Within-Day Reproducibility of an HPLC–MS-Based Method for Metabonomic Analysis: Application to Human Urine

ARTICLE *in* JOURNAL OF PROTEOME RESEARCH · SEPTEMBER 2007

Impact Factor: 4.25 · DOI: 10.1021/pr070183p · Source: PubMed

4 AUTHORS, INCLUDING:

Helen Gika
Aristotle University of Thessaloniki
39 PUBLICATIONS **1,821** CITATIONS
SEE PROFILE

Georgios Theodoridis
Aristotle University of Thessaloniki
139 PUBLICATIONS **3,798** CITATIONS
SEE PROFILE

Ian D Wilson
Imperial College London
491 PUBLICATIONS **18,163** CITATIONS
SEE PROFILE

research articles Journal of proteome research

Within-Day Reproducibility of an HPLC−MS-Based Method for Metabonomic Analysis: Application to Human Urine

Helen G. Gika,[†,‡] Georgios A. Theodoridis,[†,‡] Julia E. Wingate,[§] and Ian D. Wilson*[,†]

AstraZeneca, Department of Drug Metabolism and Pharmacokinetics, Mereside, Alderley Park, Macclesfield, Cheshire SK10 4TG, United Kingdom, Laboratory of Analytical Chemistry, Department of Chemistry, Aristotle University of Thessaloniki 541 24, Greece, and Applied Biosystems, Concord, Ontario, Canada L4K 4V8

Received March 30, 2007

Self-evidently, research in areas supporting "systems biology" such as genomics, proteomics, and metabonomics are critically dependent on the generation of sound analytical data. Metabolic phenotyping using LC−MS-based methods is currently at a relatively early stage of development, and approaches to ensure data quality are still developing. As part of studies on the application of LC−MS in metabonomics, the within-day reproducibility of LC−MS, with both positive and negative electrospray ionization (ESI), has been investigated using a standard "quality control" (QC) sample. The results showed that the first few injections on the system were not representative, and should be discarded, and that reproducibility was critically dependent on signal intensity. On the basis of these findings, an analytical protocol for the metabonomic analysis of human urine has been developed with proposed acceptance criteria based on a step-by-step assessment of the data. Short-term sample stability for human urine was also assessed. Samples were stable for at least 20 h at 4 °C in the autosampler while queuing for analysis. Samples stored at either −20 or −80 °C for up to 1 month were indistinguishable on subsequent LC−MS analysis. Overall, by careful monitoring of the QC data, it is possible to demonstrate that the "within-day" reproducibility of LC−MS is sufficient to ensure data quality in global metabolic profiling applications.

Keywords: metabolite profiling • metabonomics • metabolomics • reproducibility • quality controls • stability • urine

1. Introduction

Metabonomics, defined as "the quantitative measurement of the dynamic multiparametric response of living systems to pathophysiological stimuli or genetic modification",[1,2] has become an accepted means of finding potential biomarkers of toxicity and disease. Much of the original work in metabonomics was performed using NMR spectroscopy,[3−5] but increasingly, as a result of widespread availability of such systems, there has been a trend toward employing chromatographic techniques such as LC and GC coupled with mass spectrometry[6−10] (recently reviewed in ref 11). While NMR spectroscopy has been shown to be a relatively stable and reproducible analytical platform (for example, see ref 12), very little has been published on the reproducibility of LC− and GC−MS-based methods. However, the ability to demonstrate that metabonomics data is of high quality is crucial if these studies are to provide sound biological insights into the processes of development, disease, and toxicity. In addition, for LC−MS-based metabonomics to become a useful partner with, for examle, parallel proteomic and genomic investigations for "systems biology" studies, data quality is paramount. Similarly, high-quality data are essential if the results are to be used in subsequent submissions to regulatory authorities. It is therefore necessary to conduct studies to explore the potential, and limitations, of, for example, LC−MS-based methods to determine the important variables that must be controlled to ensure that valid data are obtained. Currently, demonstrating that the variability of metabonomic data generated by LC−MS is within acceptable limits, especially when large samples sets are analyzed in long analytical runs, represents a challenge. Here, we present the results of studies examining the reproducibility of LC−MS for the metabonomic analysis of human urine. Reproducibilty was assessed using a "biological QC"[13] approach to investigate the stability of the analytical methodology within a run as well as looking at the repeat analysis of "unknowns". The aim was to mimic the type of analysis undertaken on a relatively small clinical or toxicological study where the number of incurred samples would be expected to be in the region of 60−100. In such studies, a typical analytical run of less than

* Author for correspondence. E-mail, Ian.Wilson@astrazeneca.com; tel, +044 1625 513424; fax, +044 1625 518788.
† AstraZeneca.
‡ Aristotle University of Thessaloniki.
§ Applied Biosystems.

10.1021/pr070183p CCC: $37.00

Published on Web 07/11/2007

24 h would be performed, with each sample analyzed, in separate runs, using both positive and negative electrospray ionization (ESI). The investigations reported here, therefore, provide an indication of within-day analytical performance. Sample stability on storage was also determined.

2. Experimental Procedures

2.1. Reagents and Materials. All solvents used were of HPLC grade and obtained from Fisher Scientific (Loughborough, Leicestershire, U.K.). Water (18.2 MΩ) was obtained from a Purelab Ultra system from Elga (Bucks, U.K.). Formic acid of analytical grade was purchased from Fisher Scientific, while all other used standards were of analytical or higher grade and were obtained from Sigma-Aldrich (Gillingham, Dorset, U.K.).

2.2. Samples. Mid-stream urine samples were collected into a 500 mL sterile polypropylene screw-top vessel (Corning, Hemel Hempstead, U.K.) from both normal male and female volunteers. Samples were collected between 9:00 and 11:00 a.m. The volunteer group consisted of 30 postmenopausal females and 30 males over 50 years of age, and were fasted for at least 14 h, with the exception of water, prior to sample collection. A volume of 150 mL of urine was decanted into a separate 500 mL sterile polypropylene screw-top container, containing 1× protease tablet (Roche, Welwyn Garden City, U.K.). This container was swirled thoroughly to ensure that the tablet was completely dissolved. The samples were stored on wet-ice (for a maximum of 1 h) prior to freezing and long-term storage at −80 °C in 10 mL aliquots. When the samples were defrosted for analysis, they were first centrifuged at 13 000 rpm for 10 min to remove any particulates; then, a 700 μL aliquot of the supernatant was diluted with 2800 μL of 0.1% aqueous formic acid, and 200 μL of it was placed in a autosampler vial (Kinesis, Cambridgeshire, U.K.)

A "quality control" (QC) sample was prepared by mixing equal volumes (100 μL) from each of the 60 samples as they were being aliquoted for analysis. This "pooled" urine was used to provide a representative "mean" sample containing all the analytes that will be encountered during the analysis.

All samples were centrifuged at 3000 rpm for 10 min to remove particulates immediately prior to loading into the autosampler tray, which was maintained at 4 °C for the duration of the analysis.

Two different standard mixture solutions, each of four compounds, were used to determine initial "system suitability" and to provide a first line means of monitoring the quality and stability of the separation. The mixture used for positive ESI consisted of uridine, nicotinic acid, tryptophan, and hydroxyhippuric acid at concentrations of 16.67, 20.00, 19.33, and 15.33 μg·mL^{-1}, respectively. A mixture of raffinose, phenylalanine, glycocholic acid, and taurodeoxycholic acid at concentrations of 21.86, 20.00, 19.67, and 19.90 μg·mL^{-1}, respectively, was used for negative ESI.

2.3. LC/ESI−MS Analysis.

2.3.1. RP-Liquid Chromatography. Chromatography was performed on a Symmetry C18 3.5 μm (2.1 × 100 mm) column (Waters Ltd, Elstree, U.K.) using a Perkin-Elmer series 200 high-pressure LC micro solvent delivery system (Perkin-Elmer Life Sciences, Cambridge, U.K.) and a CTC PAL autosampler (CTC Analytical, Switzerland). Temperature was maintained constant at 40 °C with a column heater controller of ±0.2 °C temperature stability (Jones Chromatography, Hengoed, U.K.). The separation was performed using gradient elution with 0.1% formic acid in water (A) and 0.1%formic acid in acetonitrile (B) as mobile phases at a flow rate of 400 μL/min. The gradient started with 100% A for 0.5 min, changing to 80−20% A−B over the next 3.5 min and to 5−95% A−B over the next 4 min. The composition was then held constant at 95% B for 1 min, after which the solvent composition was changed back to 100% A over 0.1 min. Injection of 10 μL of each sample was done after 2 min equilibration time at 100% A. The needle filling speed was set at 5 μL/s and injection speed at 25 μL/s.

The injection system was subjected to two washing cycles with a strong (a mixture of 30:30:40 methanol/propan-2-ol/acetonitrile-0.1% formic acid) and a weak solvent (water and 0.1% formic acid) after each injection, and one cycle prior to each injection to minimize eventual carryover. Blank runs, 5 in total, were also carried out randomly between samples to allow chromatographic carryover to be examined.

2.3.2. ESI Mass Spectrometry. Mass spectrometry was carried out on a Hybrid Triple Quadrupole Linear Ion Trap system, 4000 QTRAP, (Applied Biosystems|MDS Sciex, Warrington, U.K.). Electrospray Ionization mode was applied by a TurboIonSpray inlet operating at 350 °C in positive and negative ESI modes in separate experiments. Enhanced mass scan data were acquired using the Q3 as Linear Ion Trap for 12 min over a range of 100−850 m/z with a scan rate of 1000 amu/s and a step size of 0.08 amu. TurboIonSpray Voltage was set at ±4500 V, curtain gas at 20 psi, auxiliary gases at 40 psi, declustering potential at ±80 V, entrance potential at ±10 V, and collision gas was set at medium flow, and collision energy at 10 eV. Q3 Entry Barrier was set at 8.00 V while the Ion Trap was operated in the Dynamic Fill time mode.

2.3.3. Sample Analysis. Each of the test samples were analyzed by LC−MS in both positive and negative ionization (ESI) mode to obtain metabolite profiles. An appropriate test mix of standard compounds (see above) was analyzed in the beginning, in the middle, and at the end of each run, which in total lasted 17 h. The pooled "QC" sample was injected 4 times at the beginning of the run to ensure system equilibration and then every 10 samples to further monitor the stability of the analysis. In addition to examining the QC samples using principal components analysis (PCA) (see below), the peaks collected with the peak finding software were checked to define the overall amount of variability of the run, whereas a subset of peaks in the QCs, covering a range of retention times, masses, and intensities, were examined from their ion-extracted chromatograms in more detail to evaluate the stability of the run.

In addition to investigating system stability and reproducibility via the QC samples, a separate experiment was performed on 4 samples (2 from female and 2 from male subjects) to further explore stability and reproducibility. These 4 samples were analyzed 12 times using positive ESI, and a further set of 4 samples were analyzed in the same manner using negative ESI.

2.3.4. Sample Stability. 2.3.4.1. Storage Stability at −20 and −80 °C. Sample stability was determined for up to 1 month at −20 and −80 °C. For this study, additional urine samples from 6 healthy adult males were collected and divided into 9 aliquots (of 500 μL). One set of samples was analyzed half an hour after the collection time, while the remainder was stored at either −20 or −80 °C. Sample analysis was then repeated at 1 week after the collection and 1 month. The analysis was carried out in positive and negative ESI using the same sample preparation and QC procedures as described in section 2.2.

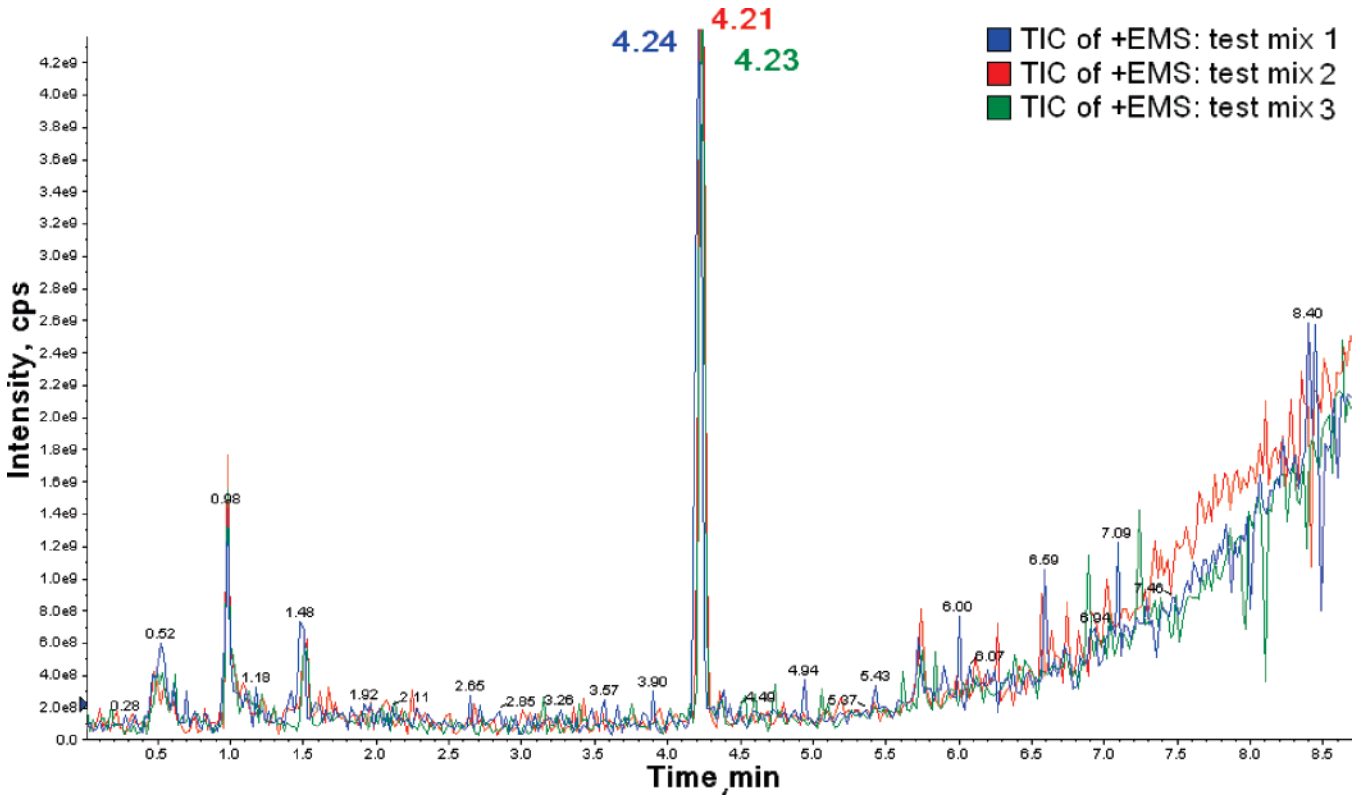


Figure 1. Overlaid EMS chromatograms of the test compounds mix analyzed in positive ESI mode at the beginning, middle, and end of the run giving an overall idea of the system stability during the run.

2.3.4.2. Freeze−Thaw Stability. Freeze−thaw stability was determined using a further subset of the urine samples collected for the 1 month stability test. Nine aliquots of each of these urine samples were stored at −20 °C and then underwent from 1 to 9 freeze−thaw cycles whereby aliquots were taken from the freezer and allowed to thaw at room temperature (ca. 3 h) before refreezing prior to analysis of all nine sets of sample as a single batch.

After the final cycle of freeze-thawing had been completed, the samples were prepared as described in section 2.2 and analyzed using both positive and negative ESI mode.

2.4. Data Analysis. MarkerView software version 1.1.0.1 was used to process the raw spectrometric data acquired by Analyst 1.4.1 software (Applied Biosystems|MDS Sciex). MarkerView allows the data from a set of samples to be compared. In the first step, peaks are located in the spectra by a spectral peak-finding algorithm. Masses belonging to the same peak 'cluster' are merged together with a resulting area. They are then aligned according to retention times if their masses differ by less than the specified tolerance. Each spectrum is first background-subtracted by the mass spectrum 10 scans before the current one such that constant background ions are not collected as peaks. The parameters for peak finding, alignment, and filtering were set as follows: noise threshold was set at 10^4, minimum spectral peak width at 0.25 amu, minimum RT peak width at 3 scans, maximum RT peak width at 100 scans, retention time tolerance at 0.5 min, mass tolerance at 0.25 amu, and maximum number of peaks at 8000.

Peak list data obtained by MarkerView were further processed by Simca P version 11 from Umetrics (Windsor, U.K.) for multivariate data analysis. Basic applications such as Principal Components Analysis and other statistical tools were implemented.

3. Results and Discussion

3.1 Assessment of the Use of QC Samples in Metabonomic Analysis. The reliable multicomponent analysis of complex biological samples such as urine and plasma via HPLC−MS-based methods provides a number of challenges with respect to obtaining valid data. Not the least of these is in ensuring that the system has not changed with regard to characteristics such as detector response and retention time during the run. Here, we have used a structured approach based on the use of standard mixtures of test compounds and a pooled urine "QC" sample to enable us to demonstrate that the LC−MS system was providing useful and reliable data. Thus, the test mixtures of pure standards provided an initial screen whereby a very rapid visual examination of the performance of the system was obtained by overlay and comparison of the 3 runs performed at the beginning, middle, and end of the batch. Clearly, any dramatic change, or deterioration, in either chromatographic or detector performance would have been immediately evident in the form of changes in retention time, peak shape, peak height/area, and mass accuracy. In this instance, signal intensity and peak shape were unchanged, and the maximum deviation in retention time for all of the test compounds was 0.03 min, which was considered to be acceptable for this type of work. Similarly, for both negative and positive ionization, the maximum mass accuracy deviation was 0.1 amu. These results were used to define the retention time and mass tolerance parameters in the MarkerView software for the alignment of the data. The LC−MS TICs obtained for the test compounds in positive ESI are shown in Figure 1.

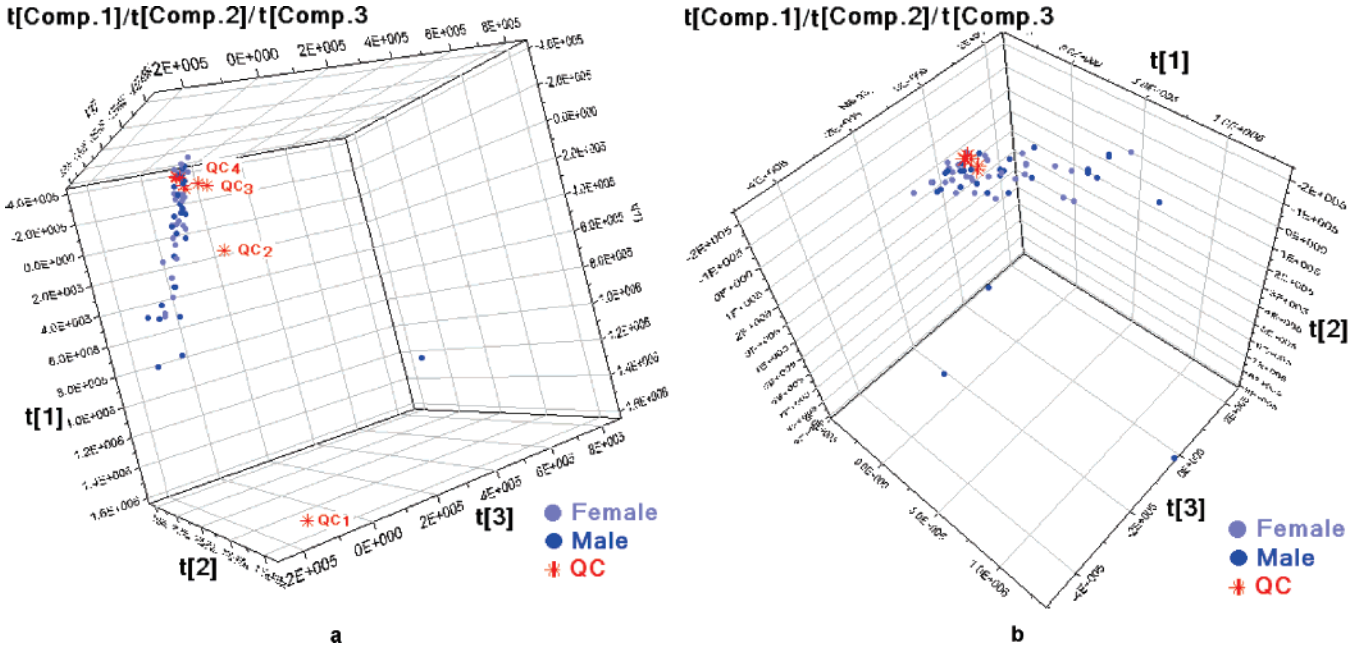


Figure 2. PCA scores plot (Comp. 1 vs Comp. 2 vs Comp. 3) of (a) all samples analyzed in positive ESI mode and (b) of the same set of samples after the 4 firsts runs were excluded.

As we have noted elsewhere,[13,14] it is our experience that, for biological samples such as urine, the LC−MS system does not usually provide reproducible results for the first few injections. For this reason, it is our usual practice to run several (usually 4 to 5) QC samples prior to the start of the main analytical run to "condition" the system and to be able to demonstrate that it has achieved stability. Presumably, the changes in retention time which were the main cause of the first few QC samples in each run being outliers were the result of the chromatographic system becoming equilibrated to the samples (possibly through the masking of active sites, etc.). Interestingly, it seems that there is an absolute requirement for the injection of biological samples to achieve retention time stability for the urinary components, at least in the system described here. Thus, the repeated injection (up to 8 cycles) of the mixture of pure standards prior to the start of analysis of the QCs was ineffective in "conditioning" the system with several subsequent injections of the pooled urine required to attain reproducible retention times (data not shown).

The instability of the system was reflected in shifts in retention times from one chromatogram to the next, and produced a very characteristic pattern for the QC samples on PCA. Thus, as shown in Figure 2a, if all of the QC samples are included in the PCA, the first block of 3 or 4 are always outliers in the PCA scores plot. Here, PCA was performed using pareto scaling of the raw data obtained in positive ESI mode. Obviously, these "outlier" samples should be excluded in the final analysis of the data, but this initial analysis does allow the facile demonstration that a degree of platform stability (in this instance retention times) had been attained prior to the commencement of the analysis of the main batch of unknown samples. PCA of the remaining QCs reveals the pattern shown in Figure 2b, which gives an indication of the reliability of the data. Thus, the assumption is that the tighter the clustering of the QC samples in the scores plot, the more repeatable the runs were. If this assumption is correct, it should then mean that differences between the test samples from different individuals are likely to reflect real differences in metabolite profiles (i.e., biological variation) rather than analytical variation. Obviously, as we have discussed elsewhere,[14] if the analytical method were perfectly reproducible, all of the QCs would be identical, and it is clear from the PCA that there is some variability between runs. In these earlier studies, we found that much of the variation encountered in the QC samples was related to variability in detector response for peaks at or near the threshold used for assessing whether a peak was present or not.[14] The current work was performed with the revised Markerview software, which is more robust with respect to this as a source of variability.

PCA was also performed on the 10 QC samples separately. As the scores represent weighted average trajectories of the original variables, exploring their time dependence gives good insight into trends and shifts with time. In Figure 3a the time series properties of the first component are depicted showing how the QC samples behaved as the run progressed. This type of result gives some confidence that the analysis was stable for the duration of the run and provides a further, pragmatic means of assessing the quality of the data, and deciding if it is sufficient to warrant further statistical analysis of the results to detect biomarkers.

In a similar way, when PCA was performed on data from the negative ESI method, the first 2 QC samples of the run were identified as outliers, while all the others were strongly clustered. As with the positive ESI data, these initial outliers were excluded from the analysis when the within-run stability of the method was examined. In Figure 3b, the time series properties of the first component of the negative ESI data are illustrated, showing a similar trend to the positive ESI data.

Another possible source of variability is the potential for late eluting components from earlier runs appearing as peaks in subsequent samples. This sort of chromatographic "cross talk" between samples obviously has the potential to be a confounding factor in this type of analysis. However, examination of the blank injections performed throughout the run did not show the presence of carryover, or long running peaks, from one run to the next for either positive or negative ionization mode.

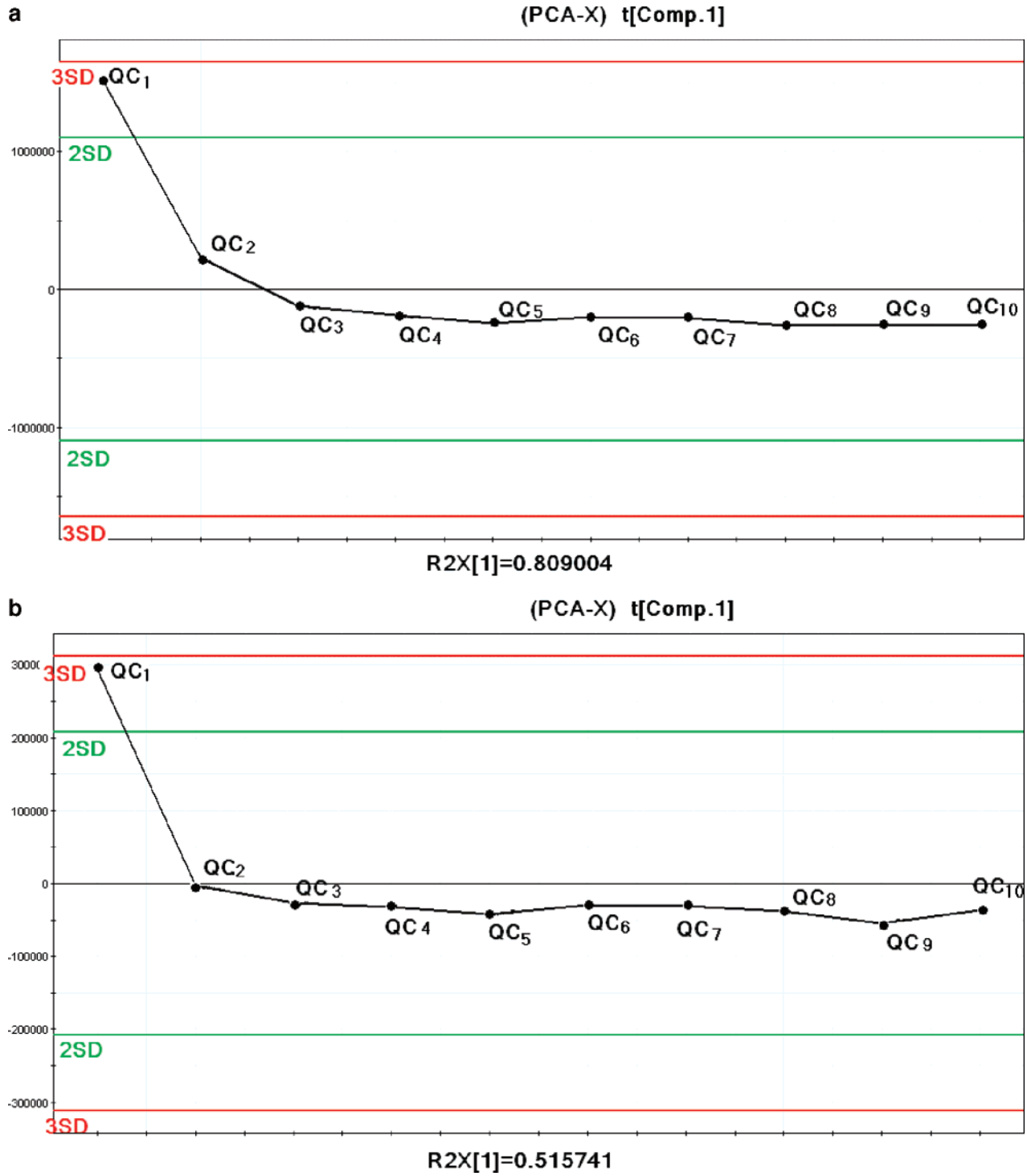


Figure 3. PCA first component for the QC samples versus time analyzed in the (a) positive and (b) negative ESI.

Finally, we considered the effect of the time allowed for re-equilibration of the column at the end of each analytical run on retention time reproducibility, particularly with respect to the early eluting peaks (i.e., those most likely to be affected by inadequate equilibration times). Clearly, there is a tension between achieving the highest analytical throughput and ensuring that the chromatographic phase has time to re-equilibrate with the starting mobile phase following the end of the previous run as this period is essentially analytical "dead time" with respect to sample analysis. The temptation is therefore to use, short, and possibly inadequate, re-equilibration conditions to maximize sample analysis. We therefore examined the effect of using re-equilibration times of 1, 2, 3, and 4 min on the retention time reproducibility of a range of peaks, and on the overall QC data throughout the run. To these, a further ca. 1 min must be added to take into account the injection cycle itself, which included several washing steps, so that the combination of both equilibration and injection cycles resulted in the column being exposed to between ca. 2 and 5 column volumes (ca. 350 μL) of the starting run buffer using 1−4 min equilibration times. These experiments showed that providing the system had been "pre-conditioned" by the

Table 1. Variation in Retention Times and *m*/*z* Values for Some Selected Peaks in Positive and Negative ESI

(a) Positive ESI Mode

	peak 1			peak 2			peak 3			peak 4			peak 5		
	RT (min)	parent ion	product ion	RT (min)	parent ion	product ion	RT (min)	parent ion	product ion	RT (min)	parent ion	product ion	RT (min)	parent ion	product ion
QC4	0.99	167.1	153.1	4.06	167.0	151.1	4.2	204.1	187.1	5.17	180.1	105.0	5.2	265.2	248.2
QC5	0.98	167.1	153.0	4.08	167.1	151.1	4.19	204.0	187.1	5.15	180.1	105.0	5.19	265.1	248.2
QC6	1.01	167.0	153.0	4.06	167.1	151.1	4.19	204.1	187.0	5.15	180.1	105.0	5.17	265.2	248.2
QC7	0.99	167.0	153.1	4.06	167.1	151.1	4.21	204.1	187.0	5.17	180.1	105.0	5.21	265.2	248.2
QC8	1.01	167.0	153.0	4.09	167.1	151.1	4.2	204.0	187.0	5.17	180.1	105.0	5.23	265.1	248.2
QC9	1.00	167.1	153.0	4.08	167.1	151.0	4.21	204.0	187.0	5.17	180.1	105.0	5.22	265.1	248.2
QC10	1.01	167.0	153.1	4.07	167.1	151.1	4.22	204.0	187.1	5.17	179.9	105.0	5.23	265.1	248.2
Variation	**0.03**	**0.1**	**0.1**	**0.03**	**0.1**	**0.1**	**0.04**	**0.1**	**0.1**	**0.04**	**0.2**	**0.0**	**0.06**	**0.1**	**0.0**

(b) Negative ESI Mode

	peak 1		peak 2		peak 3			peak 4		peak 5		peak 6		peak 7	
	RT (min)	parent ion	RT (min)	parent ion	RT (min)	parent ion	product ion	RT (min)	parent ion	RT (min)	parent ion	RT (min)	parent ion	RT (min)	parent ion
QC4	0.67	181.1	4.41	357.1	5.17	178.1	134.2	6.34	481.4	6.81	465.4	7.50	212.2	8.67	187.0
QC5	0.70	181.2	4.37	357.1	5.13	178.1	134.2	6.33	481.4	6.81	465.4	7.47	212.1	8.59	187.0
QC6	0.68	181.2	4.42	357.2	5.17	178.0	134.2	6.35	481.3	6.83	465.4	7.54	212.1	8.66	187.0
QC7	0.74	181.2	4.40	357.0	5.17	178.2	134.1	6.34	481.3	6.82	465.4	7.52	212.1	8.64	187.0
QC8	0.70	181.2	4.44	357.3	5.18	178.1	134.2	6.35	481.4	6.82	465.4	7.54	212.1	8.65	187.0
QC9	0.73	181.2	4.42	357.3	5.18	178.1	134.0	6.35	481.3	6.83	465.3	7.54	212.1	8.65	187.0
QC10	0.73	181.2	4.44	357.3	5.19	178.2	134.2	6.34	481.4	6.80	465.4	7.55	212.1	8.64	187.0
Variation	**0.07**	**0.1**	**0.07**	**0.3**	**0.06**	**0.2**	**0.2**	**0.02**	**0.1**	**0.03**	**0.1**	**0.08**	**0.1**	**0.08**	**0.0**

injection of a number of QC samples there was essentially no difference between the retention time reproducibility data for the early eluting peaks irrespective of whether a 1, 2, 3, or 4 min equilibration time was used. Similarly, the overall QC data was not affected by the length of the re-equilibration period studied here. The choice of 2 min (ca. 3 column volumes when the injection cycle is included) as the re-equilibration time used for this work was therefore based on the pragmatic basis that it was well within the envelope of the conditions examined but was nevertheless relatively short, thereby giving a good compromise between ensuring proper equilibration and sample throughput.

While the QC data, following the exclusion of the first few samples, showed good clustering, there was some variability, and we therefore undertook a more detailed examination of the results to try determining the cause(s). The obvious variables would include retention time (small random variations or systematic drift) and detector response (again random variation or systematic drift). We first examined whether peak alignment was an issue due to the selection of an inappropriate retention time window, and therefore, the effects of retention time tolerance on the subsequent data analysis were tested. Retention time tolerances of 0.2, 0.5, 0.75, and 1 min were used; however, the PC1 and PC2 scores were very similar in all cases, and the score plots obtained were almost identical. Manual inspection of these data showed that in some instances a retention time shift of 0.2 min or more, for some late eluting analytes, was occasionally observed and, as a result, a 0.5 min retention time tolerance was used for data processing. Under these conditions, ca. 4300 ions were detectable in the QC samples using positive ESI and ca. 3600 in negative ESI.

A small subset of five peaks present in the QC samples, covering a range of retention times and signal intensities, were then examined in detail via their extracted ion chromatograms as a further means of screening the QC raw data, prior to processing with the peak finding algorithm. This enabled us to determine whether the retention time, the response, and the mass changed over the course of the run. In positive ESI mode, the extracted ion chromatograms for an early peak eluting at 1 min (*m*/*z*: 167.1, 153.1), another at 4.06 min (*m*/*z*: 166.2, 151.1), the most intense peak eluting at 4.20 min (*m*/*z*: 204.1, 187.1), the hippurate peak at 5.17 min (*m*/*z*: 180.1, 105.0), and finally one at 5.23 min (*m*/*z*: 265.1, 248.2) were examined (Table 1a). This showed that the stability of the retention times for the five components over the 17 h run was good (variation from 0.03 to 0.06 min) and measured mass accuracy also acceptable (up to 0.1 amu). Thus, once the system had come to equilibrium, the main cause of variability over the 17 h of the run therefore appears to have been the detector response. However, when the responses for these peaks were examined, such variations appeared to be random as there was no obvious systematic drift in the signal over the period of the analysis. This was also observed for all ions in the list created by the MarkerView software and is illustrated in Figure 4 for a few of them. This shows the variation in the intensity of these ions in the QCs (using positive ESI) through the run (injection 4−78).

For negative ESI mode, a similar exercise was undertaken on the extracted ion chromatograms of 7 peaks, which eluted at 0.6 min (*m*/*z* 181.1), 4.4 min (*m*/*z* 357.1), 5.2 min (hippurate, *m*/*z* 178.1), 6.3 min (*m*/*z* 481.4), 6.8 min (*m*/*z* 465.4), and at 8.6 min (*m*/*z* 187.0). For these ions, the variability observed in the retention times were not higher than 0.08 min with differences in measured mass not higher that 0.2 amu (Table 1b). As with the positive ESI data, only nonsystematic variation in peak intensity was observed.

In an approach to estimate and define more clearly the variability of the MS signal, in both positive and negative ESI, all the peaks found for the QC samples by the MarkerView software (after excluding the first 3 runs) were examined. This was performed with the aim of determining some criteria for acceptance/rejection of both individual QC samples and the

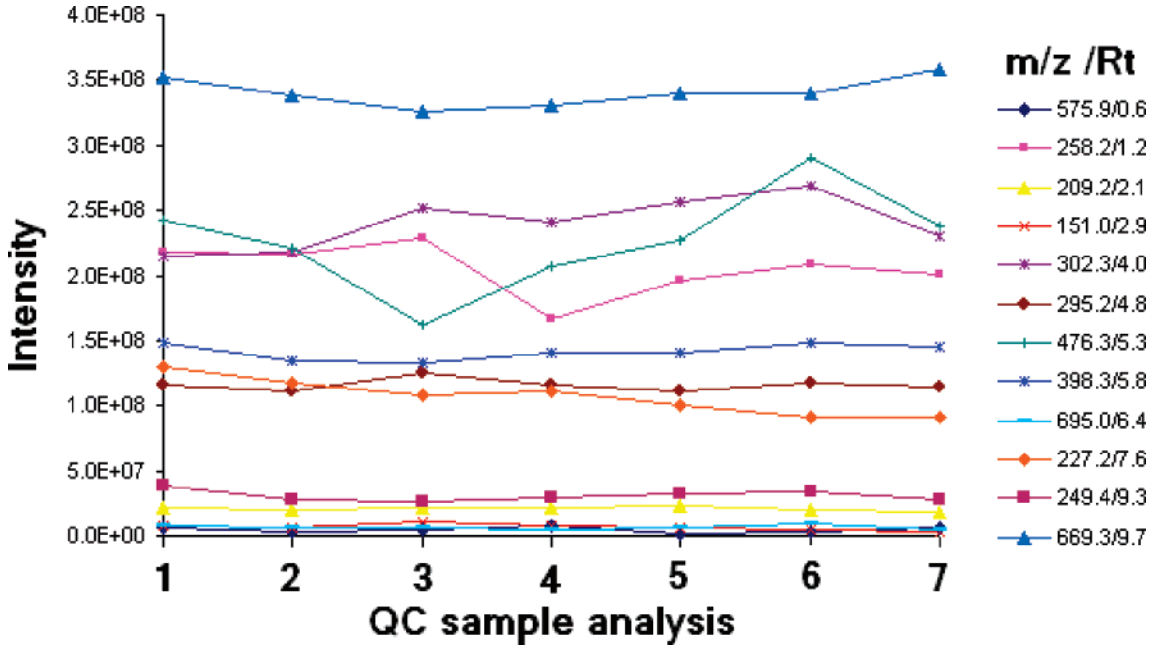


Figure 4. Intensity fluctuation of 15 randomly selected peaks [*(m/z)*/(retention time)] collected by MarkerView software in the 7 QC samples analyzed during the run.

whole analytical run. An examination of the data for the peaks detected in positive ESI mode showed that those with signal intensities ranging in magnitude from 10^7 to 10^8 were the most common (with ions of an intensity of 10^7 and 10^8 representing 41 and 32% of the total of 4288 detected peaks, respectively) (see Figure 5a). We then investigated the variability in these peaks using a range of acceptance criteria. In their guidance for analytical method validation for bioanalytical methods for drugs, the FDA suggest that a variability of ±15% of the nominal value represents an acceptable degree of reproducibility (except at the LOQ where ±20% is considered to be adequate).[15] In addition the FDA, criteria allow 2 QC samples out of a batch of 6 (i.e., 33%) to fall outside the acceptance criteria while still accepting the run. We therefore examined these data using such guidelines.

Thus, for the positive ESI data, taking a figure of ±15% of the average response for each peak as an acceptance criterion, some 64.6% of the peaks, across all of the QCs, fell within the limits (average for the 7 QC samples with a SD of 0.02). When the criteria for acceptance was increased to ±20% of the average, a total of 73% of the peaks were within the limits. As would be expected, loosening the acceptance criteria further, for example, to ±30% and ±50% raised this percentage (to 83.9% and 92.1%, respectively). While giving a broad overview of the reproducibility of the system, these figures do not show the variability of individual peaks across the run. Neither does a "global" analysis of the data like this provide an indication of the quality of the individual QC samples (although the control charts shown in Figures 3a,b provide some reassurance that, following equilibration, all of the 7 QCs used here were similar). As indicated above, the FDA guidance suggests that, in any series of 6 QCs, up to 2 can be outside the control limit with the run still being considered acceptable. Examination of the data across all 7 QCs showed that nearly 60% (57%) of the peaks were within the ±15% limit for 5 or more of the QC samples, with the figure increasing to nearly 70% (67%) if ±20% was used. However, some of the peaks showed a high degree of variability with some 24% out of the ±15% limit for 5 or more QCs (reducing to ca. 15% with a ± 20% acceptance window).

While the reproducibility of the majority of the detected ions appears broadly, indeed surprisingly, acceptable, the degree of variability was not constant over the range of signal intensities detected. By the use of the default that a peak could be considered acceptable if it was out of the set limits for no more than 2 QCs out of 7, then, with a ±15% cutoff, 67.3% of peaks of 10^7 intensity and 85% of peaks of 10^8 intensity were acceptable. At this level, all of the most intense peaks (10^9) passed. In contrast, for peaks with intensities of 10^5 and 10^6, the equivalent figures were 0 and 12%, respectively. Employing a ±20% acceptance criterion raised these percentages to 0.01, 23, 80, and 85% for the 10^5 ,10^6, 10^7, and 10^8 intensity peaks, respectively. This is illustrated graphically in Figure 5a for these positive ESI data, showing both the relative numbers of peaks at each intensity as well as the percentage of peaks found to be acceptable at the ±15 and 20% levels. Clearly, the bulk of the unacceptable results are concentrated in the minority of low-intensity peaks. Thus, the majority of the data set, which is comprised of signals with intensities of 10^7 or above, was less affected by variability, which may allow the reliable detection of differences between samples for these components. However, clearly, the same general conclusion cannot be drawn for the low-intensity analytes that show considerably more variability.

In the case of the negative ESI data, the total ion current was generally lower than that seen for positive ESI, and from the total 3565 peaks that were found by the software, the most abundant were in the intensity ranges 10^5 (27.1%), 10^6 (36.1%), and 10^7 (27.3%). As with the analysis of the positive ESI data, when ±15% of the average response for each component was taken, then 55.2% (average for the 7 QC runs with a SD of 0.06) of the negative ESI data fell within these limits increasing to 64.1, 75.4, and 86.8% when the acceptance limits were raised to ±20, ±30, and ±50%, respectively.

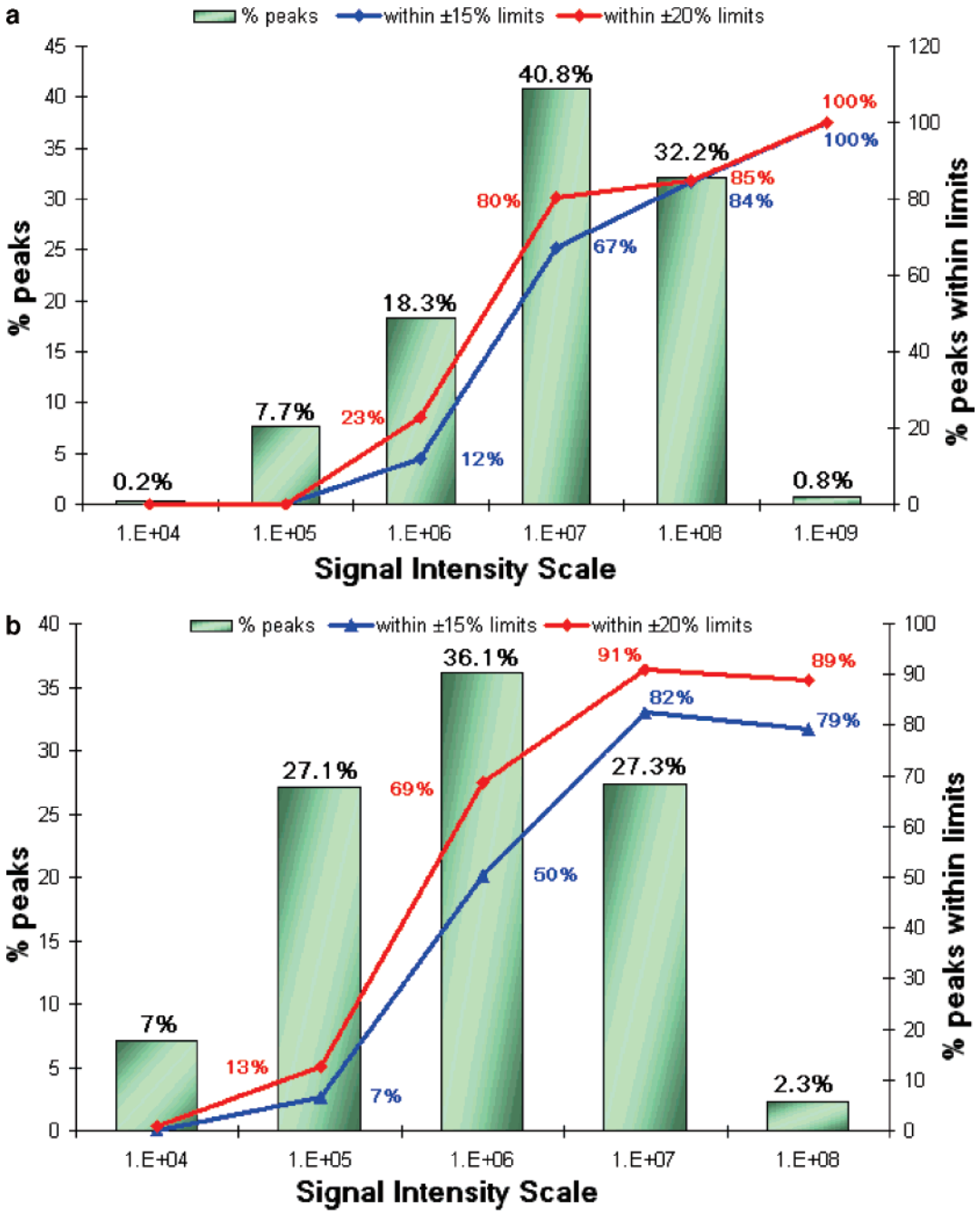


Figure 5. Distribution of peaks found by the software (noise threshold set at 10^4) together with the % acceptable peaks in QC data obtained in (a) positive ESI mode and (b) negative ESI mode determined for two critical limits (±15 and ±20 of the average response).

Across the QCs, 44% of the peaks were within the acceptance limit of ±15% for 5 or more of the 7 QC samples, rising to 55% using ±20%. However, once again, a significant number of peaks failed the ±15% acceptance value in 5 or more of the QCs (31% at ±15% falling to 23% for ±20%).

As with the positive ESI data, signal intensity was found to be the major factor in reproducibility, with much worse precision for signals of low intensity. Thus, when applying the ±15% criterion to assess peaks with intensities in the ranges 10^6, 10^7, and 10^8 some 50.2%, 82.4%, and 79.2%, respectively, fell within the acceptable range (increasing to 68.8% for 10^6, 91.1% for 10^7, and 88.9% for 10^8 when the ±-20% acceptance criterion was employed). However, as seen for positive ESI, low-intensity peaks (10^5) were highly variable with only 6.6% found to be acceptable. Even widening the acceptance criteria to ±20% only increased this to 12.7%. These results are summarized in Figure 5b, which shows the distribution of the peak intensities together with the percentage of the acceptable peaks. The overall higher variability of these results can therefore probably be attributed to generally lower intensities of the peaks obtained in HPLC−MS of these samples using negative ESI. Such results, for both positive and negative ESI, once highlighted, are not particularly surprising, and indeed would be the expected result.

All the above results were based on the analysis of the peak list data collected with the MarkerView software applying the parameters defined earlier. However, a range of different peak finding and alignment parameters were also tested to observe

effects on the quality of the data. As noted above, different retention time tolerances were tested and proved to be without effect. Similarly, a higher noise threshold of 10^5 in positive ESI mode and using a range of 10−300 scans as a chromatographic peak width were again found to have no effect on the overall results and on the amount of variability detected.

In conventional quantitative LC−MS methods, much use is made of internal standards (IS) to improve precision, and generally, these IS are a stable isotope labeled version of the target analyte(s). Clearly, this is not an easy approach to implement in multianalyte assays, especially where most of the components are unknown prior to (and often subsequent to) the analysis. Thus, any internal standard-based approach for metabonomics analysis would have to be based on the addition of only a limited number of compounds that might have little structural resemblance to many of the components of the sample. While it might be expected that the nonsystematic, compound-dependent variability highlighted in Figure 4 would confound such an approach, we nevertheless investigated the use of peaks for, for example, compounds such as hippuric acid as a surrogate IS for the other metabolites in the QC samples. Normalization of the data to such compounds, however, increased the variability in the data, and we conclude that it is difficult to advocate the use of a limited number of IS in this type of application. Similarly, using Median Peak Ratios for normalization of the data failed to improve the result.

3.2. Sample Repeatability. While clearly there was a measure of reproducibility for the QC samples, these obviously, from the way in which they were constructed from all of the sample pool, represent mean concentrations of the analytes present. To ensure that the results obtained for the QCs were representative, we therefore undertook a limited examination of a small number of randomly selected samples. Thus, two groups of 4 samples (2 male and 2 female subjects, respectively) were analyzed repeatedly (12 times) in negative ESI and 4 others in positive ESI mode (with the repeats in random order). All the peaks from the list created by Marker View software were then processed for each sample and were examined as described for the QC samples above. While it might be anticipated, as a result of the differences in metabolite proportions between the samples and QCs, that the overall performance of the HPLC−MS system might be less consistent (especially if a particular sample contained some analytes in much lower concentration), this proved not to be the case. Thus, we observed that that the percentage of peaks within the set limits of ±15% of the average response, for each set of 4 samples, was 63% in positive ESI mode and 50% in negative ESI mode, namely, similar to the percentage found when analyzing the data obtained the QC samples. Similarly, with respect to variability in mass (both positive and negative ESI) and retention time, no differences were seen between these samples and the QC data

3.3. Use of QC Data To "Validate" Data Sets. Clearly, the data derived from the QC samples provides a valuable means of showing that the analytical system has stabilized prior to analyzing the main batch of samples and of assessing the performance of the LC−MS system during an analytical run. The data obtained here, for a typical set of human urine samples, reveal that, as would perhaps be anticipated, signal intensity was the most significant source of variability rather than retention time or changes in mass accuracy. On the basis of these results, it seems reasonable to suggest that for the type of samples studied here, a fairly strict acceptance criterion for the more intense signals of ±15% seems attainable. For lower intensity peaks, a value of ±20% probably represents a fair acceptance criterion. Similarly, based on the FDA guidance, excursions outside these ranges in more than 2 QCs (out of 6) would therefore invalidate an analyte. The question of what criteria would be used to exclude all of the data from a particular QC sample is more complex. Clearly, a poor performing QC should be immediately obvious from examination of the PCA data (as illustrated in Figures 2 and 3). In our opinion, evidence of high variability within the body of the run, such that a QC was an obvious outlier, would constitute a significant reason for concern about the reliability of sample data obtained either side of that QC, and might indeed require the whole run to be repeated.

Practical examples of how the QC data could be used to "validate" the results for a potential biomarker are shown in Figure 6. Here the plot profiles of 2 ions for all the samples including the QCs are given. One of these ions (m/z 326, Figure 6a) was found to provide the most significant contribution to making one of the male subjects an obvious outlier (see Figure 2b). Examination of the behavior of this ion in the QC samples showed it to be within the ±15% limits for the whole run (0 times out of limit in 7 QC runs), and quite clearly, the amount of this particular analyte is much greater in the urine of this individual than the bulk of the population studied. On the basis of the QC data, it seems reasonable to conclude that this is a real difference and that this individual represents a genuine outlier. In Figure 6b, the intersubject variability of the ion m/z 310 is shown together with the QC data for the same ion. Once again, it can be seen that the variability of the intensity of this ion was within the ±15% limits in all QC runs suggesting that the differences observed in the amounts of this ion detected in the samples were real.

A suggested workflow for accepting LC−MS-generated metabonomics data as fit for in-depth, statistical analysis as part of biomarker discovery is shown in Scheme 1. In our view, failure to pass any of these stages should trigger a reanalysis of the sample set.

3.4. Sample Stability. A further set of criteria that must be addressed is sample stability under the conditions in which the samples were maintained during storage and analysis. To try to ensure analyte stability, the samples were kept at 4.0 °C (±0.5) in the autosampler while awaiting analysis. While this procedure cannot ensure the stability of all analytes, the absence of any time-related trends in the data obtained from the QC samples supports the view that the bulk of the sample components were stable over the time course for of the analysis (approximately 20 h) at this temperature.

In addition to stability during analysis, there is a need to establish sample stability during storage. For this, samples from 6 subjects were analyzed after storage for 1 and 4 weeks at −20 and −80 °C as described in the Experimental Procedures. PCA analysis of the data obtained from analysis of these samples, at either −20 or −80 °C, for these time periods compared with the pre-storage data showed the PCA plots to be similar (i.e., samples appeared in the same positions relative to each other before and after storage (either at −20 or −80 °C), implying that the samples were unchanged over a storage period up to 1 month. No differences were noted between samples kept at −20 and −80 °C even after 1 month of storage (Figure 7). These data suggest that storage of urine samples at −20 °C for at least 1 month is acceptable, and that short-term storage at lower temperature confers no particular advantage. Similar results for sample stability have recently been reported for ^{1}H NMR-

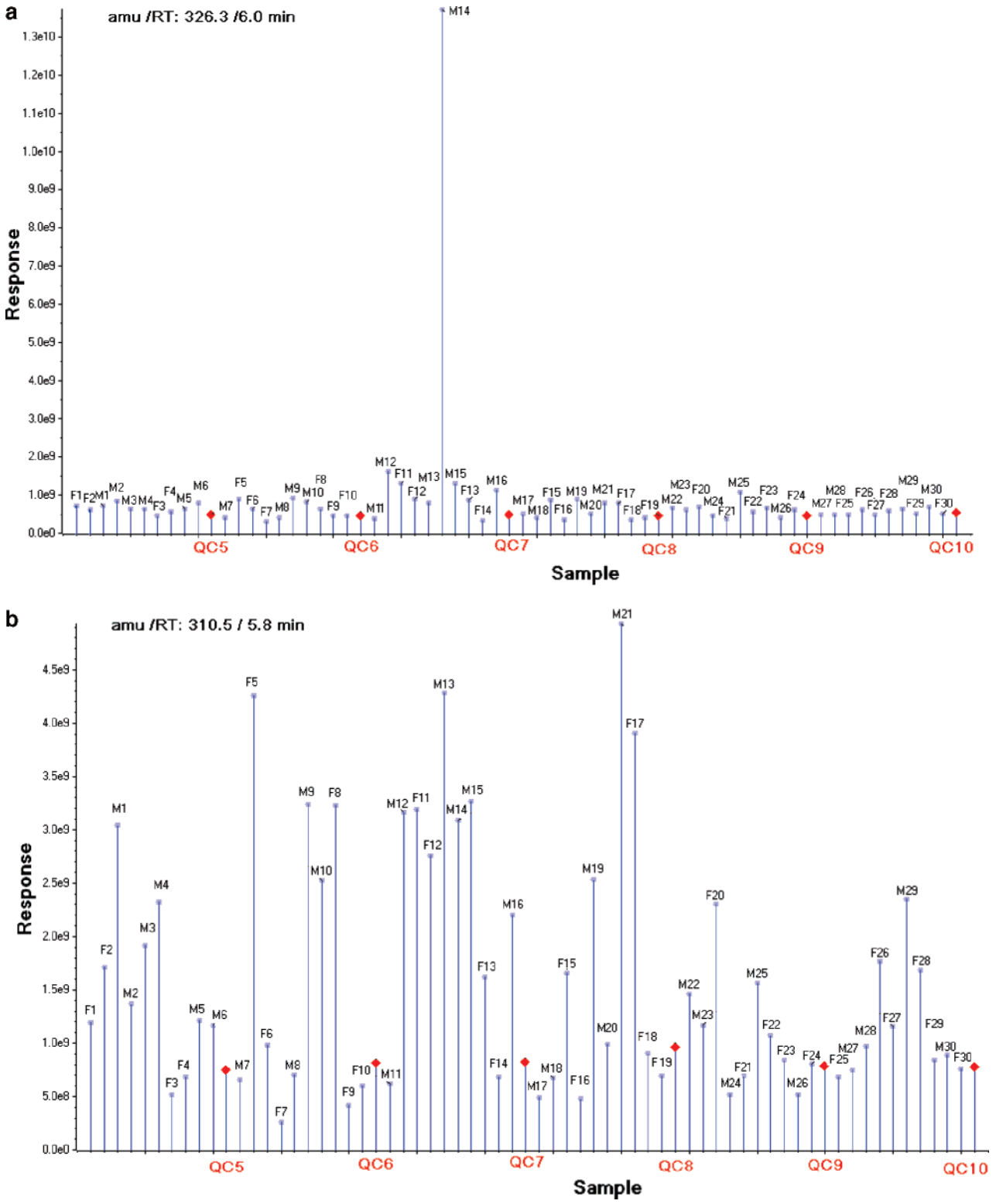
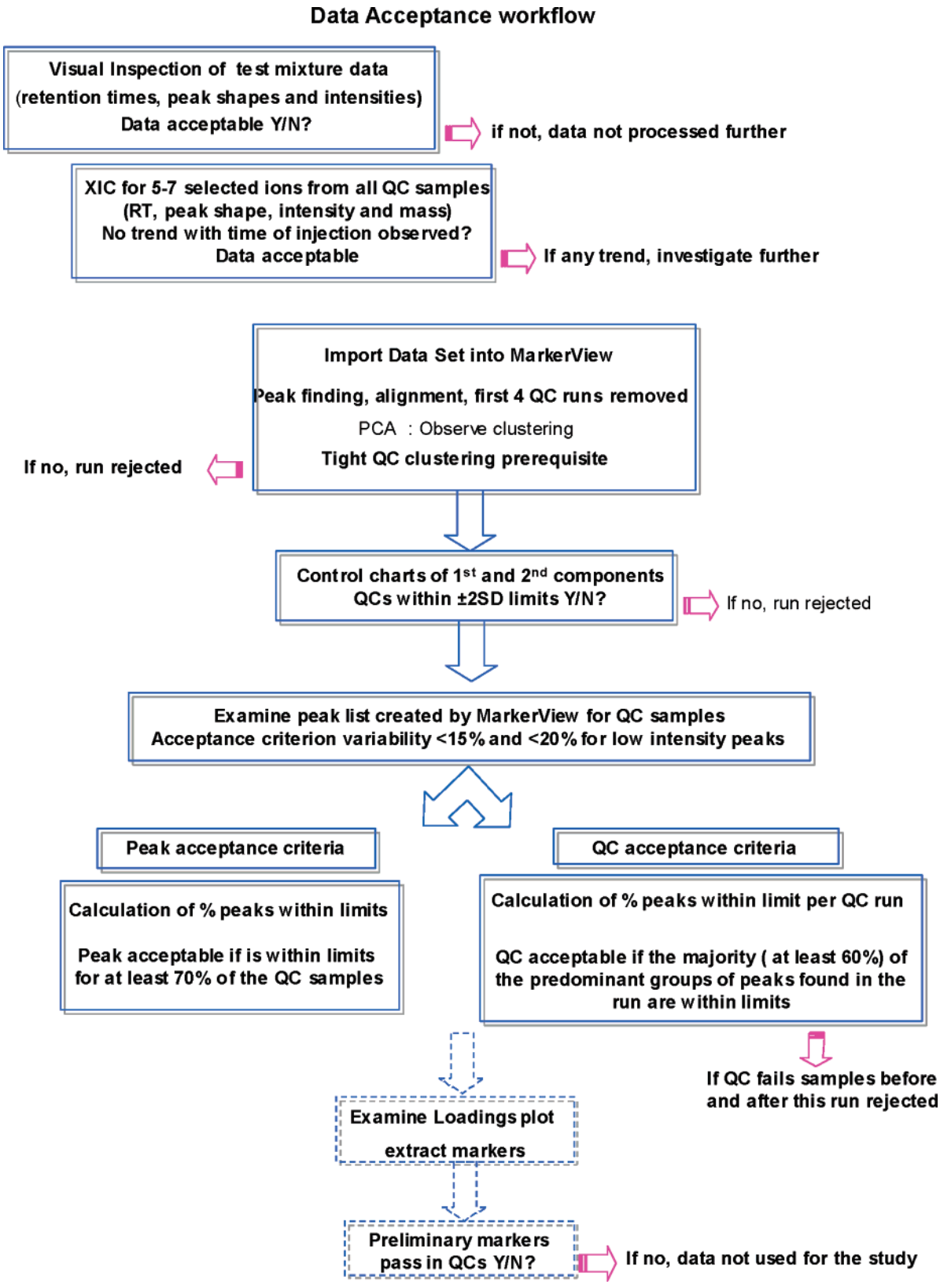


Figure 6. Intensities in all samples including QCs of (a) ion *m/z* 326.3 eluted at 6 min greatly contributing to the differentiation of male subject no. 14 and (b) ion *m/z* 310.5, eluted at 5.8 min contributing to the differentiation of samples M21, M13, F5, and F17.

based analysis,[16] which showed sample stability for human urine at −25 °C for 26 weeks. Interestingly, in the case of storage of urine for proteomic work, −70 °C was seen to give better results for the preservation of urinary exosomes compared to −20 °C, while freeze-thawing had relatively little effect for up to 4 cycles.[17]

Similarly, analysis of the data obtained from samples that had undergone from 1 to 9 freeze−thaw cycles showed that, by PCA, the number of freeze−thaw cycles did not affect how they clustered in the scores plot. This suggests that, for urine samples such as these, the overall stability of the samples with respect to freeze−thaw seems acceptable.

Scheme 1. Flow Chart of the Acceptance Criteria Suggested by the Validation Process



Obviously, all of the results with respect to stability described above apply to the "bulk properties" of the sample as determined by LC−MS and subsequent PCA. This represents a pragmatic tool for assessing stability and can be criticized as representing a rather blunt instrument. Indeed, there may well be individual components that are subject to degradation under the conditions described above that are not highlighted by this approach. Clearly, once a particular analyte, or set of analytes, is detected that may be a candidate(s) biomarker(s), a more targeted assessment of stability must be undertaken.

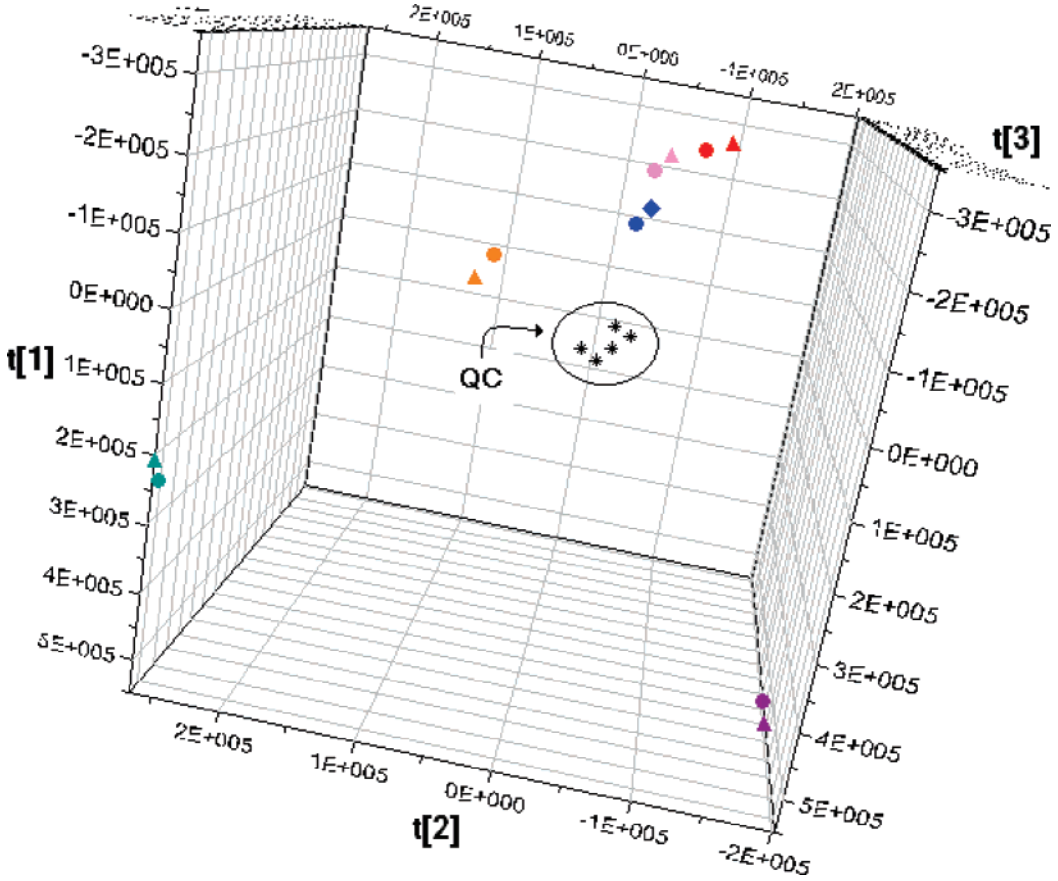


Figure 7. PCA of 6 samples analyzed after 1 month storage at either −80 °C (solid triangles) or kept at −20 °C (solid circles). The QS samples, showing the analytical variability are indicated by the "star" symbols.

4. Conclusions

The increasing emphasis on data quality in systems biology research presents great practical difficulties because of the requirement for the simultaneous measurement of multiple variables in complex samples where the identity of many of the components is unknown. Via a combination of test mixtures of known compounds and a pooled "QC" sample (constructed from aliquots of the entire of the sample set), it was possible to demonstrate that the LC−MS system had stabilized and was suitable for sample analysis. From these results, it is clear that a number of injections of urine are required to equilibrate the system prior to the commencement of the analytical run. Further, using the data from the QC samples enabled the factors that contributed to non-reproducibility between runs to be identified and for control measures to be put in place. Variability in both mass and retention time were not observed to be high (no more than 0.2 amu and 0.08 min, respectively, for randomly selected peaks), but signal intensity had a major effect on reproducibility. In particular, the variability of lower intensity peaks was significantly higher than those of higher intensity.

Samples of human urine were found to be stable in the autosampler during analysis, for up to 1 month at −20 °C, and for up to 9 freeze−thaw cycles. On the basis of these results, it seems to be possible to collect, store, and perform within-day metabonomic analysis on human urine samples with confidence that it will be possible to reliably and reproducibly detect potential biomarkers.

As LC−MS-based methods for global metabolite analysis attain a degree of maturity, it is increasingly necessary to be able to demonstrate the validity of conclusions drawn from the data. While there are various initiatives associated with the development of reporting standards,[18−19] there has been relatively little discussion in the literature on acceptable standards for method validation (for example, see refs 20 and 21). However, given the critical importance of this topic in the development of LC−MS-base metabonomics methods, this area is destined to become an increasingly important theme for such work.

Acknowledgment. This work was carried out in the context of a EU committee Transfer of Knowledge Industry-Academia partnership grant (TOK-IAP 29640). The authors thank Richard Payne for his support in technical and software matters as well as Mark Earll for the helpful discussions and suggestions in statistical issues.

References

(1) Nicholson, J. K; Lindon, J. C.; Holmes, E. *Xenobiotica* **1999**, *29* 1181−1189.
(2) Nicholson, J. K.; Connelly, J.; Lindon, J. C.; Holmes, E. *Nat. Rev. Drug Discovery* **2002**, *1*, 253−258.
(3) Nicholson, J. K.; Wilson, I. D. *Progr. NMR Spectros.* **1989**, *21*, 449−501.
(4) Lindon, J. C.; Holmes, E.; Nicholson, J. K. *Progr. NMR Spectros.* **2004**, *45*, 109−143.
(5) *Metabonomics in Toxicity Assessment*; Robertson, D., Lindon, J., Nicholson, J. K, Holmes. E., Eds.; CRC Press: Boca Raton, FL, 2005.

(6) Plumb, R. S.; Stumpf, C. L.; Gorenstein, M. V.; Castro-Perez, J. M.; Dear, G. J.; Anthony, M.; Sweatman, B. C.; Connor, S. C.; Haselden, J. N. *Rapid Commun. Mass Spectrom.* **2002**, *16*, 1991−1996.
(7) Wilson, I. D.; Plumb, R.; Granger, J.; Major, H.; Williams, R.; Lenz, E. M. *J. Chromatogr., B* **2005**, *81*, 67−76.
(8) Sabatine, M. S.; Liu, E.; Morrow, D. A.; Heller, E.; Carroll, R.; Wiegand, R.; Berriz, G. F.; Roth, F. P.; Gerszten, R. E. *Circulation* **2005**, *112*, 3868−3875.
(9) O'Hagan, S.; Dunn, W. B.; Brown, M.; Knowles, J. D.; Kell, D. B. *Anal. Chem.* **2005**, *77*, 290−303.
(10) Welthagen, C. W.; Shellie, R. A.; Spranger, J.; Ristow, M.; Zimmermann, R.; Fiehn, O. *Metabolomics* **2005**, *1*, 65−73.
(11) Lenz, E. M.; Wilson, I. D. *J. Proteome Res.* **2007**, *6*, 443−458.
(12) Dumas, M. -E.; Maibaum, E. C.; Teague, C.; Ueshima, H.; Zhou, B.; Lindon, J. C.; Nicholson, J. K.; Stamler, J.; Elliot, P.; Chan, Q.; Holmes, E. *Anal. Chem.* **2006**, *78*, 2199−2208.
(13) Sangster, T.; Major, H.; Plumb, R.; A. J. Wilson, A. J.; Wilson, I. D. *Analyst* **2006**, *131*, 1075−1078.
(14) Sangster, T.; Wingate, J.; Burton, L.; Teichert, F.; Wilson, I. D. *Rapid. Commun. Mass Spectrom.* accepted for publication.
(15) FDA Guidance for Industry, Bioanalytical method Validation, Food and Drug Administration, centre for Drug valuation and Research (CDER), May 2001. a guidance.
(16) Lauridsen, M.; Hansen, S. H.; Jaroszewski, J. W.; Cornett, C. *Anal. Chem.* **2007**, *79*, 1181−1186.
(17) Pisitkun, T.; Johnstone, R.; Knepper, M. A. *Mol. Cell. Proteomics* **2006**, *5*, 1760−1771.
(18) Lindon, J. C.; Nicholson, J. K.; Holmes, E.; Kuen, H. C.; Craig, A.; Pearce, J. T. M.; Bruce, S. J.; Hardy, N.; Sansone, S-A.; Antti, A.; Jonsson, P.; Daykin, C.; Navarange, M.; Beger, R. D.; Verheij, E. R.; Amberg, A.; Baunsgaard, D.; Cantor, G. H., Lehman-McKeeman, L.; Earll, M.; Wold, S.; Johansson, E.; Haselden, J. N.; Kramer, K.; Thomas, C.; Lindberg, J.; Schuppe-Koisetinen, I.; Wilson, I. D.; Reily, M. D.; Robertson, D. G.; Senn, H.; Krotzky, A.; Kochhar, S.; Powell, J.; van der Ouderaa, F.; Plumb, R.; Schaefer H. L; Spraul, M. *Nat. Biotechnol.* **2005**, *23*, 833−838.
(19) Castle, A. L.; Fiehn, O.; Kaddurah-Daouk, R.; Lindon, J. C. *Briefings Bioinf.* **2006**, *7*, 159−165.
(20) Wagner, S.; Scholz, S.; Sieber, M.; Kellert, M.; Voelkel, W. *Anal. Chem.*, **2007**, *79*, 2918−2926.
(21) Bottcher, C.; v. Roepenack-Lahaye, E.; Willscher, E.; Schell, D.; Clemens, S. *Anal. Chem.* **2007**, *79*, 1507−1513.

PR070183P