A Workflow For Integrated Processing of Multi-Cohort Untargeted ^{1}H NMR Metabolomics Data In Large Scale Metabolic Epidemiology

Ibrahim Karaman[1], Diana L. S. Ferreira[1], Claire L. Boulangé[2], Manuja R. Kaluarachchi[2], David Herrington[3], Anthony Dona[2,4], Paul Elliott[1], John C. Lindon[2,4], and Timothy M. D. Ebbels[2,4*]

[1] Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, St Mary's Campus, Norfolk Place, W2 1PG, London, United Kingdom

[2] Metabometrix Ltd, Bioincubator Unit, Bessemer Building, Prince Consort Road, South Kensington, London SW7 2BP UK

[3] Department of Internal Medicine, Wake Forest School of Medicine, Medical Center Boulevard, Winston-Salem, NC 27157, USA.

[4] Computational and Systems Medicine, Department of Surgery and Cancer, Faculty of Medicine, Imperial College London, Sir Alexander Fleming Building, South Kensington, SW7 2AZ, London, United Kingdom

ABSTRACT: Large scale metabolomics studies involving thousands of samples present multiple challenges in data analysis, particularly when an untargeted platform is used. Studies with multiple cohorts and analysis platforms exacerbate existing problems such as peak alignment and drift correction. Therefore there is a need for robust processing pipelines which can ensure reliable, quality controlled data for statistical analysis. The COMBI-BIO project is aimed at detection of metabolic markers of pre-clinical atherosclerosis, and incorporates plasma from 8000 individuals, in 3 cohorts, profiled by 6 assays in 2 phases using both NMR and UPLC-MS. Here we present the COMBI-BIO NMR analysis pipeline and demonstrate its fitness for purpose through statistical analysis of identical representative quality control (QC) samples interleaved with study samples throughout the analytical run. Standard 1-dimensional ^{1}H-NMR spectra were aligned using the Recursive Segment-wise Peak Alignment algorithm and normalized using the Probabilistic Quotient method. After removing interfering signals, outliers identified using Hotelling's T^2 were removed and a cohort/phase adjustment was applied, resulting in two NMR data sets for each sample. A number of quality assessment metrics were computed to assess the developed pipeline. Alignment of the NMR data was shown to increase the correlation-based $aq_{0.02}$ quality measure from 0.319 to 0.391 for CPMG and 0.536 to 0.586 for NOESY data, showing that the improvement was present across both large and small peaks. End-to-end quality assessment of the pipeline was achieved by examining the distribution of Hotelling's T^2 values across both pooled QC and biological samples. For CPMG spectra, the interquartile range decreased from 1.425 in raw QC data to 0.679 in processed spectra, while the corresponding change for NOESY spectra was 0.795 to 0.636 indicating a substantial improvement in precision following processing. PCA indicated that gross phase and cohort differences were no longer present in the final data sets. Taken together, these results illustrate that the developed pipeline produces robust and reproducible data across thousands of samples, successfully addressing the challenges of this large multi-faceted study.

INTRODUCTION

Metabolic phenotyping using ^{1}H NMR spectroscopy is becoming a widely used approach in modern molecular epidemiology. Owing to its high reproducibility and quantitative accuracy, the technique is particularly amenable to assessing the metabolic status of individuals from large epidemiological cohorts[1]. However, as study sizes increase, the challenge of obtaining high quality data from thousands of blood or urine samples becomes acute. The problems are particularly serious in untargeted assays where the conventional approach of internal standards matched to each analyte of interest cannot be used. Further complications arise from studies combining multiple cohorts, leading to systematic differences in sample composition between the groups. Thus, there is a need for efficient and robust data processing pipelines which can address large and potentially heterogeneous study designs, to ensure reliable, quality controlled data for subsequent statistical analysis.

Pre-processing is very important and challenging step in metabolic phenotyping studies, and particularly so in metabolic epidemiology[2]. Conventional pre-processing of 1-dimensional NMR data will include a Fourier Transform, apodization, baseline correction, phasing and chemical shift calibration. In metabolic phenotyping, large numbers of spectra must be made comparable using tools such as spectral peak alignment, intensity normalization and spectral binning. In addition, outlying samples and possible interfering signals need to be removed prior to statistical analysis. Large studies introduce further problems of accounting for instrument drift during long runs, batch differences, possible merging of data from multiple instruments, and the comparability of data from independent cohorts.

Validation of chemical- and data-analytic protocols is difficult in untargeted metabolomics because of the wide range and unknown identity of the metabolites assayed. However, repeated analysis throughout the run of a quality control (QC) sample has become a standard approach to monitor precision of the measurements[3,4]. QC samples can be prepared from a pool

of the study samples or by use of a representative standard reference material. Since the QC sample is of constant composition, any variation in QC measurements can be used both to monitor and correct for measurement errors.

In this paper, we present a workflow for pre-processing 1-dimensional ^{1}H NMR data from large multiple cohort studies. We focus on data from the COMBI-BIO project, in which ~8000 individuals from three cohorts were profiled with the aim of discovering serum metabolic biomarkers of pre-clinical atherosclerosis. To our knowledge, this is the largest multi-cohort, multi-platform study performed to date. Thus our suggested preprocessing pipeline will be of interest to researchers designing similar large studies using NMR.

MATERIALS AND METHODS

Study population

We used stored serum samples and associated data from randomly selected individuals from three population cohorts: LOLIPOP[5] (The London Life Sciences Prospective Population), MESA[6] (The Multi-Ethnic Study of Atherosclerosis) and Rotterdam[7] (The Rotterdam Study). The recruitment period of the participants for LOLIPOP was 2002-2008; for MESA it was 2000-2002 and for Rotterdam it was 1990-2000. The age range of the participants was 35-74 for LOLIPOP, 45-84 for MESA and 55-85 for Rotterdam. In total, 7,773 serum samples were analyzed in two phases over a period of approximately one year. Each phase corresponded to ~4,000 samples (LOLIPOP: ~1,000, MESA: ~2,000, Rotterdam: ~1,000). Samples were shipped on dry ice and stored at -80 C prior to analysis.

Preparation of samples, including quality controls (QCs)

Two types of QC samples were used to monitor the quality of the NMR data. QC1 samples were derived from a commercially available serum (human serum, off the clot, type AB, VWR catalog number BCHRS01049.2-01). QC2 samples were prepared by pooling equal 50 µl aliquots of the phase 1 LOLIPOP samples. All QC pools were aliquotted in 350 µl and stored at -80 °C prior to analysis.

Both QC and study samples were thawed on the day prior to analysis. 300 µl of each sample was mixed with 300 µl of phosphate buffer (NaHPO4, 0.075M, pH=7.4, as published previously[1] in Eppendorfs for the phase 1 analysis, and in 96 well plates for the phase 2 analysis. After centrifugation (12,000 g at 4 °C for 5 minutes), 550 µl of each sample-buffer mixture was manually transferred into SampleJet 5 mm diameter NMR tubes and kept at 4 °C until analysis. In phase 1 one QC1 sample was incorporated in each 96 tube rack. In phase 2, a single QC2 sample was run in each 96 tube rack, and a single QC1 sample was run every two racks. In the following, we call each combination of phase and cohort a 'batch', since these groups of samples were analyzed in a continuous run on the instrument.

NMR data acquisition

^{1}H NMR spectra were acquired using a Bruker DRX600 spectrometer (Bruker Biospin, Rheinstetten, Germany) operating at 600 MHz. A standard water suppressed 1-dimensional spectrum (NOESY) and a Carr-Purcell-Meiboom-Gill (CPMG) spectrum were obtained for each sample. NMR spectroscopic analysis was completed in six batches corresponding to the three cohorts and two phases.

Figure 1. Proposed pre-processing workflow for CPMG and NOESY NMR data.

PRE-PROCESSING WORKFLOW

Figure 1 presents our proposed workflow for processing NMR data acquired from large multi-cohort, multi-batch studies. In comparison to smaller, single batch studies, it is necessary to modify the steps in the processing pipeline to address the key challenges of large studies. Modified steps include chemical shift alignment, removal of interfering spectral signals and outlying samples, normalization, cohort/batch correction and, optionally, binning. The workflow was implemented in MATLAB version 8.1 (Mathworks Inc., USA).

Raw data processing and generation of one dataset from several data tables

All spectra were automatically phased and baseline corrected using Bruker instrument control software Topspin version 3.2 (Bruker Biospin, Rheinstetten, Germany). Since internal standards such as TSP exhibit significant protein binding which affects the peak shape and position, chemical shifts were calibrated to the glucose doublet at δ 5.23. The chemical shift range of the spectra was clipped to δ 0.50−δ 9.00 since no bona fide metabolite signals are observed in serum outside this region. Finally, the six data sets were concatenated to produce one large data table consisting of 34,001 variables and 7,872 samples in the CPMG dataset and 7,869 samples in the NOESY dataset including the QCs.

Spectral peak alignment

Prior to spectral peak alignment, the region δ 4.40–5.10 corresponding to residual water signals was removed. The table was split into six consecutive chemical shift slices for alignment to ameliorate high computer memory demands. The cut points (δ 1.45, 2.64, 3.33 and 6.00) were selected to be in regions containing no sharp resonances. Alignment was performed using RSPA (Recursive Segment-wise Peak Alignment[8]). This algorithm is appropriate for large data sets as it is fast and has been

shown to improve alignment of small peaks. After alignment the slices were concatenated to form a single data table.

Removal of interfering regions

In addition to the spectral region related to the water suppression residual, there may be other regions which contain interferents which can cause errors in further analysis. A common contaminant in clinical and epidemiological studies is methanol. This was also observed in our study requiring removal of the region δ 3.375–3.400. Suspected interferents in the regions δ 1.180–1.240, δ 2.244–2.261, and δ 3.660–3.710 were also removed. Selection of the interfering spectral regions may not be straightforward and may require expert-driven suggestions. After removing interfering signals, the datasets contained 30,590 data points.

Normalization

Normalization is the process of applying a spectrum-wide scale factor to each spectrum to correct for global variations in the NMR signal. These could be due to, for example, variable dilutions or small changes in instrument calibration over the course of a large run. In this study, we used probabilistic quotient normalization (PQN[9]) in which the median intensity ratio between each sample spectrum and a reference is normalized to unity. PQN has been shown to outperform earlier methods such as total intensity normalization and is fast and memory efficient. We used the median spectrum of the full data as the reference. Note that normalization using a reference must be performed after alignment, since intensity ratios prior to alignment may be influenced by uncontrolled peak shifts.

Removal of outlying samples

Before applying statistical analysis, removal of outliers is essential. Strong outliers due, for example, to instrument malfunction during measurement can be detected by investigating Hotelling T^2 values of the samples, calculated using the scores of a principal component analysis (PCA). The spectra of suspected samples are then examined as to whether they are analytical outliers and, if so, discarded from the sample set. In the present study, we constructed separate PCA models for both CPMG and NOESY datasets. We excluded 3 outliers from CPMG dataset and 4 outliers from NOESY dataset where the outliers demonstrated extreme Hotelling T^2 values ($>10T^2_{crit}$ at 95% confidence level). Excluded samples were attributable to spectra with poor water suppression and baseline distortion.

Spectral binning

Statistical analysis can be applied to high-resolution spectra so that all the information in the dataset is used. However, in order to decrease the number of variables and alleviate minor residual misalignments and peak shape differences, spectral binning can be applied. Since each approach has advantages, we decided to apply both, to produce both a high-resolution and binned version of each data set, and to apply biomarker mining approaches to each. We binned the data using statistical recoupling of variables (SRV[10]) which generates bins by searching for adjacent correlated structures in the high-resolution spectrum. The minimum number of variables to generate a bin depends on the resolution of the spectra. In this study, we chose to use a minimum of 10 variables ($2.5x10^{-3}$ ppm or 1.5 Hz) per bin. The automatic bin positions were reviewed by an expert NMR spectroscopist to ensure that the binning produced no artifacts such as split peaks, resulting in minor adjustments being made to around 2% of the bins. Figure 2 shows an example of the binning process on the mean CPMG spectrum. After binning, the number of variables for CPMG and NOESY NMR datasets decreased to 468 and 447, respectively.

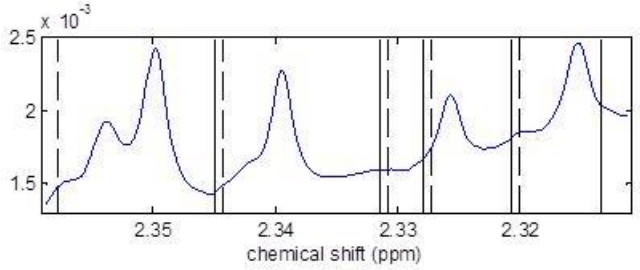

Figure 2. Representative example of SRV spectral binning on a region of the mean CPMG spectrum. Each bin starts at a solid line and ends at the position of a closest dashed line on the left.

Phase and cohort adjustment

In large-scale studies with samples with different origins, variation can be observed with respect to the origin. In the present case we have three cohorts and two phases of data acquisition. The variation due to cohorts may occur due to different sample composition, but also collection and storage conditions, whereas the phase variation is related to different time periods of NMR analysis. Given that the biomarker discovery is aimed at finding cohort-independent signatures of disease, it was deemed appropriate to remove all cohort and phase differences in mean levels prior to statistical analysis with a mean-centering operation[11]. We therefore mean-centered each of the six phase/cohort batches separately and subsequently concatenated the data back into a single table.

RESULTS AND DISCUSSION

The pre-processing workflow was applied to each of the CPMG and NOESY NMR datasets at hand. The final datasets consisted of 7,869 samples for CPMG and 7,865 samples for NOESY. High resolution versions of the NMR datasets both contained 30,590 variables, whereas binned versions of CPMG and NOESY contained 468 and 447 variables respectively. In the following we illustrate the workflow using the CPMG high-resolution data. Results for NOESY and binned data are similar and can be found in Supporting Information.

Assessment of spectral peak alignment

Figure 3 shows the results of spectral peak alignment on the CPMG data. Peak alignment quality was initially assessed visually using the visualization seen in the figure. A clear improvement can be observed in the heat maps, for example the doublet at ~1.5 ppm. Panels (c) and (d) show how much variation in position across the data set is present at each chemical shift. This clearly shows more stability (sharper peaks in the distribution) after the alignment procedure.

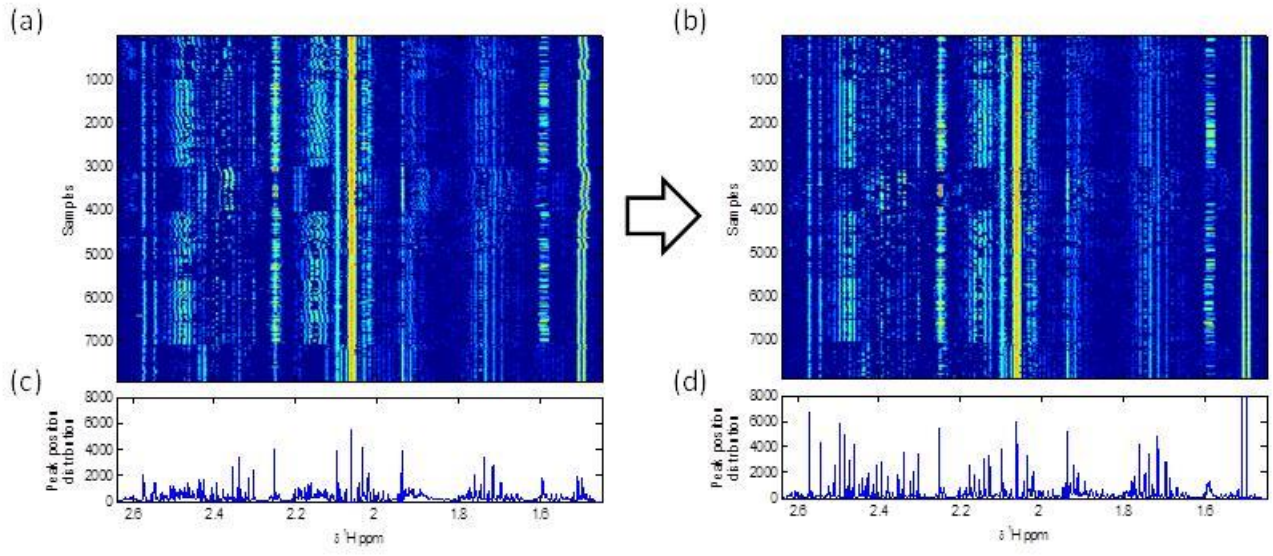

Figure 3. Illustration of spectral peaks on a representative region of CPMG NMR dataset before (a) and after (b) alignment. Panels (c) and (d) show the peak position distribution in each region. Larger peak position distribution values indicate that the peaks are better aligned.

To evaluate the quality of the alignment more objectively, we follow Veselkov et al.[8] and calculate quality measures based on correlation between appropriately scaled pairs of aligned spectra. The *i*th spectrum is first divided into a grid of *K* adjacent regions each of width *w*. To account for large variations in peak intensities, the raw intensities $\boldsymbol{s}_{ik}^{\mathbf{raw}}$ of the *k*th region in the *i*th spectrum are centered and scaled to unit variance:

$$\boldsymbol{s}_{ik} = \frac{\boldsymbol{s}_{ik}^{\mathbf{raw}} - \mu_{ik}^{\mathbf{raw}}}{\sigma_{ik}^{\mathbf{raw}}} \qquad (1)$$

where $\mu_{ik}^{\mathbf{raw}}$ and $\sigma_{ik}^{\mathbf{raw}}$ denote the mean and standard deviation of raw intensities in spectrum *i* and region *k*. The scaled intensities are reassembled $\boldsymbol{S}_i = (\boldsymbol{s}_{i1}, \boldsymbol{s}_{i2}, \ldots, \boldsymbol{s}_{iK})$ and the quality metric aq_w is defined as

$$aq_w = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{i-1} cc(\boldsymbol{S}_i, \boldsymbol{S}_j) \qquad (2)$$

where *cc* is the Pearson correlation coefficient and *n* is the number of spectra.

We assessed alignment quality through calculation of the alignment quality metric aq_w (**Error! Reference source not found.**). We chose bin sizes of w = 0.08 ppm to focus on large peaks only and w = 0.02 ppm to up-weight the contribution from small peaks. The *aq* values for the aligned data were found to be significantly higher than those for the unaligned data for both bin sizes (one-sided, paired t-test, p=0 indicating a successful peak alignment across the QC and biological sample spectra.

Assessment of the pre-processing workflow via the quality control samples (QCs)

An assessment of the overall workflow was achieved by monitoring the QC samples. Since each QC type (QC1 & QC2) is of constant composition, the dispersion in the QC measurements reveals the level non-biological variation, derived from the analytical and data analysis pipeline, which may also be present in the biological samples.

Table 1. Alignment quality (aq_w) measures for the datasets with different bin sizes in ppm.

		CPMG		NOESY	
sample type		*unaligned*	*aligned*	*unaligned*	*aligned*
QC1	$aq_{0.02}$	0.349	0.415	0.582	0.636
	$aq_{0.08}$	0.468	0.532	0.806	0.846
QC2	$aq_{0.02}$	0.407	0.456	0.654	0.677
	$aq_{0.08}$	0.524	0.570	0.872	0.887
BIO	$aq_{0.02}$	0.319	0.391	0.536	0.586
	$aq_{0.08}$	0.439	0.511	0.764	0.801

A qualitative impression of the QC measurements can be gained from observing the position of the QCs on a PCA scores plot, along with the biological samples. Figure 4 shows such a plot for the first two components of a PCA model of the mean-centered datasets. The QCs form two clusters according to the QC type and seem to become less scattered if unaligned and normalized versions are compared. To quantify the improvement in QC clustering seen in the figure, we calculated the ratio *r* of the sum of variances of the first two scores of each QC type to the sum of variances of the first two scores of the biological samples.

$$r = \frac{var(\mathbf{t}_1^{QC}) + var(\mathbf{t}_2^{QC})}{var(\mathbf{t}_1^{BIO}) + var(\mathbf{t}_2^{BIO})} \qquad (3)$$

where $var(\mathbf{t}_i^{QC})$ and $var(\mathbf{t}_i^{BIO})$ are the variances of the *i*'th score vectors for QCs and biological samples respectively. It was found that this ratio decreased from r=3.59% in the unaligned data to 1.64% after normalization for QC1. For QC2 the value decreased from 1.77% to 0.81%.

A further quality assessment can be made by analyzing the distribution of Hotelling T^2 values for the QC samples[1]. The Hotelling T^2 value is proportional to the squared Mahalanobis distance of a sample from the origin in the score space[12] and

thus takes into account all components in the PCA model. Figure 5 shows box plots of the Hotelling T^2 values for QC1, QC2 and biological samples, computed from PCA models explaining at least 95% of the variance in the data (six PCs). For the CPMG dataset, the QC1 interquartile range decreased from 1.425 in the raw data to 0.679 after alignment, normalization and outlier removal. The range for QC2 decreased similarly, from 0.471 to 0.355 after the pre-processing steps. For the NOESY dataset, the range for QC1 decreased from 0.795 to 0.636; the range for QC2 decreased from 0.763 to 0.426. Note that no information from the QCs used in pre-processing the data, so this decrease can be interpreted as a genuine effect of the processing pipeline. Therefore, it can be concluded that the pre-processing pipeline improved the quality of the data, as measured by the quality control samples.

On the PCA score plot of the pre-processed data in Figure 4c, the samples of the MESA and LOLIPOP cohorts appear to overlap whereas the samples of the Rotterdam cohort do not. Whilst there are real biological differences between the cohorts (e.g. age, diet, lifestyle), this variation might also have occurred due to sample collection and storage protocols. Although these biological and methodological effects are confounded, the impact is minor, since subsequent biomarker mining is aimed at finding consistent relationships between metabolic phenotype and clinical outcomes within, not between cohorts. In addition, within each cohort, the two phases introduce potential batch effects which should be taken into consideration. Therefore, we adjusted the data (biological samples only) for both phase and cohort as in step 7 in the workflow. After adjustment the batch differences are no longer apparent (Figure 4d).

At the end of the assessment of the pre-processing workflow, the final versions of the high resolution and binned spectra (CPMG and NOESY) were deemed suitable for further statistical analysis.

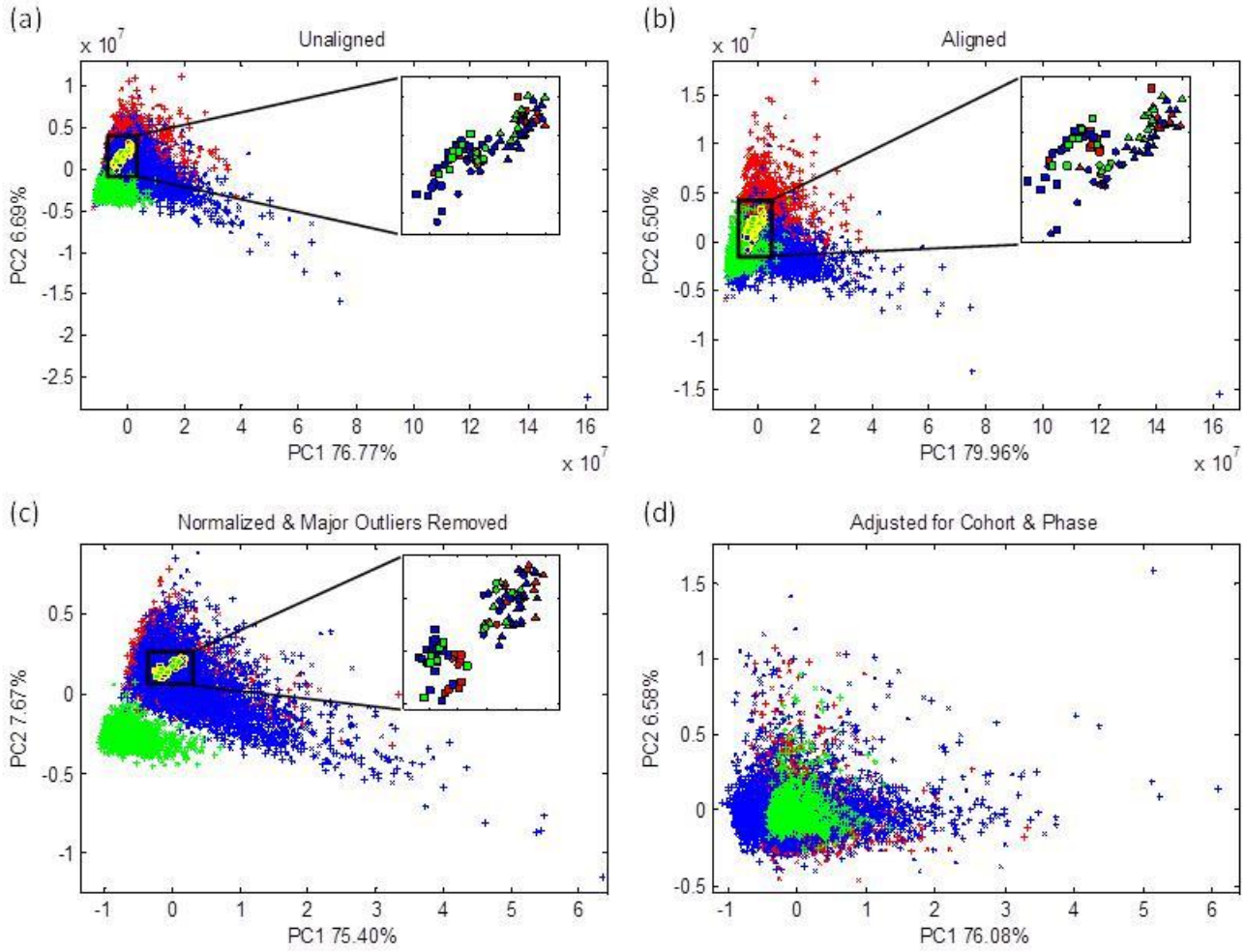

Figure 4. PCA score plots generated using a) unaligned and b) aligned CPMG data without the interfering regions. c) Score plot of the CPMG data after normalization and the removal of major outliers. d) Score plot of the CPMG data after adjustment for cohort and phase. Colors correspond to cohorts (red: LOLIPOP, blue: MESA, green: ROTTERDAM). Symbols for the biological samples vary according to phases (×: phase 1, +: phase 2). Zoomed frames show the QC samples (△: QC1 analyzed in phase 1, ○: QC1 analyzed in phase 2, □: QC2 analyzed in phase 2). Axis labels indicate the percentage variance explained by each principal component (PC).

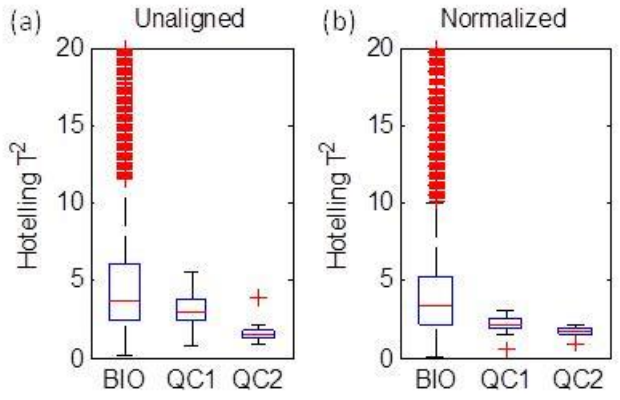

Figure 5. Box plots of Hotelling T^2 values for the CPMG dataset (QC1, QC2 and biological samples, BIO) corresponding to a) unaligned data b) aligned and normalized data with major outliers removed.

CONCLUSION

Obtaining high quality analytical data in large metabolic epidemiology studies is fraught with difficulties. We have presented a general workflow for pre-processing such large NMR data sets. To our knowledge this is the first workflow addressing the issue of multi-cohort large-scale studies in untargeted NMR metabolomics. Careful end-to-end analysis using multiple repeatedly analyzed QC samples enabled us to monitor and control the quality of the resultant data sets. Overall, the approach was able to improve the precision of the data, in several different measures of quality, as compared to that entering the pipeline. The strategy presented here will be of relevance to scientists designing large studies where large numbers of samples are to be assayed by an untargeted NMR platform.

ASSOCIATED CONTENT

Supporting Information

Supporting Information is available free of charge on the ACS Publications Website at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*Email: t.ebbels@imperial.ac.uk

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENT

The authors are grateful for access to samples and data, as well as useful discussions with the wider COMBI-BIO team. We acknowledge support from the EU COMBI-BIO project (grant agreement 305422). TE acknowledges support from the EU COSMOS project (grant agreement 312941).

REFERENCES

(1) Dona, A. C.; Jiménez, B.; Schäfer, H.; Humpfer, E.; Spraul, M.; Lewis, M. R.; Pearce, J. T. M.; Holmes, E.; Lindon, J. C.; Nicholson, J. K. *Analytical Chemistry* **2014**, *86*, 9887-9894.
(2) Tzoulaki, I.; Ebbels, T. M.; Valdes, A.; Elliott, P.; Ioannidis, J. P. *Am. J. Epidemiol.* **2014**.
(3) Sangster, T.; Major, H.; Plumb, R.; Wilson, A. J.; Wilson, I. D. *Analyst* **2006**, *131*, 1075-1078.
(4) Dunn, W. B.; Broadhurst, D.; Begley, P.; Zelena, E.; Francis-McIntyre, S.; Anderson, N.; Brown, M.; Knowles, J. D.; Halsall, A.; Haselden, J. N.; Nicholls, A. W.; Wilson, I. D.; Kell, D. B.; Goodacre, R. *Nat Protoc* **2011**, *6*, 1060-1083.
(5) Chambers, J. C.; Obeid, O. A.; Refsum, H.; Ueland, P.; Hackett, D.; Hooper, J.; Turner, R. M.; Thompson, S. G.; Kooner, J. S. *The Lancet* **2000**, *355*, 523-527.
(6) Bild, D. E.; Bluemke, D. A.; Burke, G. L.; Detrano, R.; Diez Roux, A. V.; Folsom, A. R.; Greenland, P.; Jacob, D. R., Jr.; Kronmal, R.; Liu, K.; Nelson, J. C.; O'Leary, D.; Saad, M. F.; Shea, S.; Szklo, M.; Tracy, R. P. *American journal of epidemiology* **2002**, *156*, 871-881.
(7) Hofman, A.; van Duijn, C.; Franco, O.; Ikram, M. A.; Janssen, H. A.; Klaver, C. W.; Kuipers, E.; Nijsten, T. C.; Stricker, B. C.; Tiemeier, H.; Uitterlinden, A.; Vernooij, M.; Witteman, J. M. *Eur J Epidemiol* **2011**, *26*, 657-686.
(8) Veselkov, K. A.; Lindon, J. C.; Ebbels, T. M. D.; Crockford, D.; Volynkin, V. V.; Holmes, E.; Davies, D. B.; Nicholson, J. K. *Analytical Chemistry* **2009**, *81*, 56-66.
(9) Dieterle, F.; Ross, A.; Schlotterbeck, G.; Senn, H. *Analytical Chemistry* **2006**, *78*, 4281-4290.
(10) Blaise, B. J.; Shintu, L.; Elena, B.; Emsley, L.; Dumas, M.-E.; Toulhoat, P. *Analytical Chemistry* **2009**, *81*, 6242-6251.
(11) van Velzen, E. J. J.; Westerhuis, J. A.; Van Duynhoven, J. P. M.; Van Dorsten, F. A.; Hoefsloot, H. C. J.; Jacobs, D. M.; Smit, S.; Draijer, R.; Kroner, C. I.; Smilde, A. K. *Journal of Proteome Research* **2008**, *7*, 4483-4491.
(12) Brereton, R. G. *Journal of Chemometrics* **2015**, n/a-n/a.

Table of Contents artwork