudnick *et al.* (7), several key points emerged: (1) as expected, intralaboratory variation was less than interlaboratory variation; and (2) overall, the interlaboratory variation in peptide identifications and some of the other performance metrics were comparable between instruments, although there were large differences in the average values for some metrics (*e.g.* MS1 signal intensity, dynamic sampling).

The ProteoRed Consortium initiated the ProteoRed Multicenter Experiment for Quality Control (PMEQC) (5, 6). This longitudinal QC multicenter study involved 12 institutes, and was designed to assess: (1) intralaboratory repeatability of LC-MSMS proteomic data; (2) interlaboratory reproducibility; and (3) reproducibility across multiple instrument platforms. Participants received samples of undigested or tryptically digested yeast proteins and were requested to follow strict analytical guidelines. Data analysis was centralized and performed under standard procedures using a common workflow. The study revealed that the overall performance with respect to metrics such as reproducibility, sensitivity, dynamic range etc. was directly related to the degree of operator expertise, and less dependent on instrumentation.

Several studies not specifically focused on quality control have also yielded insight into proteomic reproducibility. The HuPO plasma proteome project (HuPO PPP) distributed 20 human samples (five serum plus 3 $\times$ 5 plasma samples treated with three different anticoagulants) to 35 laboratories spanning 13 countries (8). The purpose of this large-scale study was not to assess reproducibility *per se*, but rather to generate the largest and most comprehensive data set on the protein composition of human plasma/serum. On a smaller scale, the ISB standard 18 protein mixture (purified proteins from cow, horse, rabbit, chicken, *E. coli*, and *B. licheniformus*) was also assessed between laboratories on eight different LCMS platforms (9). These data reside in a comprehensive, multiplatform database as a resource for the proteomic community. Additional interlaboratory assessments have consisted of multiple reaction monitoring-based measurements of peptides/proteins in plasma (10, 11) and protein–protein interactions at both the biochemical and proteomic level (12).

For team leaders/directors of proteomic laboratories and any researcher collaborating with such groups, major questions that may arise concerning data consistency are: how well are quality controls being implemented in the daily operations? Do the quality control measures effectively support data reproducibility? To address this, the Protein Research Group of the Association of Biomolecular Resource Facilities (ABRF-PRG) designed a study whereby LC-MSMS data obtained from the analysis of a commercially available bovine protein mixture predigested with trypsin were collected at routine intervals over a period of 9 months. Raw MS data files from a total of 64 participating laboratories were accumulated, and HPLC and MS performance were evaluated through QC metrics (13). The main impetus of the study was to recognize key sources of variability in HPLC and MS analyses under extended and routine operating conditions for each laboratory and to catalog the state of quality control in a diverse set of proteomic laboratories.

No standard operating protocol was imposed on the participants; instead, contributors were encouraged to employ

[1] The abbreviations used are: LC-MSMS, liquid chromatography tandem mass spectrometry; DDA, data-dependent acquisition; FDR, false-discovery rate; HPLC, high-performance liquid chromatography; IQR, interquartile range.; LCMS, liquid chromatography mass spectrometry; MSMS, tandem mass spectrometry; PCA, principal component analysis; PM, preventative maintenance; PSM, peptide-spectrum match; QC, quality control; SOP, standard operating procedure.

the methods that were typically applied in individual laboratories. Optimization of instrument methods on the provided sample was discouraged. A survey was conducted with each sample submission to catalog individual laboratory practices, instrument configurations, acquisition settings, including routine and nonroutine maintenance procedures. Unlike previous investigations where emphasis was placed on the preparation, distribution, and evaluation of protein standards to appraise and/or standardize LCMS platforms between laboratories, the key interest in this study was purely to determine the intralaboratory performance, reproducibility, and consistency of participating laboratories over an extended period of time.

The rapidly expanding number of proteomic laboratories have incorporated divergent HPLC systems, mass spectrometers, solvent systems, columns etc. As a result, analyzing data from a large number of laboratories necessitates tools that can accommodate data from a broad range of platforms. For example, to expect a small laboratory with a decade-old three-dimensional ion trap mass spectrometer to achieve the same sensitivity as a laboratory with a high-resolution hybrid instrument would be unfair. Correspondingly, the data analysis needs to include axes beyond simple peptide-level sensitivity. Nevertheless, the laboratory with the older instrumentation may be consistently better at maximizing performance from the chosen instrument platform compared with a laboratory with the latest high-end equipment.

The focus of this study was to estimate the degree of variability in intralaboratory performance over a 9-month period. This goal was achieved using quality metrics that are applicable to most LC-MSMS workflows. The inclusion of data from many laboratories will enable the proteomic community to determine the current state of quality control within a typical laboratory. The survey data enabled the mapping of some alterations in instrument performance to documented laboratory events, *e.g.* mass spectrometer calibration. The study was designed neither to compare one laboratory with another, nor to discriminate between classes of instrumentation.

Questions of data quality and performance in the proteomic community are appropriately aligned with the heightened awareness of a perceived lack of reproducibility of scientific findings in general (1). This community has endeavored to provide tools to assess proteomic data quality, and this study provides additional insight into the application of such tools and the quality of data within respective laboratories.

EXPERIMENTAL PROCEDURES

Recruitment and Participation—Regardless of ABRF membership, participants in the study were recruited from the wider mass spectrometry community. Anonymity throughout the study was ensured via provision of a participant code that was administered by an ABRF-assigned ombudsman. ABRF members that performed survey data collation, data reduction, or analysis were not provided access to any identifying information, and the ombudsman facilitated confidential discussions when necessary. Integrity of data uploads was verified by comparison of raw file header information, verification of IP addresses, and comparison with survey entries.

Study Design—At the launch of the longitudinal study, participants were provided with a total of ten individual vials of the lyophilized bovine six protein tryptic digest equimolar mix (1 pmol each of β-lactoglobulin (P02754), lactoperoxidase (P80025), carbonic anhydrase (Q1LZA1), glutamate dehydrogenase (P00366), β-casein (P02666), and serum albumin (P02769); Bruker-Michrom, Auburn, CA). A single lot of the mixture was procured and dissolved in 2% acetic acid. Aliquots were prepared with a high-precision automated liquid dispensing robot. Prior to distribution to participating groups, aliquots were thoroughly characterized in the laboratories of two ABRF-PRG members.

At the start of the study, each participant was asked to answer an extensive survey that covered basic demographic data plus details regarding instrumentation, typical operating parameters, etc. (survey questions are available as supplementary material). At monthly intervals, participants dissolved the contents of a fresh, individual vial; analyzed the sample by data-dependent nano-LC-MSMS; and uploaded the generated raw MS data file(s) to a central server hosted by Bioproximity, LLC (Chantilly, VA). Additionally, the participants were asked to complete a brief supplementary survey to record any major changes that had occurred since the previous month. Results were submitted under participant ID codes to ensure anonymity of the laboratories. At the completion of the study, the raw data were annotated with quality metrics and associated with the corresponding survey responses. Participant 914061 was deemed a "super-user" to reflect that the samples were analyzed on a linear trap quadrupole (LTQ) Orbitrap Velos at a higher frequency (fifteen times during the 9-month period) than the other participants. Data were acquired at two concentrations (10 and 40 fmol injected per LC-MSMS analysis) using two tandem MS fragmentation modes: LTQ collision-induced dissociation (CID) and quadrupole higher-energy collisional dissociation (HCD). Throughout the text, the "super-user" data are subdivided into four sets: cid10, cid40, hcd10, and hcd40.

Raw Data File Processing, Database Matching, and Quality Metrics—For Agilent, Bruker, and ThermoFisher instruments, ProteoWizard msConvert (14) was used to transform the raw MS data into the mz5 format. The filter settings conducted vendor library peak picking on all types of scans within the files ("peakPicking true 1-"). For Waters instruments, the same software was used, but the CantWaiT filter in ProteoWizard was used for peak picking ("peakPicking cwt msLevel = 2-") and the TurboCharger filter provided precursor charges (15). Using "ProteinPilot" peak picking, AB SCIEX .wiff files were first processed through the AB SCIEX MS Converter to produce mzML files. ProteoWizard msConvert was then applied to translate the files into the mz5 format. When available, this conversion preferred vendor-supplied algorithms for peak picking.

MyriMatch v2.1.138 (16) provided database search identifications against the RefSeq bovine protein database (downloaded 14 September 2010, containing 33,684 bovine sequences plus 71 contaminants including proteases, immunoglobulins, keratins, and fabric proteins) with each represented in both forward and reversed form. MyriMatch was configured to apply $\pm$10 ppm; $\pm$1.25 *m/z*; or $\pm$50 ppm precursor mass tolerances for accurate mass ThermoFisher instruments; other ThermoFisher mass spectrometers; and all other instruments, respectively. Irrespective of the mass analyzer, fragment ion tolerances were set to $\pm$0.5 *m/z*. Fully tryptic specificity was employed. Oxidation of methionine (+15.9949 Da), pyro-glutamylation (−17.026 Da on N-terminal Gln), and deamidation (+0.984016 Da on Gln, Asn) were allowed while assuming that all cysteine residues were modified by iodoacetic acid (+58.00548 Da). Semitryptic searching was considered as an alternative approach, but the impact on the data analysis appeared limited (data not shown).

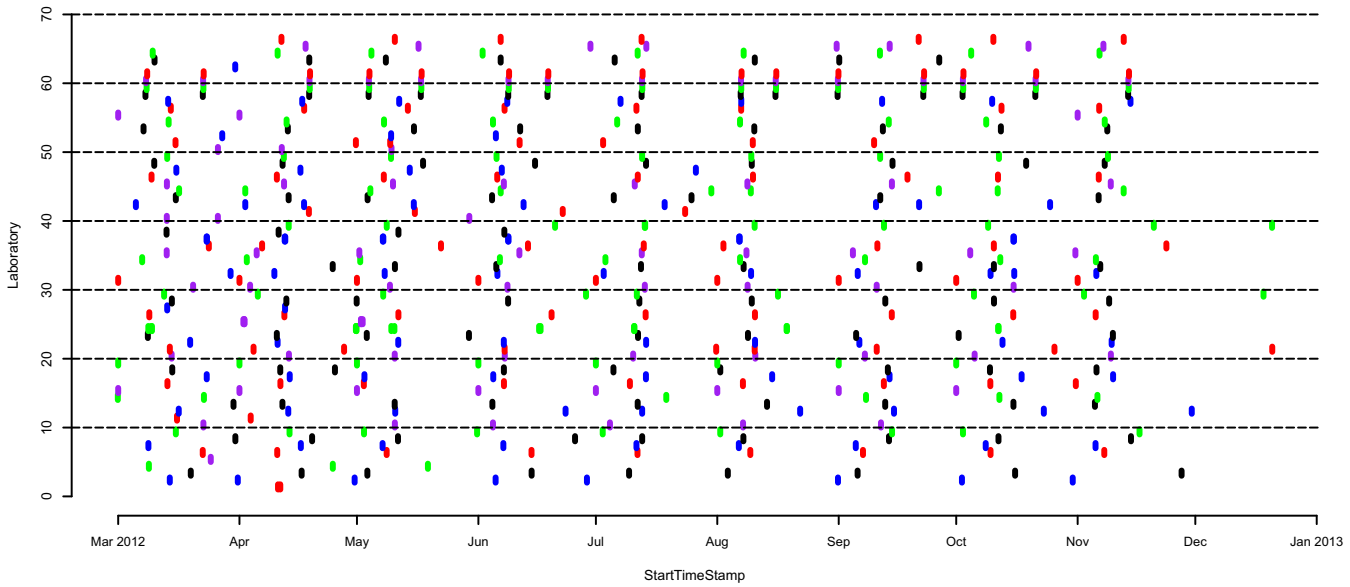

FIG. 1. **Timeline for the data generated across all 63 of the participating laboratories from March 2012 to January 2013 (participant 914061 is divided into four groups).** A total of 526 MS data files were generated during the study. Note: color coding is included to readily visualize the different laboratories.

IDPicker 3.0 (17) filtered the protein identifications and conducted parsimonious protein assembly. Peptide-spectrum matches (PSMs) were filtered at a 2% FDR (doubling reverse hits and dividing by the total passing the threshold), with two peptides required per protein for inclusion. The PSM FDR was individually computed for each LC-MSMS experiment. Based on the precursor ion charge and enzyme specificity (trypsin), PSMs were separated prior to estimating the FDR. In addition to the six known proteins in the mixture, a further six "hitchhiker" bovine proteins were treated as legitimate identifications. These were superoxide dismutase [Cu-Zn] (P00442), cationic trypsin (P00760), α-S2 casein (P02663), selenium-binding protein 1 (Q2KJ32), sulfhydryl oxidase 1 (F1MM32), and phosphatidylethanolamine-binding protein 1 (P13696). An aggregate assembly of all supplied data included a total of 212,173 spectra identified confidently to 343 distinct protein groups (reflecting that some participants experienced a degree of protein carry-over on the LCMS systems, as most of these proteins were unlikely to be present in the standard). The protein FDR for the aggregate assembly was 5.06% whereas the effective peptide-spectrum match FDR was 0.13%. Of the total, 166,721 spectra (78.58%) matched to the six proteins known to be in the defined mixture; and 188,589 spectra (88.88%) matched to either the six defined or six "hitchhiker" proteins.

QuaMeter (18) was applied to the peak-picked mz5 files in identification-independent mode to produce quality metrics (13). The "ID-Free" mode generated 46 metrics for extracted ion chromatograms, retention time, mass spectrometry, and tandem mass spectrometry characterization (see supplemental Table S1 for all metrics). The tables produced from the raw data were evaluated in the R statistical environment for robust principal component analysis (PCA) and outlier detection using 33 (supplemental Table S1, bold red) of the 46 metrics. Note that metric 1 (FileName, red) was not included in the data analysis.

Principal Component Plot and Dissimilarity Measures—PCA plots function as exploratory visualization tools for experiments. The two-dimensional PCA plots use the first two principal components, accounting for a large proportion of the variability observed in the data. PCA aids in identifying clusters in the experimental analyses and potential outlying experiments. As shown in Wang *et al.* (13), two experiments are compared using a dissimilarity measure. This measure is based on the normalized Euclidean distance between the robust PCA coordinates for each LC-MSMS experiment. The larger the dissimilarity values, the lower the similarity between the two experimental analyses. This distance measure is designed to be automatically outlier-proof, as pair-wise comparison does not require a benchmark profile. Any abnormal experiments (outliers) can be easily identified by the large distances from other experiments. Clusters of experiments can be identified by small dissimilarity values within a set.

T^2- Chart for Quality Control—A T^2-chart was applied as a multivariate quality control tool. The large number of quality metrics for each experiment were summarized by a single T^2-statistic and the values were temporally displayed. The patterns and trends are then more visible in a longitudinal experiment. More importantly, the T^2 statistic is assumed to follow a χ^2-distribution, and upper and lower control limits can be obtained with a given significance level. When compared with the dissimilarity measure, this approach provides a more rigorous evaluation of the outlier experiments with a cut-off significance level.

Change Point Analysis—Change point analysis is another statistical tool that can be used to process the data generated from a quality control study. In contrast to the χ^2 test, where the aim is to detect individual abnormal analyses with some isolated causes; the aim of change point analysis is to detect sustained temporal alterations in the pattern of the experiments. Such changes can be captured at the mean level, the variability level, or both. Such changes are usually because of sustained causes, which lead the experimental process to switch from one status to another and remain at this point for a period of time. Detection of change points can aid in examining the causes of batch effects.

RESULTS

Sixty four participants contributed a minimum of one MS raw data file; the typical participant in the study reported 6–10 years LCMS experience, used nanoflow HPLC in a two-column trapping configuration, and injected 200 fmol of the

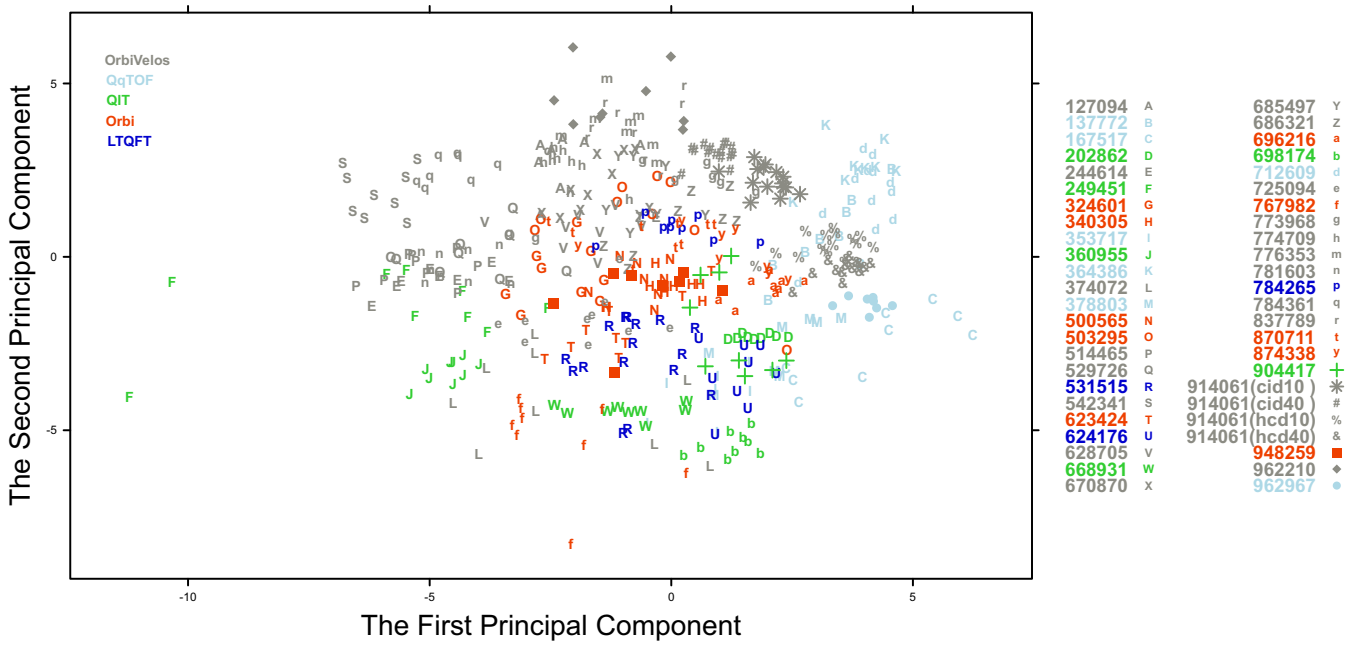

FIG. 2. **Robust PCA plot for the 458 data files from the 45 participating laboratories that submitted a minimum of eight raw data files.** The first two principal components account for 36.5% of the variability, and ~60% of the variability in the data is accounted for by the first five principal components. Data is color-coded according to mass spectrometer type.

digest mixture for each analysis. Fig. 1 shows the time plots for the 526 MS files generated by 63 of these participants from March 2012 to January 2013 (laboratory 627604 was removed from further analysis as the raw data contained MS1 information only). The color coding is included purely to visualize the different laboratories. Forty-five participants uploaded at least eight individual experiments to yield a large repository of 458 raw data files collected longitudinally, within and across multiple laboratories. Instruments included several mass spectrometer models from AB SCIEX (4); Agilent (1); Bruker (2); ThermoFisher (35); and Waters (3); and were categorized as either a: (1) quadrupole ion trap (QIT); (2) quadrupole time-of-flight (QqTOF); (3) linear trap quadrupole Fourier Transform (LTQFT); (4) LTQ Orbitrap classic (Orbi); and (5) LTQ Orbitrap Velos (OrbiVelos). HPLC instruments were from Agilent (3); Dionex (13); Eksigent/Sciex (10); Michrom (2); Proxeon/ThermoFisher (4); and Waters (12). HPLC gradients ranged from 22 to 160 min, and the total number of MSMS acquisitions varied from less than 100 to nearly 30,000 tandem mass spectra per experiment. No effort was undertaken to standardize the LC-MSMS conditions between laboratories. Consequently, the number of identifications produced from the data sets also ranged widely.

Most laboratories reported performing some type of preventative maintenance and calibration during the course of this study. In addition, >90% of the participating groups also reported performing quality control analyses as part of routine LC-MSMS operation. The type, frequency, and methods of evaluation, however, varied dramatically. Approximately 20% of the laboratories reported at least one major service event (*e.g.* HPLC pump rebuild, electron multiplier replacement, or control board replacement) during the course of the study, and most groups reported other minor corrective changes (*e.g.* replacement of emitters, replacement of leaking HPLC unions). Only one participant reported changes to the LC or mass spectrometer data collection or operating parameters between data uploads. An unusually high variance in column life was noted with the reported number of injections ranging from 5–1058. Interestingly, mean column life appeared to be a feature unique to each individual laboratory (see supplemental Tables S2 and S3 for all the information collected during the survey for all participants and a reduced list for the 45 participants that uploaded at least eight individual experiments, respectively).

PCA based on identification-independent quality metrics (Supplemental Table S1) visualized all participants in a single plane and showed that for most of the contributors, submitted experiments clustered together (Fig. 2). Experiments from the same type of instrument also tended to cluster. In some cases, however, the LC-MSMS experiments were broadly dispersed, revealing significant variation in the data, *e.g.* 374072 (L) that was generated on an OrbiVelos; and 167517 (C) and 767982 (f) that were both generated on Orbi mass spectrometers. Additionally, some LC-MSMS experiments showing abnormal performance were isolated from the majority of the analyses from the same participant, *e.g.* the data from 249451 (F) that was generated on a QIT. These dispersions were re-evaluated in the context of the survey results provided by the contributing laboratories and are discussed later.

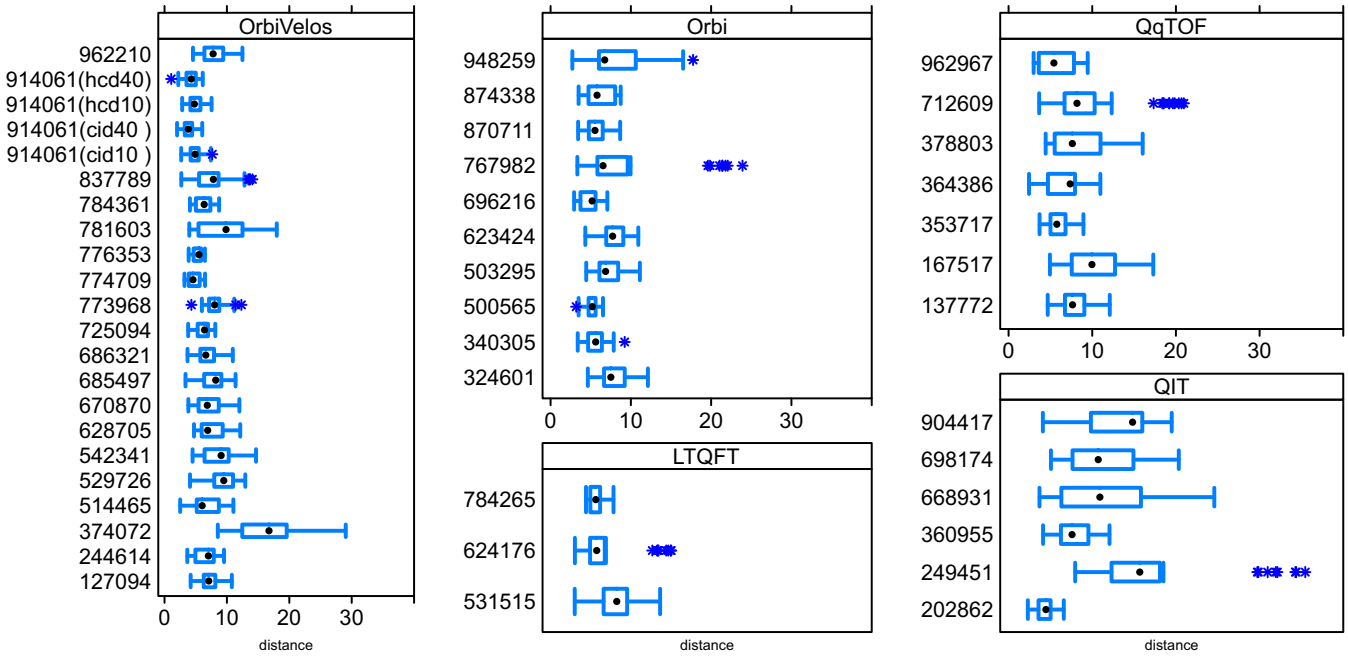

FIG. 3. **Dissimilarity measures between pairs of experiments at each of the 45 participating laboratories that submitted a minimum of eight raw data files (participant 914061 is divided into four groups).** Asterisks represent extreme pair-wise distances within the data, *i.e.* outside Q1–1.5 IQR or Q3 + 1.5 IQR.

Shown in Fig. 3 are the pair-wise dissimilarity measures within each of the participating laboratories and are grouped by the type of mass spectrometer. The figure is based on normalized Euclidean distances using robust PCA scores. The dissimilarity measures were used to evaluate the variability across either the participants or the type of instrument. The data analysis revealed that in terms of median dissimilarity, experiments conducted on a QIT exhibited the greatest variability. The four other types of mass spectrometer showed similar variability and were ranked as: OrbiVelos < QqTOF < LTQFT < Orbi (see Supplemental Table S4). Note that these comparisons were not controlled for sample size or operator experience. Fig. 3 also aided examination of outlying LC-MSMS experiments for each participant. The asterisks in the plots represent extreme pair-wise distances within the data from each participant (*i.e.* outside the interquartile range [IQR] by a factor of 1.5*IQR). If the dissimilarity measures related to a specific experiment appear several times as extreme values in the plot; then it follows that the LC-MSMS analysis is possibly an outlier and the quality metrics are quite different from the majority of those provided by the participant. For example, if one LC-MSMS analysis is only extremely dissimilar from one of the other experiments, it is much less likely to be an outlier compared with an LC-MSMS analysis that is extremely dissimilar from five or six experiments. From 458 experiments, 60 exhibited extreme dissimilarity measures. From these 60, seven outlying LC-MSMS analyses were identified based on the frequency that an experiment appeared to be dissimilar from another. The seven LC-MSMS analyses included one from each of the following participants: 249451, 624176, 712609, 767982, 773968, 837789, and 948259. The seven points labeled with asterisks in Fig. 4 (described in detail in the following section) are the same outlying experiments. Six out of the seven were also identified as outlier experiments by the T^2-chart.

Plotted in Fig. 4 are the T^2-statistics for all LC-MSMS analyses. A family error rate of 0.01 was used for each participant. The filled pink circles are LC-MSMS analyses that were determined as outliers, whereas the blue filled circles represent data that was within control. Outlier experiments were classified as either isolated or sustained. Classification as one or the other was dependent on whether or not any neighboring experiment was also an outlier. For isolated outlier experiments the filled pink dots are encircled in black. Change point analysis was performed on the remaining experiments for both in control and sustained outlier experiments. Here the assumption was that only the average values of the T^2-statistics are altered before and after a change point. The R library changepoints was used. Batch means that are separated by change points are indicated with black horizontal lines. Several participants (n = 29) had one change point in the temporal data. Such change points can be used to study early and late batch differences. Both types of outlier events were examined together with the participant survey information to trace possible causes and are discussed later.

Participants 202862 (QIT); 137772 (QqTOF); 696216, 500565, 340305 (Orbi); 784265 (LTQFT); and 774709, 725094, 914061 (OrbiVelos) showed the most consistent within control LC-MSMS analyses over the 9-month period. The "superuser" (914061) also showed stable, longitudinally controlled

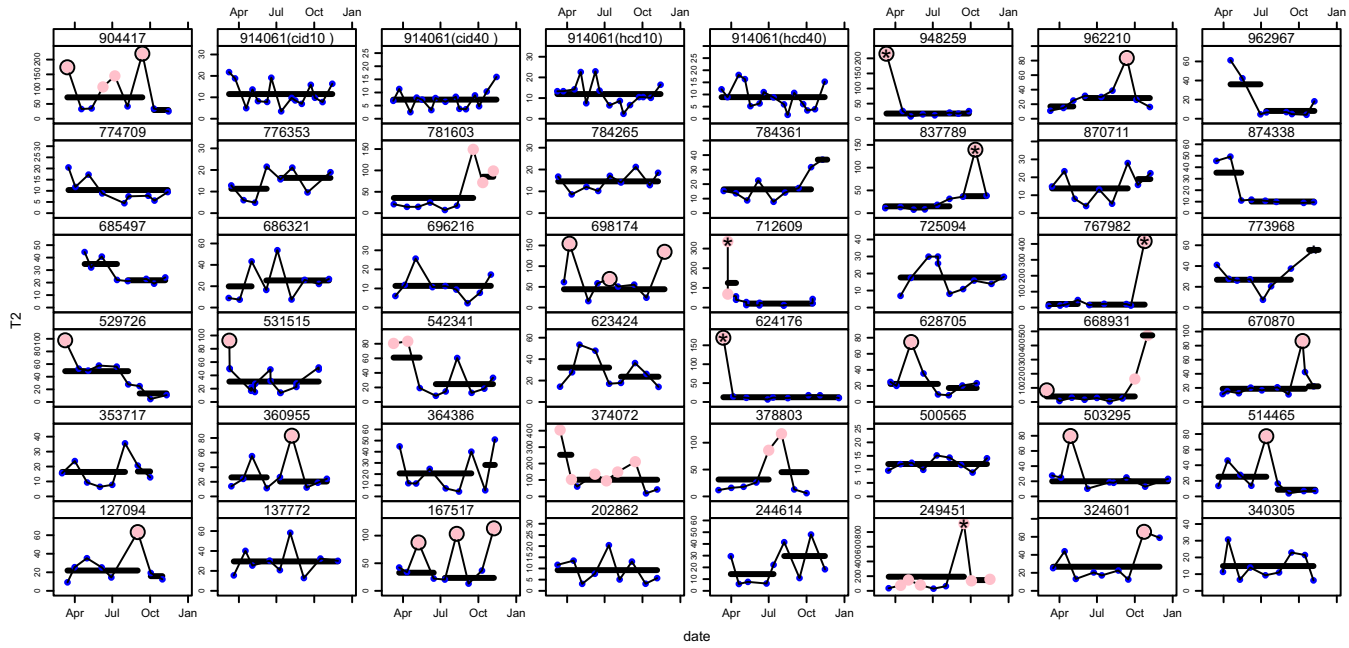

FIG. 4. χ^2-**quality control chart based on T^2-statistics.** The filled pink and blue circles represent analyses that are outlier and within control experiments, respectively. The family error rate used for each participant was 0.01. Filled pink dots encircled in black are isolated outlying experiments, which were not considered in the detection of change points. Asterisks indicate outliers that were identified by dissimilarity measures. Batch means identified from the change point analysis are indicated by the black horizontal segments.

analyses for the two selected sample concentrations analyzed with the two peptide fragmentation methods. Participants that exhibited within control experiments over the course of the study, but additionally had early and late batch differences were: 244614, 685497, 686321, 773968, 776353, 784361 (OrbiVelos); 353717, 364386, 962967 (QqTOF); and 623424, 870711, 874338 (Orbi). Participants 948259 (Orbi) and 624176 (LTQFT) are examples of the data sets with a single, outlying experiment. The change point analysis revealed that longitudinally there were no batch effects, and the data were very consistent. Only the first time point at the initiation of the study fell outside the controlled data analysis limits. Additionally, participants 324601, 503295 (Orbi); and 531515 (LTQFT) had a single time point that was outside the margins of the controlled data. These events occurred at time points 8, 3, and 1, respectively.

Participants that showed in control batch effects with only a single, isolated outlier experiment were: 360955 (QIT); 767982 (Orbi); and 127094, 514465, 529726, 628705, 670870, 837789, 962210 (OrbiVelos). The single outlying data points for 767982 and 837789 were extreme outliers. Participants that showed sustained outlying experiments with both early and late batch effects and more than one outlier experiment were: 378803; 712609 (QqTOF); 542341, 781603 (OrbiVelos); and 668931 (QIT). Participant 712609 also had an extreme outlier at the first time point. Five participants revealed rather erratic longitudinal data. These were: 167517 (QqTOF); 249451, 698174, 904417 (QIT); and 374072 (OrbiVelos). Participant 698174 did not have an early *versus* late batch effect; however, data from time point 2, 5, and 9 were outside the margins of the controlled experiment. Participants 167517, 249451, 374072, and 904417 showed early *versus* late batch effects and had 3, 6, 6, and 4 data sets, respectively, that were classified as outlier experiments. Additionally, 249451 showed a divergent outlier at time point 7. Interestingly, these five participants all noted in the survey that a system suitability or QC test was run prior to analyzing the PRG sample. Indeed, only ~10% of the laboratories answered negatively to this survey question.

Closer examination of the survey data in the context of the outlying data points was quite revealing. For the six laboratories that only had a single outlying event, no survey entries were recorded for 781603 and 948259; and no survey data correlated with the outlying event for 767982. Participant 624176 reported a generic problem with the instrument, 529726 indicated that a preventative maintenance (PM) was performed, and 542341 recorded that the instrument had a major service requiring the replacement of a control board. Nine laboratories had two outlier occurrences. One participant only recorded minimal information. For the other eight, these events were correlated with: (1) a HPLC column leak and PM (324601); (2) awareness of an instrument issue (360955); (3) a trap column change and PM (514465); (4) a column change and PM (628705, 962210); (5) a PM between time points 8 and 9 (668931); (6) awareness that the instrument was in need of cleaning (670870); and (7) column overpressure and recuperation by column shortening and a PM (837789).

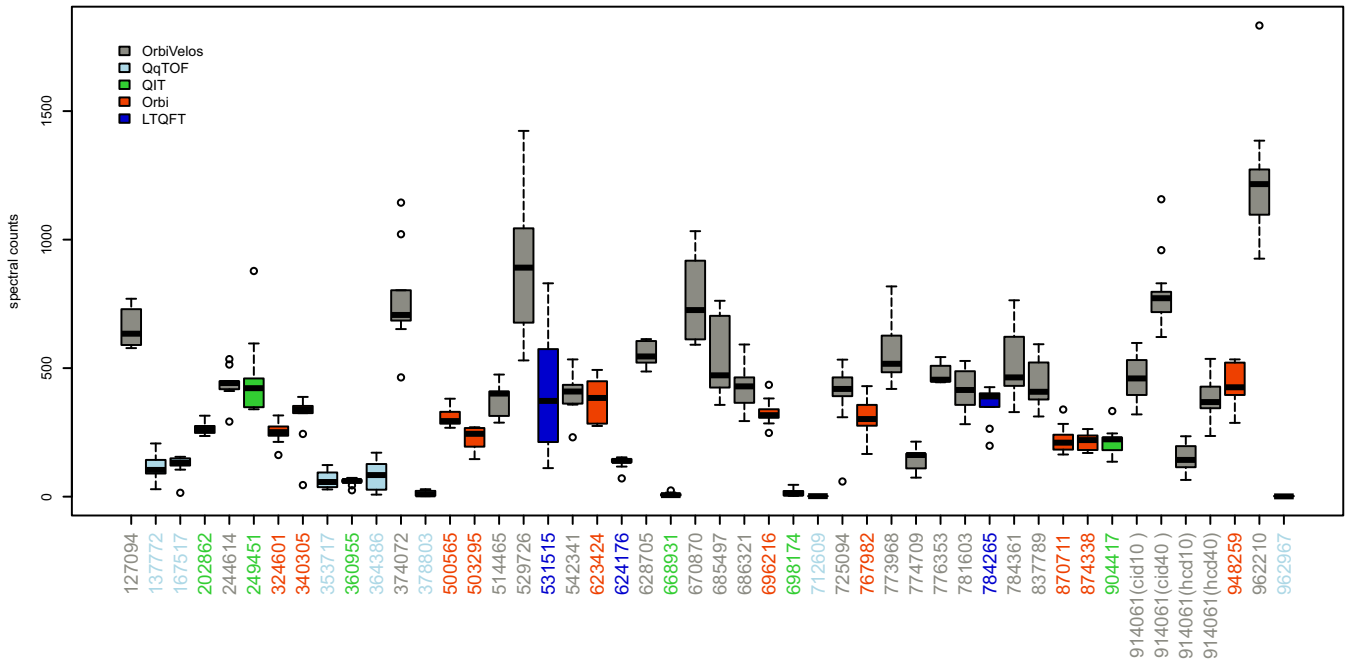

FIG. 5. **Spectral counts identified in each file from the 45 laboratories that submitted a minimum of eight raw data files.** Data is color-coded according to mass spectrometer type.

For the remaining five laboratories that had more than two outlying data points, no entries were recorded for participant 167517. A PM together with an electron multiplier exchange prior to the first LC-MSMS analysis, and another PM plus a column change prior to the seventh time point were recorded by 904417. No survey event correlated with the outliers for analyses four and five. Participant 689174 reported a PM prior to the second time point; and a column change prior to the ninth analysis. Investigation of the survey data for two of the participating laboratories that had several outlying data sets, namely 249451 (QIT) and 374072 (OrbiVelos) revealed that 374072 reported both PM and HPLC maintenance for the first, second, and fourth analyses. In addition, generic maintenance was performed between analyses six and seven, but no survey event correlated with any of the other outlying data sets. For participant 249451, no survey event correlated with any of the out of specification LC-MSMS analyses.

Fig. 5 reports the number of spectral counts obtained for the six known bovine proteins plus the additional six "hitchhiker" proteins (see experimental procedures) for each participant that submitted a minimum of eight raw MS files. The results are color-coded according to mass spectrometer type. The data revealed that overall, the QqTOF data (light blue) generated the lowest number of spectral counts (see participants 137772, 167517, 353717, 364386, 378803, 712609, and 962967); however, the variability in the data sets was minimal. For the three participants that analyzed samples on an LTQFT (dark blue), the median spectral counts for 624176 $<$ 531515 $<$ 784265; however, the variability for 531515 was among the highest of all participants enrolled in the study. Participants 202862, 249451, 360955, 668931, 698174, and 904417 analyzed the samples on a quadrupole ion trap (green). As a general observation, the median number of spectral counts for the 12 bovine proteins was relatively low; although three participants from this group (202862, 249451, and 904417) produced median spectral counts higher than the quadrupole time-of-flight data (light blue). For the 10 participants that analyzed the samples on an Orbitrap (orange), the median number of spectral counts were in the range of 300 to 500. The variability was, in general, low.

Finally, the remaining 19 participants that used an Orbitrap Velos (gray) to analyze the bovine mixture showed differences in the degree of variability; and overall, the median number of spectral counts was higher for the data analyzed on this machine type. In particular, participant 962210 had the highest number of spectral counts with a median of ~1200. For the "super-user," the 40 fmol samples had an increase in median spectral counts *versus* the 10 fmol sample, and the collision-induced data had higher spectral counts than either of the higher-energy collisional activation data sets. Fig. 6 shows that the variability in spectral counts is relatively stable compared with the variability in the quality control metrics. No statistically significant correlations were apparent. Note that there were a low number of spectral counts from the data for the participants shown in blue.

DISCUSSION

This ABRF-PRG study provides a large resource of longitudinally collected raw MS data files from more than 60 participating laboratories world-wide. Through the analysis of a

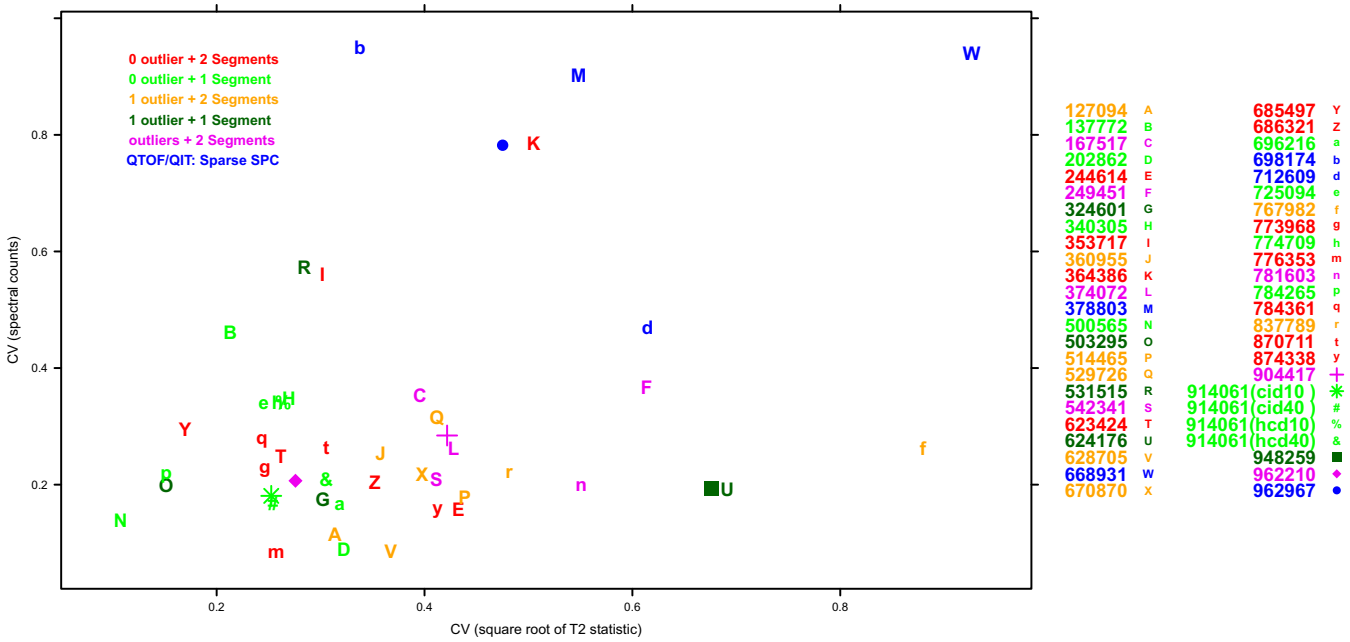

FIG. 6. **Coefficient of variation based on spectral counts and quality metrics.** SPC, spectral counts.

supplied, standard protein digest, the study was designed to assess the temporal degree of internal variability or consistency of each laboratory via an examination of 33 identification-independent quality metrics. From the data, factors such as estimating typical variation, revealing outlier experiments, and inferring change points in instrument performance were ascertained. The repository of data collected over the 9-month period contains the raw mass spectrometry data files, information on such metrics as mass spectrometer and liquid chromatography system operation and performance, workflows, and operator experience. As no constraints were placed on participants, the data may actually closely emulate daily operations in service and research-orientated proteomic laboratories.

Our attempts to correlate specific outlier events in the longitudinal data with an occurrence in the daily operations of a laboratory proved to be more challenging than anticipated. Overall, 58% of the final 45 participants meticulously recorded eight to nine entries during the study; however, 42% had less than eight entries with 13 laboratories only recorded from zero to two entries (supplemental Table S2 and S3). It appeared from the accompanying survey information that few participants accurately recorded details or were unaware of any alterations that were made between the analyses for specific time points. When the survey information was combined with the final outcome of the study, several participants with outlier data either: 1) did not record any survey entry changes at all, *e.g.* 167517, 781603, and 948259; 2) entered sparse information, *e.g.* 378803; or 3) not a single survey event could be correlated with an observed outlying data point, *e.g.* 249451 and 767982. Such missing pieces of information suggested that participants were either unaware of changes in the workflow, or did not convey observed changes through the survey. Despite high quality and reproducible data, participants 202862 (QIT, zero entries); 137772 (QqTOF, one initial entry); 500565 (Orbi, one initial entry); and the "super-user" 914061 (OrbiVelos, two entries) failed to enter complete survey information. At the other end of the spectrum, participants that had erratic data but diligently supplied survey entries were 249451 and 698174 (QIT, nine entries each); 904417 (QIT, seven entries); and 374072 (OrbiVelos, eight entries). Another interesting observation was related to performing an internal QC prior to analyzing the ABRF-PRG sample. From the participating laboratories, 90% recorded a positive pre-QC assessment; however, this affirmation was not reflected in the questioning the diligence of the internal quality control assessments made. Thus, for the many of the observed change points and outlier events, it was impossible to specifically correlate a recorded event to account for the observed differences. The single most outstanding observation that could be made from the correlation of the survey data results with outlying data points, however, was that these events showed a tendency to occur directly following a preventative maintenance on the instrument. This suggests that an extra degree of scrutiny following a PM may become necessary to maintain quality within a laboratory.

Beyond the single laboratories that participated in this study, the R changepoint analysis revealed interesting findings on proteomic laboratories in general. Twenty percent of the groups did not have any outliers over the 9-month period, and the experiments fell within a single segment, *i.e.* there were no apparent batch effects. Although data from some laboratories separated into two segments indicating an early and late batch effect, an additional 24.4% of the groups were

also without any outlying experiments. Thus, in total ~45% of the participants could show consistent and reproducible longitudinal data over the course of the study. Further, approximately one third of the participants had only a single outlying experiment, with 11.1% and 22.2% of these groups showing data segregation into one and two segments, respectively. At the other end of the spectrum, however, 20% of the participants exhibited relatively poor reproducibility and consistency. These groups had more than one outlying data point and the results were also segregated into the two early and late segments. Overall, around 80% of the groups that were involved in the ABRF-PRG study performed reasonably well to excellent. This indicates that indeed the majority of proteomic service and/or research laboratories perform to an equally high standard. These findings appear to be independent of factors such as the chosen instrument platform, age of the equipment, amount of sample injected, or experience of the operator. Interestingly, the study also revealed that there was no particular bias toward any of the five mass spectrometer types (LTQFT, Orbi, OrbiVelos, QIT, and QqTOF) used in the study. Highly consistent, reproducible data and low-quality, inconsistent data were obtained from all instrument types, suggesting that the operator and/or the performance of the HPLC system coupled to the mass spectrometer were the deciding factors in the data quality.

Evaluation of the PCA in the context of QuaMeter metrics revealed that there were eight features that led to the most outliers. These were: (1) RT.MSMS.Q2; (2) RT.MSMS.Q3; (3) MS1.TIC.Change.Q2; (4) MS1.TIC.Q2; (5) MS1.Density.Q3; (6) MS2.Count; (7) MS2.Freq.Max; and (8) MS2.Density.Q3, RT.MSMS.Q2, and RT.MSMS.Q3 refer to the fraction of retention time during which the second and third quartiles of the MSMS scans appear. As this can be considered the "sweet spot" for peptide identification; a wide time range for these two quartiles (a sum over 0.5, or 50% of the time) would be ideal. This observation is highly related to a comparable metric underscored by Rudnick *et al.* (7). In addition, MS2.Count rises when the instrument is generating a high number of MSMS spectra, so this metric is positively correlated with an increase in peptide identification rates. MS2.Freq.Max will also rise when identification rates are high.

The other four metrics were more difficult to correlate with the effectiveness of peptide identification, however, MS1.TIC.Change.Q2 and MS1.TIC.Q2 are interrelated (the former is essentially the first derivative of the latter). The MS1.Density.Q3 and MS2.Density.Q3 metrics are concerned with the number of peaks that appear in the spectra. Thus, the assumption is that most of the outlying experiments have low values for RT.MSMS.Q2, RT.MSMS.Q3, MS2.Count, and MS2.Freq.Max. We are confident in our assessment of the data as the natural expectation is that the best peptide identification performance occurs when all LCMS parameters are optimal; and conversely, poor identification performance is likely to result if any one (or more) parameter(s) are out of specification.

In summary the conducted study was reassuring in that the ABRF-PRG could conclude that the majority of proteomic laboratories analyzing peptide samples via data-dependent on-line LC-MSMS were performing within the bounds of medium to high reproducibility and consistency. The one major factor for loss of reproducibility appeared to correlate with a preventative maintenance on the instrument. Perhaps more concerning was the apparent failure of many laboratories to follow instructions on properly tracking and auditing alterations in instrumentation through the accompanying survey. Based on this study, the ABRF-PRG recommends that carefully scrutinized quality audits should follow maintenance activities to ensure instruments operate at optimal reproducibility. In addition, the ABRF-PRG suggests that the scientific community strives to address the issue of accurate recording and observing changes in QC metrics by more intensive and thorough training of instrument operators; using dedicated staff that have an intimate knowledge and understanding of the assigned system; optimizing the hand-over procedure of a system from one operator to another; and regardless of how minor and insignificant a detail may seem, considerably improving the recording of system information. Here, instrument vendors may be able to assist by developing software alert systems that can be programmed with specific QC settings by the team leaders/directors of proteomic laboratories. Ultimately, these combined measures should markedly aid in maintaining consistency and reproducibility within a laboratory. With this capability comes an increasing confidence from collaborators that the data received is of the highest quality possible.

Acknowledgments—The PRG2012/2013 group gratefully acknowledge Michrom Bioresources for the generous donation of the six protein bovine mix and Brian Blagley at Bioproximity, LLC for contributing data storage and transfer capabilities. We also thank Paul A. Rudnick from the National Institute of Standards and Technology; Jason G. Williams from the National Institute of Environmental Health Sciences; and Christian Knoll from CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences for invaluable advice and assistance during the course of the study. The mass spectrometry data files, QuaMeter metric tables, and IDPicker assembly have been deposited into the ProteomeXchange Consortium via the MassIVE services at the University of California San Diego Center for Computational Mass Spectrometry with the accession code PXD002114.

* This work was supported by ABRF. D.L.T., X.W., and M.C.C. were funded by NCI U24 CA159988. X.W. was also supported by the 2014 UC LEAF Career Branch Awards at the University of Cincinnati. R.L.M. was funded in part by NIGMS Grant No's 2P50 GM076547 and GM087221. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

S This article contains supplemental Information and Tables S1 to S4.

e Current address: SciKon Innovation, Inc., 2 Davis Drive/FFVC, Research Triangle Park, NC 27709

§§§ To whom correspondence should be addressed: Department of Biomedical Informatics, Vanderbilt University, U9211 Learned Lab/MRB III, 465 21st Ave S, Nashville, TN 37232-8575. Tel.: +1-615-936-0380; Fax: +1-615-343-8372; E-mail: david.l.tabb@vanderbilt.edu; CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences Lazarettgasse, 14, AKH BT 25.31090 Vienna, Austria. Tel.: +43-1-40160-70010; Fax: +43-1-40160-970000; E-mail: kbennett@cemm.oeaw.ac.at, and SciKon Innovation, Inc., 2 Davis Drive/FFVC, Research Triangle Park, NC 27709. Tel.: +1-919-593-0056, Maureen K. Bunger, VP, Director of Business Development and Marketing, Director of Product Development; E-mail: maureen.bunger@gmail.com.

REFERENCES

1. Kocher, T., Pichler, P., Swart, R., and Mechtler, K. (2011) Quality control in LC-MS/MS. *Proteomics* **11,** 1026–1030
2. Tabb, D. L. (2013) Quality assessment for clinical proteomics. *Clin. Biochem.* **46,** 411–420
3. Bell, A. W., Deutsch, E. W., Au, C. E., Kearney, R. E., Beavis, R., Sechi, S., Nilsson, T., and Bergeron, J. J. (2009) A HUPO test sample study reveals common problems in mass spectrometry-based proteomics. *Nat. Methods* **6,** 423–430
4. Paulovich, A. G., Billheimer, D., Ham, A. J., Vega-Montoto, L., Rudnick, P. A., Tabb, D. L., Wang, P., Blackman, R. K., Bunk, D. M., Cardasis, H. L., Clauser, K. R., Kinsinger, C. R., Schilling, B., Tegeler, T. J., Variyath, A. M., Wang, M., Whiteaker, J. R., Zimmerman, L. J., Fenyo, D., Carr, S. A., Fisher, S. J., Gibson, B. W., Mesri, M., Neubert, T. A., Regnier, F. E., Rodriguez, H., Spiegelman, C., Stein, S. E., Tempst, P., and Liebler, D. C. (2010) Interlaboratory study characterizing a yeast performance standard for benchmarking LC-MS platform performance. *Mol. Cell. Proteomics* **9,** 242–254
5. Campos, A., Díaz, R., Martínez-Bartolomé, S., Sierra, J., Gallardo, O., Sabidó, E., López-Lucendo, M., Casal, J. I., Pasquarello, C., Scherl, A., Chiva, C., Borras, E., Odena, A., Elortza, F., Azkargorta, M., Ibarrola, N., Canals, F., Albar, J. P., and Oliveira, E. (2015) Multicenter experiment for quality control of peptide-centric LC-MS/MS analysis – a longitudinal performance assessment with nLC Coupled to Orbitrap MS analyzers. *J. Proteomics in press*
6. Díaz, R., Martínez-Bartolomé, S., Gallardo, O., Canals, F., Albar, J. P., de Oliveira, E., ProteoRed, I. C., and Campos, A. (2012) ProteoRed multicenter experiment for long-term quality control evaluation of proteomics core facilities. *J. Biomol. Tech.* **23,** S52-S52
7. Rudnick, P. A., Clauser, K. R., Kilpatrick, L. E., Tchekhovskoi, D. V., Neta, P., Blonder, N., Billheimer, D. D., Blackman, R. K., Bunk, D. M., Cardasis, H. L., Ham, A. J., Jaffe, J. D., Kinsinger, C. R., Mesri, M., Neubert, T. A., Schilling, B., Tabb, D. L., Tegeler, T. J., Vega-Montoto, L., Variyath, A. M., Wang, M., Wang, P., Whiteaker, J. R., Zimmerman, L. J., Carr, S. A., Fisher, S. J., Gibson, B. W., Paulovich, A. G., Regnier, F. E., Rodriguez, H., Spiegelman, C., Tempst, P., Liebler, D. C., and Stein, S. E. (2010) Performance metrics for liquid chromatography-tandem mass spectrometry systems in proteomics analyses. *Mol. Cell. Proteomics* **9,** 225–241
8. Omenn, G. S., States, D. J., Adamski, M., Blackwell, T. W., Menon, R., Hermjakob, H., Apweiler, R., Haab, B. B., Simpson, R. J., Eddes, J. S., Kapp, E. A., Moritz, R. L., Chan, D. W., Rai, A. J., Admon, A., Aebersold, R., Eng, J., Hancock, W. S., Hefta, S. A., Meyer, H., Paik, Y. K., Yoo, J. S., Ping, P., Pounds, J., Adkins, J., Qian, X., Wang, R., Wasinger, V., Wu, C. Y., Zhao, X., Zeng, R., Archakov, A., Tsugita, A., Beer, I., Pandey, A., Pisano, M., Andrews, P., Tammen, H., Speicher, D. W., and Hanash, S. M. (2005) Overview of the HUPO Plasma Proteome Project: results from the pilot phase with 35 collaborating laboratories and multiple analytical groups, generating a core dataset of 3020 proteins and a publicly-available database. *Proteomics* **5,** 3226–3245
9. Klimek, J., Eddes, J. S., Hohmann, L., Jackson, J., Peterson, A., Letarte, S., Gafken, P. R., Katz, J. E., Mallick, P., Lee, H., Schmidt, A., Ossola, R., Eng, J. K., Aebersold, R., and Martin, D. B. (2008) The standard protein mix database: a diverse data set to assist in the production of improved peptide and protein identification software tools. *J. Proteome Res.* **7,** 96–103
10. Addona, T. A., Abbatiello, S. E., Schilling, B., Skates, S. J., Mani, D. R., Bunk, D. M., Spiegelman, C. H., Zimmerman, L. J., Ham, A. J., Keshishian, H., Hall, S. C., Allen, S., Blackman, R. K., Borchers, C. H., Buck, C., Cardasis, H. L., Cusack, M. P., Dodder, N. G., Gibson, B. W., Held, J. M., Hiltke, T., Jackson, A., Johansen, E. B., Kinsinger, C. R., Li, J., Mesri, M., Neubert, T. A., Niles, R. K., Pulsipher, T. C., Ransohoff, D., Rodriguez, H., Rudnick, P. A., Smith, D., Tabb, D. L., Tegeler, T. J., Variyath, A. M., Vega-Montoto, L. J., Wahlander, A., Waldemarson, S., Wang, M., Whiteaker, J. R., Zhao, L., Anderson, N. L., Fisher, S. J., Liebler, D. C., Paulovich, A. G., Regnier, F. E., Tempst, P., and Carr, S. A. (2009) Multisite assessment of the precision and reproducibility of multiple reaction monitoring-based measurements of proteins in plasma. *Nat. Biotechnol.* **27,** 633–641
11. Prakash, A., Rezai, T., Krastins, B., Sarracino, D., Athanas, M., Russo, P., Ross, M. M., Zhang, H., Tian, Y., Kulasingam, V., Drabovich, A. P., Smith, C., Batruch, I., Liotta, L., Petricoin, E., Diamandis, E. P., Chan, D. W., and Lopez, M. F. (2010) Platform for establishing interlaboratory reproducibility of selected reaction monitoring-based mass spectrometry peptide assays. *J. Proteome Res.* **9,** 6678–6688
12. Varjosalo, M., Sacco, R., Stukalov, A., van Drogen, A., Planyavsky, M., Hauri, S., Aebersold, R., Bennett, K. L., Colinge, J., Gstaiger, M., and Superti-Furga, G. (2013) Interlaboratory reproducibility of large-scale human protein-complex analysis by standardized AP-MS. *Nat. Methods* **10,** 307–314
13. Wang, X., Chambers, M. C., Vega-Montoto, L. J., Bunk, D. M., Stein, S. E., and Tabb, D. L. (2014) QC metrics from CPTAC raw LC-MS/MS data interpreted through multivariate statistics. *Anal. Chem.* **86,** 2497–2509
14. Holman, J. D., Tabb, D. L., and Mallick, P. (2014) Employing ProteoWizard to convert raw mass spectrometry data. *Curr. Protoc. Bioinformatics* **46,** 13 24 11–19
15. French, W. R., Zimmerman, L. J., Schilling, B., Gibson, B. W., Miller, C. A., Townsend, R. R., Sherrod, S. D., Goodwin, C. R., McLean, J. A., and Tabb, D. L. (2015) Wavelet-based peak detection and a new charge inference procedure for MS/MS implemented in ProteoWizard's msConvert. *J. Proteome Res.* **14,** 1299–1307
16. Tabb, D. L., Fernando, C. G., and Chambers, M. C. (2007) MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *J. Proteome Res.* **6,** 654–661
17. Zhang, B., Chambers, M. C., and Tabb, D. L. (2007) Proteomic parsimony through bipartite graph analysis improves accuracy and transparency. *J. Proteome Res.* **6,** 3549–3557
18. Ma, Z. Q., Polzin, K. O., Dasari, S., Chambers, M. C., Schilling, B., Gibson, B. W., Tran, B. Q., Vega-Montoto, L., Liebler, D. C., and Tabb, D. L. (2012) QuaMeter: multivendor performance metrics for LC-MS/MS proteomics instrumentation. *Anal. Chem.* **84,** 5845–5850