Anal. Chem. **2006,** *78,* 2199−2208

Assessment of Analytical Reproducibility of ^{1}H NMR Spectroscopy Based Metabonomics for Large-Scale Epidemiological Research: the INTERMAP Study

Marc-Emmanuel Dumas,*,†,‡ Elaine C. Maibaum,†,‡ Claire Teague,† Hirotsugu Ueshima,§ Beifan Zhou,|| John C. Lindon,† Jeremy K. Nicholson,† Jeremiah Stamler,⊥ Paul Elliott,# Queenie Chan,# and Elaine Holmes*,†

Biological Chemistry, Biomedical Sciences Division, Imperial College London, Sir Alexander Fleming Building, South Kensington, London UK, Department of Health Science, Shiga University of Medical Science, Otsu, Japan, Department of Epidemiology, Fu Wai Hospital and Cardiovascular Institute, Chinese Academy of Medical Sciences, Beijing, People's Republic of China, Department of Preventive Medicine, Feinberg School of Medicine, Northwestern University, Chicago, IL 60611, and Department of Epidemiology and Public Health, Imperial College London, St Mary's campus, London UK

Large-scale population phenotyping for molecular epidemiological studies is subject to all the usual criteria of analytical chemistry. As part of a major phenotyping investigation we have used high-resolution ^{1}H NMR spectroscopy to characterize 24-h urine specimens obtained from population samples in Aito Town, Japan (n = 259), Chicago, IL (n = 315), and Guangxi, China (n = 278). We have investigated analytical reproducibility, urine specimen storage procedures, interinstrument variability, and split specimen detection. Our data show that the multivariate analytical reproducibility of the NMR screening platform was >98% and that most classification errors were due to urine specimen handling inhomogeneity. Differences in metabolite profiles were then assessed for Aito Town, Chicago, and Guangxi population samples; novel combinations of biomarkers were detected that separated the population samples. These cross-population differences in urinary metabolites could be related to genetic, dietary, and gut microbial factors.

With the current emphasis on research in "systems biology", there is a need to develop appropriate tests of data quality and reproducibility, e.g., of gene expression arrays, 2-D gels for protein analysis, or metabolite profiles. This is inherently difficult owing to the simultaneous measurement of multiple variables where identity of many components may be unknown. Thus, in ^{1}H NMR-based metabonomics, an integrative approach in systems biology, NMR spectroscopy of biofluids, generates many hundreds of signals from low molecular weight metabolites without selection of specific analytes.[1−3] The data are then reduced and interpreted by use of multivariate statistics. Metabonomics has considerable potential as a means of rapidly providing metabolic fingerprints of individuals. These fingerprints alter in a reproducible and characteristic manner in response to various physiological or pathological challenges and can reflect multiple influences of both genetic (such as inborn errors of metabolism)[4−6] and environmental (diet or alcohol use)[7−9] factors. Such holistic metabotyping may prove valuable in the future for assessing the presence of disease (or risk factors for disease), e.g., as already exemplified for coronary heart disease.[10] The technology has been shown to be stable in interlaboratory animal toxicology studies conducted with strict environmental control on genetically similar animals,[11] but has not yet been assessed in the more challenging role of detecting diet- or disease-related metabolic signatures in human populations.

The International Study of Macro/micronutrients and Blood Pressure (INTERMAP) was launched in 1996 to investigate the

* To whom correspondence should be addressed. E-mail: elaine.holmes@imperial.ac.uk. Tel.: +44(0)20 7594 3220. Fax: +44(0)20 7594 3226. E-mail: m.dumas@imperial.ac.uk. Tel.: +44(0)20 7594 1698. Fax: +44(0)20 7594 3226.

† Biological Chemistry, Imperial College London.

‡ Contributed equally to this work.

§ Shiga University of Medical Science.

|| Fu Wai Hospital and Cardiovascular Institute.

⊥ Northwestern University.

Department of Epidemiology and Public Health, Imperial College London.

(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*, 153−161.
(3) Nicholson, J. K.; Wilson, I. D. *Nat. Rev. Drug Discovery* **2003**, *2*, 668−676.
(4) Griffin, J. L.; Williams, H. J.; Sang, E.; Clarke, K.; Rae, C.; Nicholson, J. K. *Anal. Biochem.* **2001**, *293*, 16−21.
(5) Holmes, E.; Caddick, S.; Lindon, J. C.; Wilson, I. D.; Kryvawych, S.; Nicholson, J. K. *Biochem. Pharmacol.* **1995**, *49*, 1349−1359.
(6) Holmes, E.; Foxall, P. J.; Spraul, M.; Farrant, R. D.; Nicholson, J. K.; Lindon, J. C. *J. Pharm. Biomed. Anal.* **1997**, *15*, 1647−1659.
(7) Solanky, K. S.; Bailey, N. J.; Beckwith-Hall, B. M.; Davis, A.; Bingham, S.; Holmes, E.; Nicholson, J. K.; Cassidy, A. *Anal. Biochem.* **2003**, *323*, 197−204.
(8) Gavaghan, C. L.; Wilson, I. D.; Nicholson, J. K. *FEBS Lett.* **2002**, *530*, 191−196.
(9) Bollard, M. E.; Holmes, E.; Lindon, J. C.; Mitchell, S. C.; Branstetter, D.; Zhang, W.; Nicholson, J. K. *Anal. Biochem.* **2001**, *295*, 194−202.
(10) Brindle, J. T.; Antti, H.; Holmes, E.; Tranter, G.; Nicholson, J. K.; Bethell, H. W.; Clarke, S.; Schofield, P. M.; McKilligin, E.; Mosedale, D. E.; Grainger, D. J. *Nat. Med.* **2002**, *8*, 1439−1444.
(11) Keun, H. C.; Ebbels, T. M.; Antti, H.; Bollard, M. E.; Beckonert, O.; Schlotterbeck, G.; Senn, H.; Niederhauser, U.; Holmes, E.; Lindon, J. C.; Nicholson, J. K. *Chem. Res. Toxicol.* **2002**, *15*, 1380−1386.

10.1021/ac0517085 CCC: $33.50
Published on Web 03/08/2006

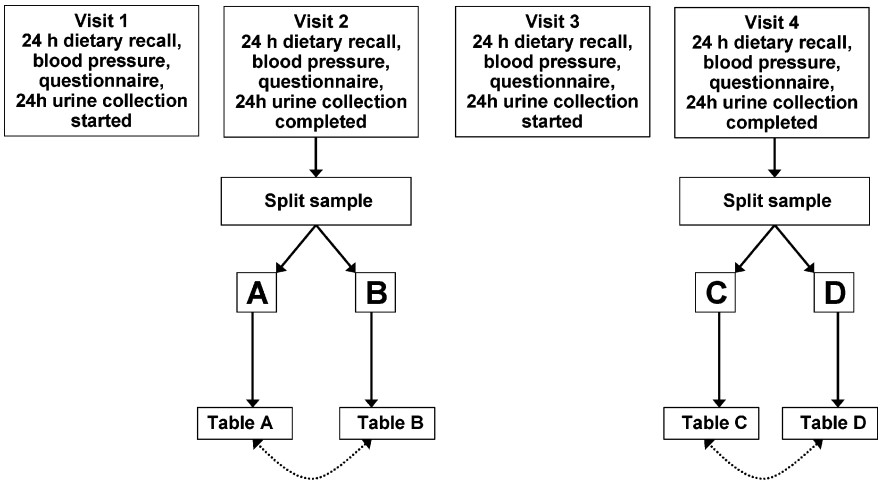

Figure 1. Summary of INTERMAP study design serving as basis for ^{1}H NMR metabonomic urinalysis protocol. Collection of 24-h urine specimens during visits 2 and 4, respectively, 1 day after visit 1 and 3. 8.3% of the specimens split at source and given different identification labels. All the specimens were analyzed by ^{1}H NMR.

relationship of multiple dietary variables to blood pressure.[12,13] The 4680 adult men and women, aged 40−59, who participated in the study were selected from 17 population samples in Japan, the People's Republic of China, the United Kingdom, and the United States. In-depth characterization of their nutrient intakes was assessed by four 24-h dietary recalls and two timed 24-h urinary collections over a 3−6-week period. Pregnant women were not included since both diet and blood pressure are affected by pregnancy.

Here we present data in a first-phase metabonomic analysis of urine specimens obtained from male and female participants from three population samples; Aito Town (Japan), Guangxi (China), and Chicago (USA).[13] In this first study, the repeatability and accuracy of the metabonomics methodology are explored, as well as variation in urinary metabolite profiles across populations. A series of benchmark tests are used to address the following: (i) the robustness of the method based on repeated analysis of aliquots from a large pool of quality control samples prepared to assess NMR stability over the 7-month data acquisition period; (ii) interinstrument variation in analytical reproducibility; (iii) ability of NMR-based metabonomics to identify samples split at source within each population sample for double-blind quality control assessment, using a hierarchical clustering approach; (iv) analytical variation across split samples by computing canonical *Rv* coefficients; and (v) intersample differences in urinary metabolite profiles.

EXPERIMENTAL SECTION

Collection and Preparation of Quality Control (QC) Samples. The 24-h urine specimens (∼1 L) were obtained from Three volunteers; a Caucasian woman, 29 years old (QC1), a Caucasian man, 32 years old (QC2), and a Chinese man, 45 years old (QC3), all U.K. residents, at Imperial College. Boric acid was added to the collection vessels as a preservative, the specimens were mixed thoroughly and stored in aliquots of ∼5 mL at −40 °C.

Population Samples and Study Design. Methods of data collection for both nondietary and dietary data in the INTERMAP study have been described.[13,14] Briefly, each individual attended a specified clinic center on four occasions; the first two and last two visits were consecutive days and there was a period of 3−6 weeks between pairs of visits (Figure 1). Data collection included blood pressure on eight occasions (2 per visit), four 24-h dietary recalls including dietary supplement consumption, two 7-day records of daily alcohol intake, medical diagnoses, and extensive data on demographic traits, smoking, and other variables from a questionnaire. At the first and third visits, a 24-h urinary collection was initiated and completed the following day in the clinic, according to a standardized protocol. Boric acid preservative was included in the urine collection bottles given to each participant. Completed urine collections were mixed thoroughly and total volumes recorded. Urine collections of less than 250 mL or where the participant indicated incompleteness of the collection on the questionnaire were rejected and a further 24-h collection was requested (or a substitute participant of the same age (5 year group) and sex was recruited). Aliquots (5 mL) were taken, frozen at −20 °C within 24 h, and shipped frozen to the INTERMAP Central Laboratory in Leuven, Belgium, for multiple standardized analyses (Na, K, Ca, Mg, creatinine). The samples were stored at −30 °C, before being shipped on dry ice to the Metabonomic Laboratory at Imperial College. This publication relates to 852 first and 852 repeat visits (426 women, 426 men), urinary collections from centers in Aito Town (Japan), Guangxi (People's Republic of China), and Chicago (USA) (Table 1).

Analysis of Duplicate Aliquots for Assessment of Method Performance. To assess precision of NMR metabonomic analysis, 134 randomly selected urine specimens (8.3%) were split at the clinical center and sent to the Metabonomics Laboratory with different identification numbers. Therefore, each INTERMAP individual from the split sample data set was characterized by four NMR spectra: a split pair for the first visit and a split pair for the

(12) Beevers, D. G.; Stamler, J. *J. Hum. Hypertens.* **2003**, *17*, 589−590.
(13) Stamler, J.; Elliott, P.; Dennis, B.; Dyer, A. R.; Kesteloot, H.; Liu, K.; Ueshima, H.; Zhou, B. F. *J. Hum. Hypertens.* **2003**, *17*, 591−608.
(14) Dennis, B.; Stamler, J.; Buzzard, M.; Conway, R.; Elliott, P.; Moag-Stahlberg, A.; Okayama, A.; Okuda, N.; Robertson, C.; Robinson, F.; Schakel, S.; Stevens, M.; Van Heel, N.; Zhao, L.; Zhou, B. F. *J. Hum. Hypertens.* **2003**, *17*, 609−622.

Table 1. Identification of Split Specimens by Hierarchical Clustering

population sample		visit 2			visit 4			total		
no.	name	no. of urine specimens	no. of split pairs	split pairs identified	no. of urine specimens	no. of split pairs	split pairs identified	no. of urine specimens	no. of split pairs	split pairs identified (%)
14	Aito Town	259	24	24	259	23	23	565	47	97.9
41	Chicago	315	24	23	315	24	23	678	46	95.8
21	Guangxi	278	23	17	278	24	17	603	34	70.8
all three samples		852	71	64	852	71	63	1846	127	88.2

repeat visit. Data tables A, B, C and D (Figure 1) were created using the spectra from the specimens obtained on the first visit (A), repeat visit (C), or aliquots from both of these visits (B and D, respectively) and used to populate the four spectral data tables. A second Chinese sample, Beijing, was subsequently investigated with the same protocol to evaluate results from the Guangxi analysis.

Preparation of Urine Specimens for ^{1}H NMR Spectroscopy. Urine specimens were thawed completely before mixing 500 μL of urine with 250 μL of phosphate buffer for stabilization of the urinary pH 7.4 ($\pm$0.5) and 75 μL of the sodium 3-trimethylsilyl-(2,2,3,3-$^2H_4$)-1-propionate (TSP) in D_2O (final concentration 0.1 mg/mL) solution; TSP acted as internal chemical shift reference (δ0.0) D_2O provided NMR lock signal for the NMR spectrometer. This preparation was placed into a 96-well plate for analysis. The remaining urine specimen was refrozen. The 96-well plate was then left to stand for 10 min before centrifuging at 4000 rpm for 10 min to remove any precipitate.

Monitoring of Quality Control Samples. To ascertain that contamination from previous specimens in the NMR flow injection system was not occurring, blanks were included at the beginning and the end of each well plate. The spectra from the blanks showed no significant carryover of metabolites, i.e., no cross contamination. To assess analytical stability over time, continuous monitoring of the study was achieved using aliquots of the three large 24-h urine collections obtained from three healthy volunteers (see above). Two aliquots of each quality control specimen were interspersed among INTERMAP specimens on every well plate in random order, i.e., six QC samples per well plate. Altogether, 24 well plates were measured over 7 months, i.e., 144 QC specimens altogether (48 aliquots per QC specimen). Position and dispersion parameters such as mean (μ) and standard deviation (SD, σ) were computed for each variable (i.e., NMR spectral region). Variation coefficients were then computed for each spectral region by dividing σ by μ.

Boric Acid Preservative. Boric acid, used as a preservative in the urine specimens, is known to bind covalently to vicinal diols and some amino acids. Such reactions are slow but can result in the formation of novel adducts in urine. We have investigated the effect of borate on urine for NMR studies and have found that, although it may confound quantification of compounds convolved with citrate, it does not impair classification of the urine specimens by chemometric techniques.[15]

^{1}H NMR Spectroscopic Analysis of Urine Specimens. Spectra were obtained using a Bruker (Bruker Biospin, Rheinstetten, Germany) DRX600 spectrometer operating at 600 MHz in flow injection mode. Specimens were automatically delivered to the spectrometer with a Gilson robot incorporated into the BEST (Bruker Efficient Sample Transfer) system. One-dimensional ^{1}H NMR spectra of urine were acquired using a standard 1-D pulse sequence (recycle delay$-90°-t_1-90°-t_m-90°-$acquisition) with water presaturation during both the recycle delay (2 s) and the mixing time (t_m, 150 ms). The 90° pulse length was adjusted to ~10 μs and t_1 was set to 3 μs, providing an acquisition time of 2.73 s and a total pulse recycle time of 4.73 s. Sixty-four free induction decays (FIDs) were collected into 32K data points using a spectral width of 20 ppm. To ascertain effect of differences in interinstrument acquisition parameters and instrumental variability, a randomly selected subset of urine specimens was also measured on a separate spectrometer operating under similar conditions but using a spectral width of 12 ppm and a recycle delay of 1.5 s giving a total acquisition time of 3.73 s. For both sets of data, FIDs were multiplied by an exponential weighting function corresponding to a line broadening of 0.3 Hz and data were zero-filled by a factor of 2 prior to Fourier transformation.

^{1}H NMR Spectral Processing. Baseline correction and phasing of the spectra was achieved with in-house software (T. Ebbels and H. Keun). Each spectrum was reduced to a series of integrated regions of equal width (0.04 ppm, standard bucket width) corresponding to the regions of δ0.16$-$9.76 inclusive. To remove effects of variation in water resonance suppression, saturation transfer effects to the urea signal, and borate$-$citrate interaction shifts, spectral regions were removed between δ4.50$-$5.98 and the citrate peak regions (δ2.69$-$2. 73 and δ2.53$-$2.58). Each spectrum was then normalized to unit area, to remove the effect of differences in urinary concentrations between specimens. The analysis was repeated using smaller integrated regions of 0.01 ppm (improved resolution of model coefficients).

Statistical Analysis of NMR Spectroscopic Data. Statistical analysis of urinary ^{1}H NMR spectral data was performed with S-PLUS 6.1 (www.insightful.com), according to several data analysis strategies, as described below.

Identification of Split Specimen by Hierarchical Clustering Trees (HCTs).[16] Hierarchical clustering has been used to discover groupings among a data matrix. HCTs are based on a measure of similarity or dissimilarity between observations, computed from the data matrix. Dissimilarity coefficients are symmetric, i.e., $d(A,B) = d(B,A)$ and nonnegative i.e., $d(A,A) = 0$ and can be displayed as dendrograms. The complete-linkage method used here employs the maximum Euclidean distance between two

(15) Teague, C. Ph.D. thesis, Imperial College London, University of London, London, 2004.

(16) Venables, W. N.; Ripley, B. D. *Modern applied statistics with S-PLUS*, 4th ed.; Springer-Verlag: New York, 2002.

specimens as the dissimilarity measure. To magnify the low distance values (denoting a high similarity or nearness of the specimens), the logarithm of the distance matrix was used for building the clustering tree model.

Measure of Table Similarity Using the *Rv* Coefficient.[17] Each data set corresponding to the two urine specimens obtained for each participant with the corresponding split specimens was used to populate four different tables (first and repeat visit tables: $\mathbf{X}_A$ and $\mathbf{X}_C$, split sample tables $\mathbf{X}_B$ and $\mathbf{X}_D$) each with different variance (Figure 1). The *Rv* coefficient expresses the homothetic (similar geometrical) relationship between the pairs of tables (A and B) and (C and D), since samples in table B were duplicate aliquots of samples in table A, etc. The *Rv* coefficient for a (column-centered) data matrix (with p variables/columns) **X**, and the regression of these columns on a k-variable subset, is defined by Robert and Escoufier:

$$Rv(\mathbf{X}_A, \mathbf{X}_B) = \mathrm{tr}(\mathbf{X}_A\mathbf{X}_A'\cdot\mathbf{X}_B\mathbf{X}_B')/[(\mathrm{tr}(\mathbf{X}_A\mathbf{X}_A')^2\cdot\mathrm{tr}(\mathbf{X}_B\mathbf{X}_B')^2]^{1/2}$$

where $\mathbf{X}_A$ and $\mathbf{X}_B$ are mean-centered matrices, tr is the trace operator (sum of diagonal elements), and $'$ is the transposition operator. $Rv(\mathbf{X}_A,\mathbf{X}_B) = 0$ if the variables from $\mathbf{X}_A$ and $\mathbf{X}_B$ tables are not correlated, and $Rv(\mathbf{X}_A,\mathbf{X}_B) = 1$ if the $\mathbf{X}_A$ and $\mathbf{X}_B$ tables are identical.

Pattern Recognition across Samples by Linear Discriminant Analysis (LDA).[16] One of each pair of split specimens was removed from the data set such that each of the 852 participants was represented by a single specimen per clinic attendance prior to performing LDA on the principal component matrix for the spectral data. LDA is a supervised PR method, using the **Y** response vector (in this case a class assigned on the basis of center), to structure the information within **X** data (NMR or metabolic descriptors). LDA maximizes the distance between the centroids of different population samples. LDA seeks an optimal linear combination **Xp** of the input variables leading to separation of the class means relative to the within-class variance, i.e., maximizing $\mathbf{p}^T\mathbf{B}\mathbf{p}/\mathbf{p}^T\mathbf{W}\mathbf{p}$ ratio, in which **B** is the between-group covariance matrix and **W** the within-class covariance matrix (centered on the class mean). The linear combination is solved by taking **p** to be the eigenvectors of **B**. LDA has to be performed on full rank matrices. With NMR spectral data, each metabolite may generate more than one signal; thus, there is redundancy of information within the data matrix. This is commonly solved by precompressing the data, for example by principal components analysis (PCA).[18]

Data Compression by Principal Components Analysis.[18] PCA is a pattern recognition method for multivariate data, independent of any knowledge of class membership (unsupervised). It performs a bilinear decomposition of the NMR data **X** represented in *K*-dimensional space (*K* equal to the number of variables), known as singular value decomposition. It reduces **X** to a set of orthogonal principal components (or latent variables) describing the main directions of variation within the data. Each direction (principal component) is attached to an "eigenvalue" λ, summarizing the amount of variance explained, and a dual set of values (bilinear): (i) the "scores" **t** describing the statistical objects (participants), and (ii) the "loadings" **p** or "eigenvectors" highlighting the influence of input variables (NMR descriptors) on **t**.

Full Cross-Validation of the Models.[16] Model parameters, e.g., number of principal components, were optimized by a full 3-fold cross-validation strategy. After random selection, approximately 2/3 of observations were used for model calibration (calibration set), whereas the remaining 1/3 of observations was used to test the model (test set).

RESULTS AND DISCUSSION

Intrinsic Reproducibility of ^{1}H NMR Spectroscopy of Urine. As noted in the Experimental Section, to evaluate long-term reproducibility of the ^{1}H NMR-based metabonomic technology in large-scale human studies, three pairs of QC specimens were analyzed in each 96-well plate flow injection NMR (48 aliquots for each of three QC individuals, altogether 144 QC specimens). Using PLS regression analysis, no significant correlation between the variances in the QC spectral profiles with time was found, indicating that no biochemical degradation had occurred over the 7-month study period and that the analysis was sufficiently robust (a Q^2 value of -0.057 was obtained for a one-component model and -0.163 for a two-component model). This is concordant with earlier studies which concluded that the analytical approach and specimen storage are inherently stable.[11,19]

The three QC specimens generated spectra that were intrinsically similar to each other in terms of the urinary metabolites present in high concentrations, with each of the three specimens dominated by creatinine and trimethylamine-*N*-oxide (TMAO). Regions corresponding to the signals from dominant species in urine provide high average values and also high standard deviations. These results are in agreement with an earlier study on variability in rat urine.[20]

The mean, standard deviation, and coefficients of variation of each NMR variable from the QC specimens provide initial estimates of NMR variable stability; these are depicted for each NMR variable in a spectral format in Figure 2 (a spectral integral corresponding to a discrete region of 0.01 ppm). The average spectra derived from the mean values of the measured aliquots for each volunteer (Figure 2A) give a good estimate of the original spectrum of the original QC specimen used for this purpose, whereas the standard deviation spectra (Figure 2B) give an estimate of the instability of the QC samples over the 7-month measurement period.

Calculation of the Stability of the Spectral Profile. The mean and standard deviation were computed over the full spectral window (δ0.16−9.76). Mean and standard deviation plots for most biological variables (in the range δ1.0−8.0) display a similar pattern as exemplified in Figure 2A and B. To estimate better the variation across the spectrum, the coefficient of variation is preferable, as it does not bias toward spectral regions containing high-concentration metabolites (Figure 2C). The spectral regions usually covered by urinary signals were relatively flat compared with the surrounding regions at the edges of the spectrum that

(17) Robert, P.; Escoufier, Y. *Appl. Stat.* **1976**, *25*, 257−265.
(18) Eriksson, L.; Johansson, E.; Kettanah-Wold, N.; Wold, S. *Multi and Megavariate Data Analysis. Principles and Applications*; Umetrics AB: Malmo, Sweden, 2001.
(19) Holmes, E.; Foxall, P. J.; Nicholson, J. K. *J. Pharm. Biomed. Anal.* **1990**, *8*, 955−958.
(20) Ebbels, T. M.; Holmes, E.; Lindon, J. C.; Nicholson, J. K. *J. Pharm. Biomed. Anal.* **2004**, *36*, 823−833.

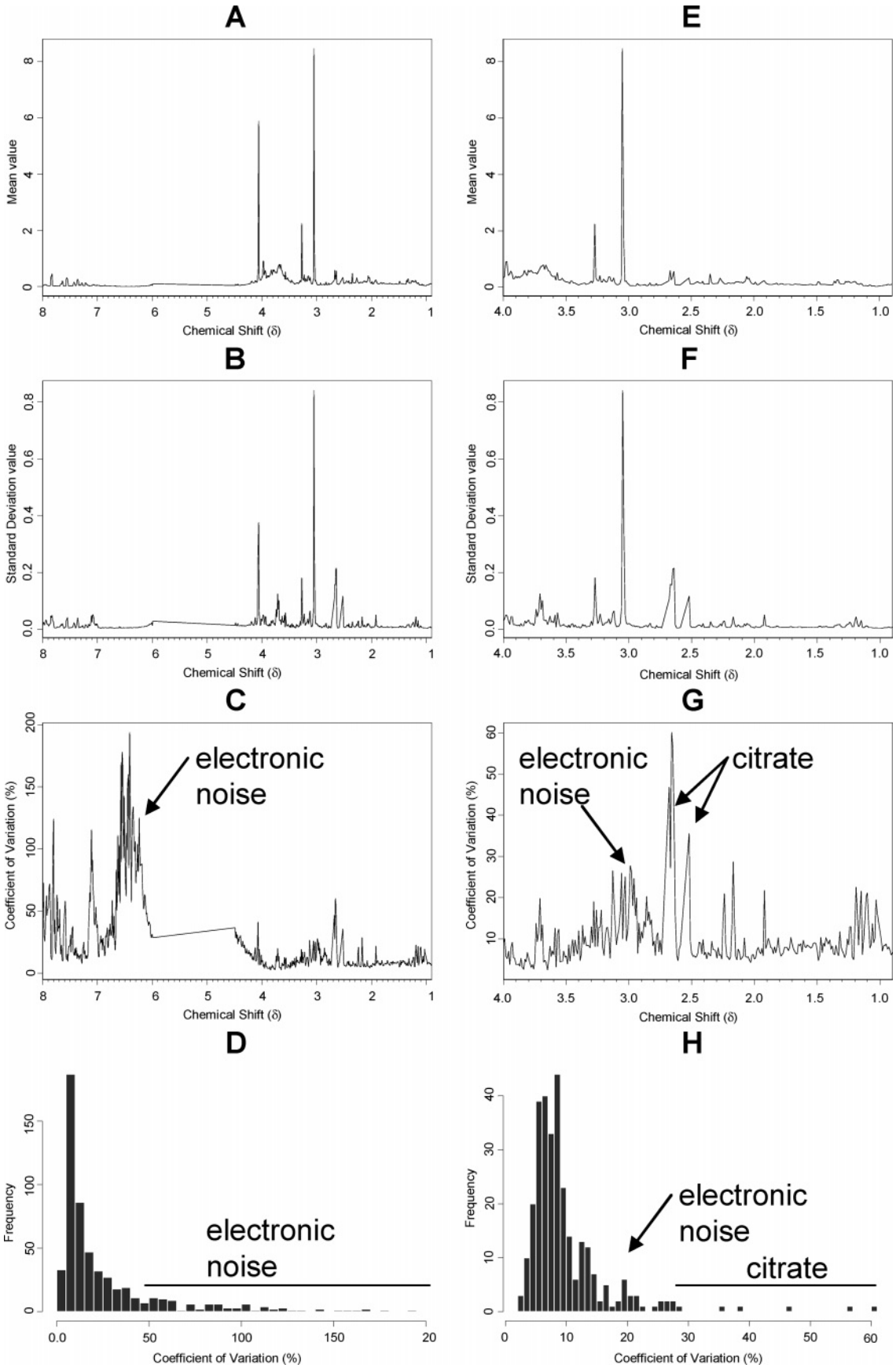

Figure 2. Position and dispersion parameters and coefficients of variation for the quality control urine specimens. (A) Mean spectrum in the δ1.0−8.00 range, (B) standard deviation spectrum in the δ1.0−δ8.00 range, (C) coefficient of variation spectrum in the δ1.0−δ8.00 range, (D) distribution of the coefficients of variation in the δ1.0−δ8.00 range, (E) mean spectrum in the δ1−δ4 range, (F) standard deviation spectrum in the δ1−δ4 range, (G) coefficient of variation spectrum in the δ1−δ4 range, and (H) distribution of the coefficients of variation in the δ1−δ4 range.

did not contain metabolite signals. For ^{1}H NMR spectra of urine, the regions above δ6.0 and below δ6.6 correspond mostly to electronic noise; SD and mean tend to zero. Small variations around zero in mean and SD can lead to chaotic behavior for a coefficient of variation (CV) computed by dividing the SD by the mean. The distribution of the CV across multiple peaks in the spectrum is dominated by the electronic noise variables and not by peaks that correspond to true metabolite signals (Figure 2D). Such nonreproducible regions can be discarded from further statistical calculations, since they represent only a very small proportion of the NMR intensity and can introduce "noise" into the analysis. Reducing the spectral range to δ1−4 did not significantly affect the mean (Figure 2E) and standard deviation (Figure 2F) spectra but provided an improved synopsis of the reproducibility across the spectrum based on the coefficients of variation (Figure 2G,H). Here the scale for coefficient of variation has been expanded for the purpose of illustration.

Other regions of relative instability in the spectra were associated with variation in the efficiency of water suppression around δ6 and δ4, commonly encountered due to factors such as minor fluctuations in laboratory temperature and differences in ionic strength between samples and around δ2.5 and δ2.7 corresponding to variation in the intensity/chemical shift of citrate, which resulted in a CV above 30%. Slight differences in the spectral regions containing citrate have been well documented.[21,22] Although the urine specimens had been buffered, the intrinsic buffering capacity of urine is strong and precipitation of certain components will occur over time, varying with temperature, and influencing the exact pH of the sample. This accounts for slight shifts in metabolites with a pK_a close to the buffered pH. Additionally, citrate was shown to be the most vulnerable metabolite to complexation with boric acid. The region δ1−4 excluding δ2.5−δ2.7 was found to produce the optimal NMR profile in terms of signal reproducibility with CV below 5% (Figure 2G,H).

Identification of Duplicate Aliquots Using Hierarchical Clustering Trees: Blind Assessment. Based on the assessment of stability of the QC specimens, HCTs were constructed using the spectral regions between δ1 and δ4 (excluding δ2.5−δ2.7) for the rest of the samples to assess ability to identify 24 pairs of split aliquots hidden within each population sample (China, Japan, USA). In this blind assessment, identification of split pairs of urine specimens provides an additional measure of reproducibility of the technology and also assesses the sensitivity of the method in being able to differentiate between specimens from one individual obtained on the same occasion and specimens obtained from the same person (hence possibly similar) at different clinic visits.

The hierarchical clustering tree approach consists of the measurement of the overall dissimilarity index between the specimens; it uses the distance metric to build the dendrograms (see Experimental Section for details) (Figure 3). Such a clustering strategy relies on the ability of clustering trees to aggregate split specimens at the bottom of the tree because of their spectral similarity. For example, the clustering tree of the American female population sample clearly shows pairs of samples aggregated with very low distance measures (Figure 3A). Split samples are correctly aggregated at the bottom of the dendrograms and have very low aggregation levels, whereas intraindividual variability generates higher aggregation levels, with a qualitative gap in the distribution of aggregation levels. A fixed length has been assigned to the leaves of the clustering tree, and this can lead to a visual artifact: for specimens aggregated at the bottom of the tree, the beginning of the leaf, corresponding to the aggregation value used in the algorithm to cluster the specimens, is positive whereas the tip of leaf apparently starts at a negative value. Two of the specimens from duplicate aliquots, marked with an asterisk in Figure 3A, have been plotted in Figure 3B together with their difference spectrum. This clearly shows their near identity and delineates them as replicate (split) specimens. Slight differences in the spectral regions containing citrate and TMAO can be noted.

This clustering approach, undertaken blind, successfully identified 98% of Japanese, 96% of American, and 71% of Chinese split samples when using spectral segments of 0.04 ppm width (Table 1). The more accurate identification rate obtained for Japanese and American population samples, using regions of 0.04 ppm in the δ0.16−δ9.76 spectral window, contrasts with the somewhat lower one obtained for the Chinese population sample. Smaller spectral segments provide an enhancement of resolution and, therefore, more accuracy in identifying pairs of split specimens. The variation between splits (same specimen, different aliquot) and first and repeat visits was most likely related to variation in the consistency of specimen handling, but may have been partly attributable to the strong consistency between first and repeat specimens from the Chinese population. To investigate the source of this relatively low detection rate for split specimens in the Guangxi population, urine specimens from a second Chinese population, Beijing, were analyzed using the same protocol. From 272 first and 272 repeat visits, 39 of the 42 pairs of splits (93%) were correctly identified; i.e., greater reliability than for Guangxi population sample, suggesting sample handling inconsistencies. The identification of split specimens and the confidence level associated with these findings was used to establish the variation arising from experimental methods and data handling.

Reasons for failure to identify pairs of split specimens (misclassification) can be divided into three categories: (A) acquired physicochemical differences between specimens, e.g., pH or osmolarity, as a result of variation in sample preparation procedures; (B) chemical or biochemical degradation; (C) instrumental variation or differences in NMR acquisition parameters.

Type A Misclassifications. Variation in physicochemical parameters produced 36% of the split specimen identification error. Interspecimen differences in pH and ionic strength lead to inconsistency of chemical shift of metabolites (e.g., citrate, taurine, alanine, and histidine); small pH difference between paired specimens was the most common cause for this error. Also, specimens were preserved using boric acid; the main effect of this in urine is to complex citrate. Differential chelation of citrate and boric acid was also noted.

Type B Misclassifications. A total of 32% of the misidentified split specimens resulted from biochemical degradation of specimens, mostly from the Guangxi population sample. Many specimens from this sample were high in ethanol content, which was

(21) Silwood, C. L.; Grootveld, M.; Lynch, E. *J. Biol. Inorg. Chem.* **2002**, *7*, 46−57.

(22) Cloarec, O.; Dumas, M. E.; Craig, A.; Barton, R. H.; Trygg, J.; Hudson, J.; Blancher, C.; Gauguier, D.; Lindon, J. C.; Holmes, E.; Nicholson, J. *Anal. Chem.* **2005**, *77*, 1282−1289.

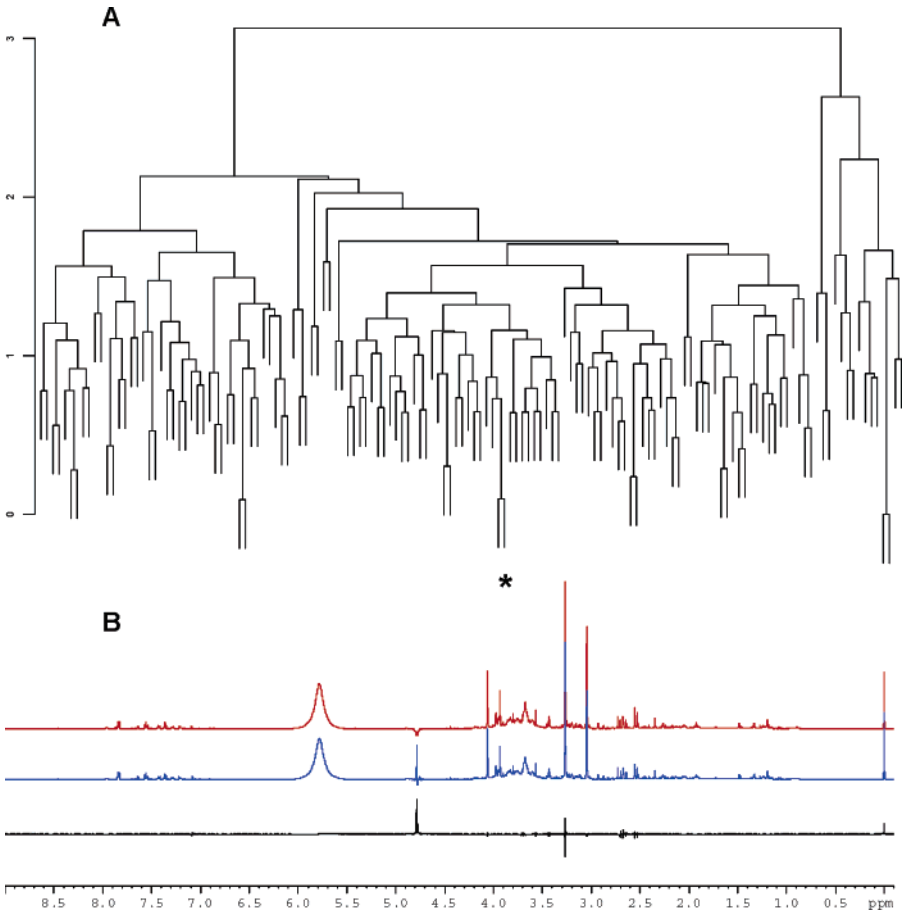

Figure 3. Identification of split urine specimens by hierarchical clustering trees of 600-MHz ^{1}H NMR spectra. (A) Single linkage dendrogram for the American female 40−49 year group for the total spectrum, (B) NMR spectra of the two replicate specimens marked with * in (B) and their difference. The vertical axis in the dendrogram corresponds to the aggregation value computed, as described in the Experimental Section, summarizing the measure of dissimilarity (≥ 0) between clusters or initial samples. To identify the split urine specimens, a fixed length has been assigned to the "leaves" (i.e., individual specimens) of the clustering tree. Specimens that are very similar (i.e., splits) aggregate close to 0 (i.e., low dissimilarity).

prone to volatilization. In some cases, the ethanol content in one of the split pair was only 50% of its matched split while the urinary concentration of ethanol metabolites such as ethyl glucoside was constant between the split aliquots. This finding suggests that differences in specimen handling had occurred at some stage prior to shipment of the specimens. For two of the specimens, bacterial contamination was the cause of split pair differences, as manifested by relatively high concentrations of benzoic acid, acetate, and lactate with lower glucose in the aliquots.

Type C Misclassifications. Variation in analytical instrumentation or conditions resulting from factors such as efficiency of water suppression or variation in relaxation delay resulted in 32% of misclassified split specimens where specimens were measured using two different acquisition protocols. In particular, quantitative differences were observed in metabolites with relatively long T_1 relaxation times, e.g., TMAO and pyruvate between split samples measured under different acquisition protocols. The percentage of nonidentified specimens attributable to NMR differences fell to 1% of the total number of split specimens when split specimens were measured using identical acquisition parameters. These results show that, even with interinstrument differences, the methodology is highly reproducible and therefore appropriate for population screening. However, these results also indicate that, for quantitative accuracy, it is important to standardize acquisition parameters such as relaxation delays.

The reproducibility analysis reported here assesses not only spectroscopic reproducibility but also specimen handling consistency and specimen quality. The ^{1}H NMR urine spectra contain information relating to biological and chemical degradation of specimen and differences in physicochemical parameters possibly caused by variation or inconsistency in freezing or specimen preparation. Thus, it goes far beyond assessment of instrumental validation. In brief, although ^{1}H NMR spectroscopy has been recognized for decades for being analytically robust, such "in the field" assessment makes possible assessment of stability in terms of biological degradation as well as consistency of NMR spectrometers in time and irregularities in instrumentation running automation mode, such as magnetic field quality and parameter optimization.

Analysis of Variation between Split Samples by Deriving *Rv* Coefficients.[17] The *Rv* coefficient was used to assess the canonical links between tables A−D containing spectra from two separate visits to the clinic for each person with a (blinded split) aliquot for each visit (Figure 1, Table 2). Specifically, it describes the isomorphic relationship between the tables representing the duplicate samples. Together, this analysis yields an *Rv* coefficient

Table 2. Correlation between Split Specimens Based on *Rv* Coefficients[17]

split sample *Rv* coefficient	visit 2	visit 4
Japan (Aito Town)	0.975	0.994
China (Guangxi)	0.909	0.916
USA (Chicago)	0.868	0.941
all three populations	0.917	0.964

of 0.917; analysis of each population sample separately generates *Rv* coefficients between 0.868 (Chicago, IL, USA) and 0.994 (Aito Town, Japan).

Overall Analytical Reproducibility. Analytical reproducibility was demonstrated in three different ways: (i) by coefficients of variation below 5% for genuine metabolic signals of the NMR spectrum, (ii) by blind identification of 71−97% of split specimens, and (iii) by *Rv* coefficients between the split specimen tables providing a similarity measure of 0.994 (where a coefficient of 1 would indicate identical tables). The coefficients of variation obtained for the NMR metabolic signals range 0−10%, with exclusion of citrate signal tails and electronic noise responsible for high CVs in Figure 2G,H. This is better reproducibility than obtained for the validation of a pyrolysis metastable atom bombardment time-of-flight mass spectrometer that had been specially designed for metabolic fingerprinting with coefficients of variation in the range of 20−250%.[23] Moreover, the CV percentage for NMR is favorable when compared to the reproducibility of gene microarray data, yielding coefficients of variation in the interval 20−30%.[24] Other studies in metabolic profiling have also displayed similar high analytical reproducibility. For example, gas chromatography−MS analysis of plant extracts produced an analytical error for several metabolites of 8% compared to a biological variance estimated at 26−56%.[25] The analytical reproducibility in metabolic fingerprinting across a long time scale based on *Rv* coefficients is consistent with previous work assessing NMR reproducibility on urine aliquots analyzed in both the U.K. and Switzerland on NMR spectrometers operating at different frequencies (respectively 600 and 500 MHz), with correlation coefficients for individual metabolites above 0.95.[11] Here we compute higher *Rv* coefficients for NMR signals, indicating that the reproducibility of NMR-based metabonomics makes it highly suitable for research in a medical and epidemiological setting.

Classification of Urine Specimens Based on Population Samples. One ^{1}H NMR spectrum of the urine from one man in each of the three population samples (Aito Town, Japan, Guangxi, China, and Chicago, USA) were selected based on good prediction of population specimen. These spectra are shown in Figure 4. Metabolite identification was based on literature values[26] and on the addition of authentic standards to the urine specimens. A PC-LDA model of population samples based on urinary ^{1}H NMR data using 0.01 ppm segments in the δ0.16−δ9.76 range was built with 29 precompression principal components and assessed by full 3-fold cross-validation (Figure 4, Table 2, and Table 3). PC-LDA maps the individuals by finding the orthogonal directions of maximal variation explaining the differences related to between-group variance (in this study, differences across three population samples: American, Chinese, and Japanese). The score plot corresponding to average profile PC-LDA model (n = 852) shows the discrimination among the three population samples (Figure 4). PC-LD1 segregates the Chinese population (negative scores) from the other two populations (positive scores). These two later American and Japanese populations are respectively discriminated along PC-LD2 with positive and negative scores (Figure 4). The prediction ability of this PC-LDA model was assessed by a full 3-fold cross-validation. Two-thirds of the 852 samples were randomly selected to calibrate the PC-LDA model, and the remaining third was used for independently testing the accuracy of its predictions. The confusion matrix for the cross-validation of this model shows a high prediction rate of 96% (Table 3). Investigation of the PC-LDA loadings indicated several spectral regions were dominating the discrimination of specimens on the basis of geographical origin (population sample). The major metabolites identified in the loadings are given in Table 4. High urinary concentrations of TMAO were particularly dominant in the Aito Town population, consistent with the high dietary intake of fish. TMAO is naturally synthesized as an antifreeze agent in many types of deep-sea fish and is thus accumulated in the tissues

Table 3. Confusion Matrix Derived by 3-Fold Cross-Validation of Pattern Recognition of Population by PC-LDA

observed vs predicted population	Aito Town (Japan	Guangxi (China)	Chicago (USA)
Aito Town (Japan)	245	1	13
Guangxi (China)	2	273	3
Chicago (USA)	8	7	300
prediction rate (%)		96	

Table 4. Selected Biomarkers as Predictors of Population Samples Derived from the Loadings of the PC-LDA Model (n = 852)

metabolite	hierarchy[a] (highest to lowest across the 3 samples)
β-aminoisobutyric acid	Gx only
creatinine	C > AT > Gx
creatine	Gx > C > AT
glucose	AT > C > Gx
hippurate	C > AT > Gx
histidine/methylhistidine	C > Gx = AT
methylated polyols	Gx only
N-methylnicotinic acid/ *N*-methylnicotinamide	C > Gx = AT
taurine	AT > C > Gx
trimethylamine-*N*-oxide	AT > C > Gx
ethanol	Gx[b] > AT > C
ethyl glycoside	Gx[b] > AT > C

[a] Key: C, Chicago; Gx, Guangxi; AT, Aito Town. [b] High concentrations of urinary ethanol are a feature of male Guangxi participants but not of female participants from this population sample.

(23) Dumas, M. E.; Debrauwer, L.; Beyet, L.; Lesage, D.; Andre, F.; Paris, A.; Tabet, J. C. *Anal. Chem.* **2002**, *74*, 5393−5404.
(24) Piper, M. D.; Daran-Lapujade, P.; Bro, C.; Regenberg, B.; Knudsen, S.; Nielsen, J.; Pronk, J. T. *J. Biol. Chem.* **2002**, *277*, 37001−37008.
(25) Fiehn, O.; Kopka, J.; Dormann, P.; Altmann, T.; Trethewey, R. N.; Willmitzer, L. *Nat. Biotechnol.* **2000**, *18*, 1157−1161.
(26) Nicholls, A. W.; Mortishire-Smith, R. J.; Nicholson, J. K. *Chem. Res. Toxicol.* **2003**, *16*, 1395−1404.

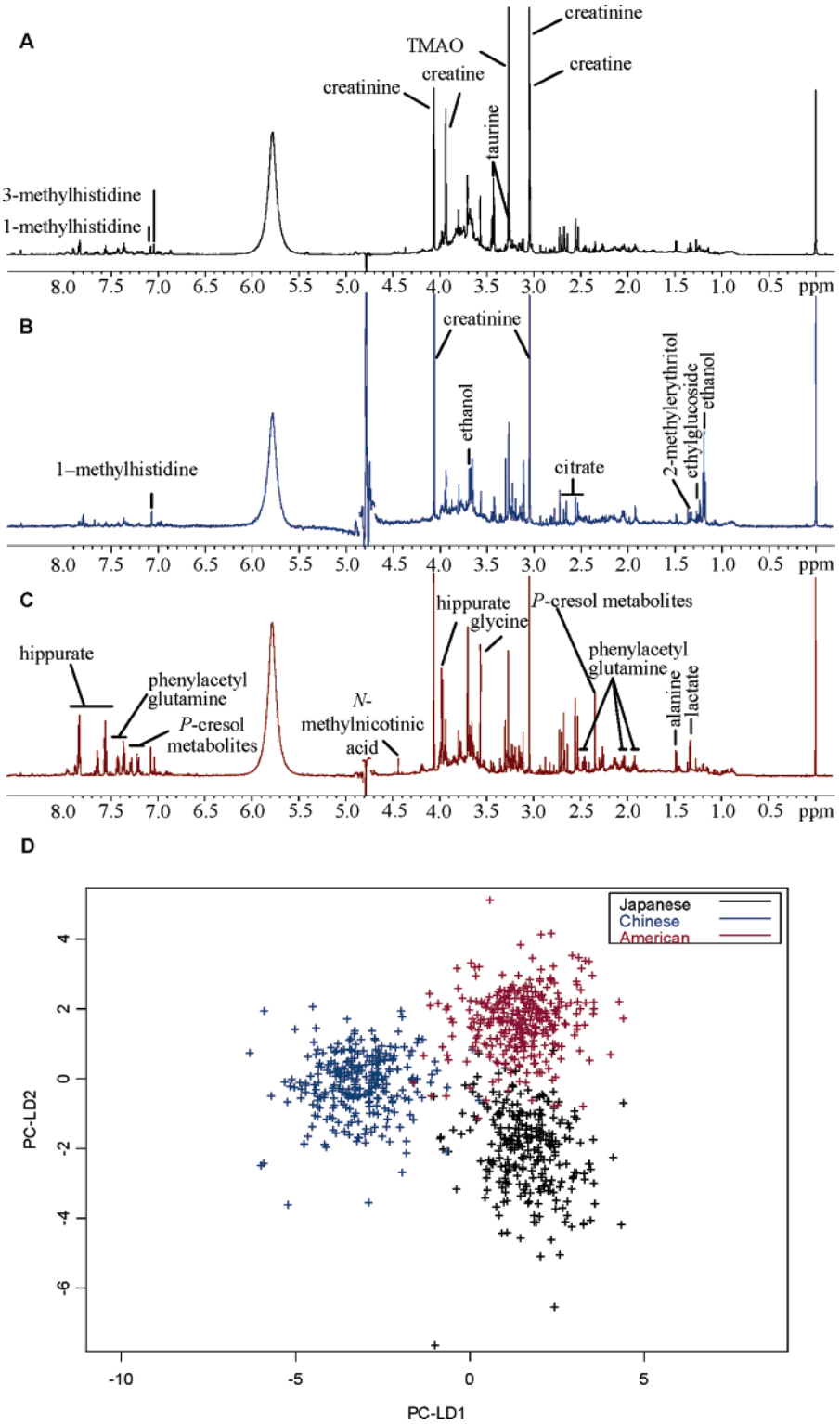

Figure 4. 600-MHz ^{1}H NMR spectra of human urine from (A) a Japanese, (B) a Chinese, and (C) an American man. Principal assignments are marked. Key: TMAO − trimethylamine-*N*-oxide (the inset spectra are shown to give an indication of the spectral complexity); D) Pattern recognition of geographical origin using PC-LDA (n = 852). This model was computed with 0.01 ppm buckets in the δ0.16−9.76 range using 29 principal components for pre-compression before LDA.

in high levels.[27] Guangxi specimens had higher urinary excretion of β-aminoisobutyric acid and ethanol (in male participants only) than American and Japanese specimens. NMR signals from hippuric acid and other phenolic compounds were positively associated with the American population sample. Many of these compounds are derived from gut bacteria rather than of mammalian origin[26,28] and suggest consistent differences in gut microflora across populations samples. Creatinine concentrations were

(27) Yancey, P. H.; Rhea, M. D.; Kemp, K. M.; Bailey, D. M. *Cell Mol. Biol. (Noisy-le-grand)* **2004**, *50*, 371−376.

also higher in the Chicago specimens then those from either Aito Town or Guangxi, due to greater body mass.[29]

CONCLUSION

The results here show that NMR spectroscopy of biofluids combined with multivariate pattern recognition is a robust and precise approach for metabonomics studies, outperforming other "-omic" technologies in terms of reproducibility. Such findings make metabonomics suitable for high-throughput long-term epidemiological studies. This reproducibility also confirms the robustness and the appropriateness of the study protocols, which have been optimized in terms of data acquisition and processing. It implies that large-scale, multi-instrument (and presumably multilaboratory) epidemiological metabonomic studies have a high degree of tolerance to unavoidable differences in environment and specimen handling, without hampering the classification power to identify the source and effects of any such confounding variation. The analyses reported here were at least as reproducible as the most precise functional genomic methods at the protein level,[30] metabolic profiling,[25] and clearly more precise than any differential analysis of gene expression using cDNA/mRNA microarrays.[24] Since ^{1}H NMR enables accurate pattern recognition of populations (96% correct prediction rate), ^{1}H NMR-based metabonomic analysis has the potential to discover metabolic biomarkers possibly related to present-day major public health challenges, hence opening the field of metabonomic epidemiology.

(28) Williams, R. E.; Eyton-Jones, H. W.; Farnworth, M. J.; Gallagher, R.; Provan, W. M. *Xenobiotica* **2002**, *32*, 783−794.

(29) Zhou, B. F.; Stamler, J.; Dennis, B.; Moag-Stahlberg, A.; Okuda, N.; Robertson, C.; Zhao, L.; Chan, Q.; Elliott, P. *J. Hum. Hypertens.* **2003**, *17*, 623−630.

(30) Gygi, S. P.; Rist, B.; Gerber, S. A.; Turecek, F.; Gelb, M. H.; Aebersold, R. *Nat. Biotechnol.* **1999**, *17*, 994−999.

ACKNOWLEDGMENT

We are grateful to the U.S. National Heart, Lung, and Blood Institute, Bethesda, MD, for their financial support of the project. The INTERMAP Study is supported by grant 2-RO1-HL50490, US NHLBI; the INTERMAP Metabonomics Study is supported by grant 5-RO1-HL71950-2. National and local agencies in the four countries have also contributed support to INTERMAP, including Japan Ministry of Education, Science, Sports and Culture grant in Aid for Scientific Research [A] 090357003. M.-E.D. is funded by the Biological Atlas of Insulin Resistance (BAIR) consortium (www.bair.org.uk) (Wellcome Trust Functional Genomics Initiative grant 066786), E.C.M. by INTERMAP, and C.T. by Unilever. NMR signal processing in-house software was developed by Dr. T. Ebbels and Dr. H. Keun (Imperial College London). We also thank L. Smith, Dr. O. Cloarec, and Dr. T. Ebbels for helpful discussion. For a listing of many of the colleagues who contributed importantly to the INTERMAP study, see ref 13.

Received for review September 23, 2005. Accepted January 31, 2006.

AC0517085