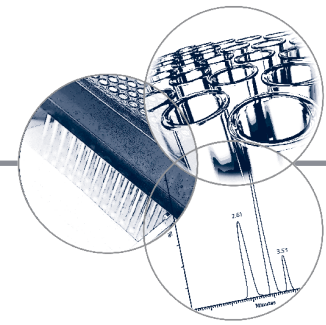
Review
Special Focus Issue: Metabolomics



The importance of experimental design and QC samples in large-scale and MS-driven untargeted metabolomic studies of humans

The metabolic investigation of the human population is becoming increasingly important in the study of health and disease. The phenotypic variation can be investigated through the application of metabolomics; to provide a statistically robust investigation, the study of hundreds to thousands of individuals is required. In untargeted and MS-focused metabolomic studies this once provided significant hurdles. However, recent innovations have enabled the application of MS platforms in large-scale, untargeted studies of humans. Herein we describe the importance of experimental design, the separation of the biological study into multiple analytical experiments and the incorporation of QC samples to provide the ability to perform signal correction in order to reduce analytical variation and to quantitatively determine analytical precision. In addition, we describe how to apply this in quality assurance processes. These innovations have opened up the capabilities to perform routine, large-scale, untargeted, MS-focused studies.

The environment in which humans live is diverse and provides many different physical and (bio)chemical challenges to the human body. Although we view the genome as a static library of genes, variability between individuals is observed as a result of, for example, single nucleotide polymorphisms [1] and epigenetic changes [2]. This variability results in great diversity in our appearance, how we act and how we respond to environmental and other stimuli. The complex interaction of the human genomes and their environment are imprinted in the collection of xenobiotics and biochemicals in the human body; this constantly changing dynamic picture of interaction is defined as the phenotype. Clearly, the phenotype of each individual is different and there is large inter- and intra-individual variability. To statistically define the large phenotypic variability in the human population there is the requirement to sample large numbers from that population; generally hundreds or thousands of subjects are investigated to acquire a statistically robust result, depending on the fold change observed, the study's objective and **experimental design**. Only in specific examples, where the genetic and environmental variability is controlled to a significantly higher level (e.g., rodent models of disease applied in the laboratory [3] or in studies collecting multiple time-point samples from the same subjects [4]), can lower sample numbers be used. Even with careful selection of a suitably large and representative cohort of individuals, there are numerous additional sources of bias (otherwise known as confounding factors) that can lead to failure to discover anything of true significance, or more seriously, report spurious findings that prove impossible to validate. Collection of data related to demographic, lifestyle and physiological factors (e.g., diet, gender, ethnicity, age and BMI) can assist in identifying confounding factors and ensuring the appropriate data can be incorporated into the experimental design and data analysis processes.

The quantitative collection of endogenous and exogenous metabolites, defined as the **metabolome**, provide an appropriate route to determine the phenotype(s) of individuals or populations and can be defined as the metabotype [5]. The study of the metabolome is defined as metabolomics and has been reviewed recently with a focus on mammalian investigations [6]. In humans the metabolome is large; the Human Metabolome Database is the most comprehensive definition of the human metabolome and describes approximately 7900 metabolites [7], although other less common metabolites, complex lipids and exogenous metabolites are currently not described [6]. The metabolome is the final downstream product of gene transcription

Warwick B Dunn*[1], Ian D Wilson[2], Andrew W Nicholls[3] & David Broadhurst[4]

[1]Centre for Advanced Discovery & Experimental Therapeutics, Institute of Human Development, University of Manchester & Manchester Academic Health Sciences Centre, Central Manchester NHS Foundation Trust, York Place, Oxford Road, Manchester, M13 9WL, UK
[2]Biomolecular Medicine, Department of Surgery & Cancer, Faculty of Medicine, Sir Alexander Fleming Building, Imperial College London, London, SW7 2AZ, UK
[3]Investigative Preclinical Toxicology, GlaxoSmithKline, David Jack Centre for Research and Development, Park Road, Ware, Hertfordshire, SG12 0DP, UK
[4]Department of Medicine, Katz Group Centre for Pharmacy & Health, University of Alberta, Edmonton, Alberta, Canada
*Author for correspondence:
Tel.: +44 161 7010239
Fax: +44 161 7010242
E-mail: warwick.dunn@manchester.ac.uk

FUTURE SCIENCE part of fsg

Key Terms

Experimental design: Design of a study to acquire data related to a specific biological question while ensuring that covariants or confounders are not present or are well characterized.

Metabolome: The total qualitative and quantitative collection of metabolites present in a defined sample. Many metabolomes are present in the human body including the serum, urine, endothelial cell and liver tissue metabolomes.

Untargeted metabolomic studies: Holistic study of crude sample extracts applying chromatography and/or MS or NMR spectroscopy with limited *a priori* knowledge of metabolome composition; applied in discovery studies.

Quality assurance: Planned process activities to ensure the quality of data produced meets a specified acceptance level.

and protein translation and so closely defines the phenotype [8]. The dynamics of the human body are mirrored in the metabolome; for example, after eating a meal the quantitative composition of blood and urine metabolomes change in timescales of minutes to hours [9]. The holistic investigation of the metabolome allows high-throughput studies to be performed at a relatively low cost per sample in comparison with transcriptomics and proteomics [6]. Metabolomics has been applied to investigate how lifestyle (e.g., exercise [10]) and the environment (e.g., food intake [11]) impact on humans, how we develop diseases [12] and to develop prognostic or diagnostic biomarkers or risk factors [13,14]. In small studies, biofluids [15], cells [16] and tissues [17,18] are applied; the limited number of samples collected does not place too high a demand for skilled and coordinated collection for the more difficult to collect samples, most notably tissues. In large-scale studies involving hundreds to thousands of different subjects, it is technically demanding to collect tissue and/or cell samples, although tissue banks are being constructed; therefore, most studies involve the investigation of biofluids as these are easily collected. Common biofluids studied include blood plasma [13,14] and urine [19].

Metabolomic studies are operated with one of three different analytical strategies: targeted, semi-targeted and **untargeted metabolomic studies** [6]. Targeted methods study a limited number of predefined metabolites (typically fewer than 20) with high levels of specificity, precision and accuracy to define absolute amounts of each metabolite. For example, the quantitation of amino acids using LC–triple-quadrupole MS [20]. Semi-targeted methods apply the same or different analytical platforms to quantify predefined metabolites, but the number of metabolites studied is increased to the low hundreds [13,21]. Untargeted methods define the relative concentrations of hundreds or thousands of metabolites with fit-for-purpose precision, although with a lower analytical specificity. The metabolites are not predefined and the identification of biologically important metabolites is performed post-data acquisition, which is a significant bottleneck [22]. The method applied is highly dependent on the study objectives and available analytical platform(s). In true discovery studies, where the biologically important metabolites defining the phenotype are not known, untargeted studies provide the greatest opportunity to identify unexpected changes through the application of methods that detect the largest number of metabolites. One advantage of untargeted studies is the ability to observe changes in unknown metabolites or in metabolites not commonly reported or detected. In untargeted studies, data provide relative comparisons between samples (metabolite concentrations are not reported) compared with targeted studies that provide quantitative data related to metabolite concentrations. Whichever analytical strategy is used, significant attention to detail and **quality assurance** (QA) is required for all analytical strategies (targeted and untargeted), especially for large-scale studies.

The most frequently applied analytical platforms in untargeted metabolomics studies are NMR spectroscopy [10,19] and MS via direct injection, or more commonly hyphenated to a chromatographic technique, such as GC–MS [23,24] and LC–MS (along with associated developments including UPLC) [14,23]. MS and NMR platforms provide advantages and limitations in their application in metabolomic studies, as has been previously discussed [6]. The integration of multiple analytical platforms provides greater scientific power to these studies [6,23]. Until recently, NMR spectroscopy was the chosen analytical platform for large-scale untargeted studies; the platform provided reproducible data across multiple analytical experiments [19,25]. MS could not provide this reproducibility in large-scale untargeted studies composed of many hundreds or thousands of sample analyses, although was operated routinely in small- and large-scale targeted studies and small-scale untargeted studies. In targeted studies, comparison of data to calibration curves in each separate analytical experiment provided reproducible and quantitative data that could be easily integrated across different analytical experiments. However, in the last 5 years the development of innovative experimental protocols and data processing methodologies has allowed GC–MS and UPLC–MS to be applied in large-scale untargeted studies of the human population [23,26]. All untargeted metabolomics studies should apply analytical methods that have been developed and validated to provide reproducible and robust data. A number of protocols have been published that have defined in accurate detail technical aspects including sample preparation and chromatography–MS operation (e.g., routine maintenance). This important aspect of any analytical study will

not be discussed further here and the reader is directed to a number of protocol papers for further information (for serum/plasma and urine see [23,27]).

In this article we will discuss, with a focus on untargeted and large-scale studies:

- The difficulties of applying chromatography and MS platforms to large-scale metabolomics studies;
- The importance of experimental design;
- The integration of QC samples into studies and their role in QA;
- The importance of robust data preprocessing.

All of these factors, rather than a single factor, are important in performing robust and biologically significant large-scale, MS-focused, untargeted metabolomics studies of the human population.

Categories of large-scale metabolomic studies in the human population

Metabolomic studies of the human population usually fall into three basic categories: predictive biomarker discovery, with the aim of translation into diagnostic/prognostic tests; pathogenesis studies, where the aim is to uncover mechanism of disease; or association studies, where the aim is to find correlations between the human metabolome and factors such as demographic, lifestyle and physiological factors.

It is important to understand that although all three of these study categories may be based on identical core patient populations and involve similar biological samples, the study designs may differ considerably. There is a huge difference between uncovering a statistically significant association between a biomarker and disease compared with finding a biomarker of disease with the potential to be useful as a discriminator in a clinical setting. An odds ratio of approximately two may be sufficient to suggest an association that may help elucidate disease etiology or pathogenesis; whereas, it is realistic to expect an odds ratio in the region of >20 (or an associated AUC of >0.9) for a diagnostic biomarker [28]. That said, in multifactorial diseases, it is often the combination of multiple 'weak' individual markers into a single 'strong' multivariate model that provides the required high levels of discrimination. Unfortunately, the use of multivariate methodologies can easily be abused, with the very high possibility of over-fitting to random associations, thus giving a false impression of the true predictive ability of the candidate biomarker signature. Careful cross-validation procedures are imperative [29,30].

When a multivariate biomarker is expected, then a structured design of the experiment is necessary and important. Studies need to be sufficiently powered to produce meaningful measures of specificity and sensitivity and the use of cross-validation procedures means that test subjects need to be carefully selected to ensure that they are sufficiently homogeneous with regard to demographic/lifestyle/physiological factors (e.g., sex, ethnicity, age and BMI) such that hold-out sets (a representative selection of the complete dataset, typically 10–30% of samples) applied in model validation are equally representative [29]. In addition, data on potential confounding factors between cases and control should also be collected and incorporated into the study design and subsequent statistical analysis. If all the above considerations are effectively applied, it is inevitable that the size of discovery-phase studies must grow. For example, if we consider the power calculations alone using the standard inferential approach described by Arkin and Wachtel [31], for a study in which we hypothesize that a clinically effective screening test will be observed to have a sensitivity of at least 0.85 with a corresponding specificity of 0.95, and assuming a 95% CI in sensitivity of ±0.05 is sufficiently precise, we will require 195 cases. If, in addition, we match four controls to each case to ensure good representation of the target population for cross-validation, then the number rapidly increases to 780 patients.

For association studies, again sample numbers need to be large. Here, the aim of a given study may be to find subtle, but significant, associations between disease, clinical factors and the metabolome. Basic statistical theory dictates that as the size of the main effect decreases, the number of measured subjects needed for that effect to be significant increases nonlinearly. In addition, if modern methods for correcting for confounding factors are used, such as logistic regression, the number of required subjects increases again, particularly for categorical confounders such as gender or ethnicity. In order for such studies to be clinically relevant, they also have to be cross-sectional (i.e., subjects need to be representative of the population from which they are

Key Term

Standard operating procedure: Written procedure available to all scientists to allow experimental tasks to be performed at different sites and by different people following suitable training.

sampled). Indeed, study design must follow the robust protocols set out by the epidemiology community [32], where studies typically involve many thousands of subjects.

What are the factors to consider in the experimental design of large-scale untargeted metabolomics studies?

The appropriate design of scientific studies is critical to ensure robust scientific conclusions are reached. By definition, experimental design is the plan constructed to perform data-gathering studies; in other words, providing the appropriate foresight to plan a study to ensure that the variation related to the biological observations are significantly greater than process variation – the variation introduced by performing the study. Without foresight and design, large experimental datasets can be acquired that provide no relevance to the biological objectives or that provide data wthat are not robust and can lead to false observations and biological conclusions.

■ Reproducibility of sample collection & preparation across single or multiple sites & over long periods of time

In studies of the human population, samples are frequently collected at multiple sites either within one country or across different countries and continents. It is common that two separate researchers will collect and process a sample differently unless the sampling and processing steps are clearly defined in a **standard operating procedure** (SOP) and the users are fully trained. Differences in how samples are collected, processed, stored and transported have the potential to impact on the metabolic profile determined [33]. This is especially true for blood serum and plasma, which, in comparison with urine, contain concentrations of enzymes that provide the capability for metabolism to operate post sample collection. Without the appropriate quenching of metabolism, the metabolic profile of the sample analyzed will differ to the metabolic profile of the sample at the time of collection. Therefore, the application of a single validated SOP and training across all sampling sites is essential to minimize any intra- or inter-researcher and inter-site process variation and its impact on the metabolic profile. Validation of sample collection and transport SOPs for large-scale population sampling have been performed and indicate that with appropriate processes in place the variation introduced during the processing and transport of samples is small compared with the inter-individual variation associated with the subject samples [34,35].

It is recommended that sample tubes from a single manufacturing batch are applied for a complete study where possible and that chemical contamination checks of a random selection of tubes are performed to ensure that no chemicals are present that can interfere with the assays applied [23]; most chemical contaminants are low-molecular weight species with similar chemical and physical properties to metabolites. In 'case–control' studies, it is often tempting to use control samples from a previous or independent study in which collection procedures are unknown, and therefore cannot be duplicated for collecting case samples. This is not recommended as it introduces a potential bias that is impossible to correct for in subsequent statistical analysis.

In large-scale studies, samples are collected across a single or multiple sites and are then transported to a single center to perform the metabolomics research. When the sample size is large (typically n > 100), the biological study is divided into smaller analytical experiments with each experiment comprising sample preparation, data acquisition and data preprocessing. The complete set of analytical experiments can be performed across weeks, months and years. Therefore, and as for sample collection, SOPs and training for all researchers involved is essential to ensure minimal variation is introduced during the sample preparation process, as well as subsequent steps, which are discussed below. Without these steps to control variability, intra-experiment variance (where one person performs the sample preparation on a single day) will be small, but inter-experiment variability (performed on different days/weeks/months and potentially by more than one researcher) can be large. The SOP should accurately define volumes, times and temperatures to apply in sample preparation.

■ The requirement for multiple analytical experiments & routine MS maintenance

In untargeted studies applying MS, where the sample population is large (hundreds to thousands of samples), analysis of all samples cannot be achieved in a single analytical experiment. To achieve an appropriate analysis of the sample, the large biological study is divided into smaller analytical experiments, typically comprising 50–150 samples. Data from these multiple analytical experiments are combined

post data acquisition. This requirement to divide the biological study into smaller analytical experiments is necessary because of the drift in chromatographic and mass spectrometric performance over time [36]. Samples are analyzed following minimal sample preparation, are complex and contain high levels of matrix components and metabolites. Matrix components and metabolites physically interact with chromatography and MS platforms and are the primary reason for degradation in analytical performance. Typically, the greatest issue is a change in sensitivity (normally a decrease) as sample components aggregate in the GC injector (affecting volatilization of metabolites) or electrospray ion source (affecting ion transmission from atmospheric to vacuum regions of the MS). The build-up of matrix and metabolites on GC and UPLC columns can also cause changes in chromatographic performance and introduce variability in retention-time data. Lipophilic metabolites (e.g., triglycerides and phospholipids) can accumulate on reversed-phase UPLC columns and hydrophilic metabolites (e.g., salts at high concentrations, including phosphate and ammonium) can accumulate on HILIC columns. Other non-sample-based effects include drift of operation of electronic components and, for example, its influence on mass calibration.

'Drift' or variability in measured variables (response, *m/z* and retention time) is low in the first 100–150 injections following routine maintenance, but generally becomes significant after this point depending on the sample type and the analytical method applied (examples for UPLC–MS can be observed in [36,37]). The authors recommend multiple injections (up to 150) of the same sample (this can be a QC sample) when developing a new analytical method or analyzing a new sample type to ascertain the appropriate number of injections before the degradation in instrument performance is unacceptable. An example for a single metabolite, tryptophan, present in plasma and analyzed applying UPLC–MS is shown in Figure 1. Without routine maintenance (as described below) the variability introduced into the data during sample analyses will become significant and will equal or be greater than inherent biological variability in the dataset. In NMR spectroscopy, where samples are introduced into the spectrometer in sealed glass tubes, there is no sample–instrument interaction resulting in the ability to analyze large sample numbers in a single analytical experiment without a requirement for routine instrument maintenance. However, degradation of sample extracts present in an autosampler for many days has to be considered in an appropriate experimental design for studies applying NMR spectroscopy.

Due to the build-up of sample components and low-level instability in electronic systems, routine maintenance of chromatography and MS platforms is performed to return the instrument to optimal or near-optimal performance. For GC–MS, this will involve replacing injector liners and gold seals, removal of the top 5–20 cm of the column to eliminate contamination and tuning/calibrating the MS. For UPLC–MS this involves cleaning the UPLC column and electrospray source components to remove contamination and tuning/calibrating the MS.

Applying multiple analytical experiments of an appropriate number of injections and routine instrument maintenance ensures that the within-experiment variability introduced is small compared with the biological variability in the samples being studied. In the following sections, we will discuss the role of QC samples in determining the precision for each metabolite/metabolic feature detected. However, routine maintenance does not necessarily return the instrument to the same level of operation as observed at the start of the previous analytical experiments and that would allow simple integration of data from different experiments. Step changes in response can be observed between analytical experiments performed on alternate days. Larger step-changes can be observed after annual maintenance of instruments. This introduces significant levels of analytical variation into the dataset and has a large impact on the quality of the data. However, the recent integration of data from QC samples and the development of innovative data preprocessing algorithms have enabled, for the first time, the robust integration of data from multiple analytical experiments. This will be discussed below.

■ Randomization of sample preparation & analysis

In human studies, sample collection is not randomized in relation to many different sources of variability in the human population, including age and gender as two examples. Samples are collected as the subjects attend a collection site and this is not generally controlled. However, it is recommended that investigators subsequently randomize both sample preparation and sample analysis orders in small- and large-scale

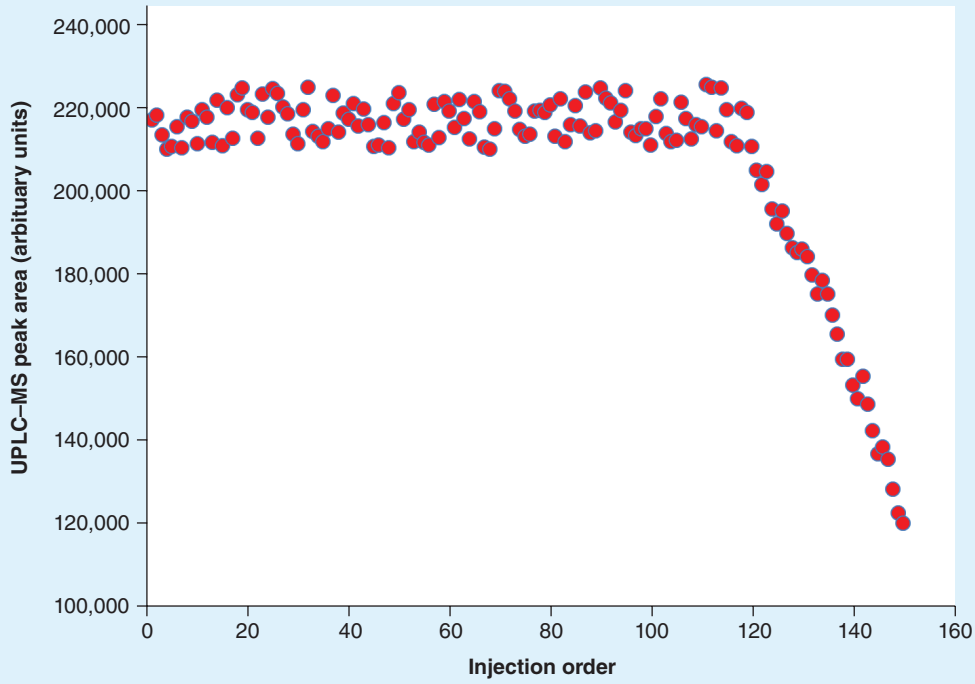


Figure 1. A plot showing the UPLC–MS peak area of tryptophan for the replicate injection of a single QC sample pooled from metabolic footprint samples acquired from the culture of placental tissue. A Waters Acquity UPLC™ system coupled to a Thermo Scientific LTQ Orbitrap Velos™ MS was applied to acquire the data using a previously defined method [23]. The data representing the first ten equilibration injections have been removed from these data. The data show the low level of variance present for the first 120 samples followed by a significant degradation in the stability of the peak following 120 samples.

studies [6]; samples should be randomized and blinded to the analyst before chemical analysis to ensure that no bias is introduced at this stage by the analyst and, therefore, reduce the chance of time-of-analysis becoming a confounding factor. Randomization is performed to ensure that there is no correlation between demographic/lifestyle/physiological factors (e.g., gender, ethnicity, age and BMI) and preparation or analysis order. Any correlation could introduce bias into the study. Appropriate randomization of samples ensures that no bias is introduced as part of sample preparation and data acquisition. For example, randomization should ensure that an even distribution of male and female subject samples across the analytical experiment is present. If all or a large proportion of the male subject samples were analyzed first, then changes in MS response over time will impact differently on the data acquired for the subpopulations of female and male subjects. This highlights the importance of collecting meta-data related to demographic/lifestyle/physiological factors during sample collection to allow the assessment of bias, but also that data related to confounding factors can be incorporated into the experimental design and statistical analysis.

In large-scale studies involving multiple analytical experiments, valid randomization within and between analytical experiments is essential. The distribution of subjects across the single biological study should be closely represented in each analytical experiment, where possible, to ensure that bias is not introduced. For example, if the male:female ratio in the biological study is 60:40 then this same ratio should be present, to an approximation, in each analytical experiment. In studies involving thousands of samples, data acquisition may be required to commence before sample collection is completed. Here it is recommended that randomization is performed and that the distribution of subject samples in each analytical experiment matches that of the currently collected study samples as closely as possible. In larger studies, with sample and data

collection performed over months or years, then constraints can be applied. For example, a maximum period of time between sample collection and data preparation or data acquisition can be applied to ensure sample instability is not a concern in studies performed over many years.

It is important to note that in any human metabolomic study, either small or large, it is appropriate to statistically assess the proposed experiment design for confounding factors before proceeding with its execution. Often simple randomization is not sufficient to avoid imbalance within, and between, analytical batches. This is particularly true for low-prevalence outcomes where the proportion of 'case' samples is small. It may be appropriate to perform stratified proportional randomization. That is, separate the subjects into groups based on outcome, and then randomly pick samples from each group in proportion to the number of members in each group. For example, for an unbalanced case–control study with four controls for each case, the sampling protocol will randomly select (without replacement) four subjects from the control group followed by one subject from the case group. This process is then repeated until all subjects have been selected and, therefore, ordered. For studies with a population containing a heterogeneous mixture of demographic/lifestyle/physiological factors, it may be necessary to repeat the selection and confounder assessment process multiple times, until a suitably balanced design is achieved. Alternatively, a nested stratified proportional randomization procedure can be used; for example, in a multicenter study it is recommended to ensure that both outcome and center are evenly stratified across the analytical experiment. In this case, the subjects are grouped into multiple outcome/center groups before proportional random selection.

■ QC samples & QA

In simple terms, QA is a process followed to statistically assess the performance of a process and to ensure the process meets predefined acceptance criteria. In targeted analytical studies applying MS platforms, accuracy (or recovery) and precision are calculated to determine the quality of the data and assess whether it meets the acceptance criteria defined before the experiment started. This is a regulatory body requirement in many industries, including pharmaceuticals (see the US FDA guidelines for an example [38]). QC samples are routinely applied in these targeted experiments to enable accuracy and precision to be monitored. However, the inclusion of QC samples into untargeted metabolomics studies applying MS platforms is currently limited. The authors highly recommend their inclusion in untargeted studies to enable QA processes to be implemented [23,39] as well as to provide capabilities to improve the quality of data in combination with data processing algorithms [23]. QC samples can also be applied to assess metabolite recovery during sample preparation procedures and to assess variability introduced by different analysts, although these will not be discussed further here; QC samples will be discussed in detail below.

What are QC samples?

QC samples, in the context of metabolomics, should be representative of the qualitative and quantitative composition of the subject samples being analyzed in the study; ideally they are an average of the composition of all samples studied. In analytical experiments a QC sample (of the same composition) or samples (each with a different composition) are analyzed intermittently. In untargeted studies, a single QC sample is analyzed and as the composition of each sample injected is equal, in theory, all data acquired for QC samples should be identical [26]. However, small levels of variation are introduced during the analytical process (e.g., injection volume or ion-transmission efficiency), which leads to minor and random variation being observed in the data obtained for the QC samples. These data can be applied in QA processes to ensure the data acquired are fit for purpose or meet predefined acceptance criteria.

In untargeted metabolomic studies, QC samples were, and still are, primarily applied to assess and ensure that the analytical processes being performed are appropriate and meet predefined acceptance criteria. In simple terms, QC samples are applied to ensure that the data acquired are of a suitable quality for data analysis to be performed and can produce valid and robust datasets and biological conclusions [23,27,39]. A single 'pooled' QC sample is applied for a complete biological study with aliquots of this sample processed through sample preparation, data acquisition and data preprocessing. The variation in the data acquired from different aliquots of the pooled QC sample reflects all process variations from the point of QC sample introduction; typically, sample preparation,

data acquisition and data preprocessing steps are assessed. Although the primary application of QC samples is to assess process variation, they are also applied to fulfill other roles in data acquisition and data preprocessing and these will be discussed below.

Preparation of pooled QC samples

A QC sample should qualitatively and quantitatively mimic the sample matrix and metabolite composition of the samples to be investigated in the scientific study. As the QC sample closely represents the study samples, any technical variation introduced during sample preparation, data acquisition and data preprocessing will be represented in the data acquired for the QC samples. It is important that the qualitative and quantitative composition is mimicked as closely as possible to allow representative measures of process variation to be assessed; in cases where the sample composition is not mimicked, then differences in the variation introduced by matrix components may not represent variation introduced into the data acquired for the study samples. One example is the presence or absence of phospholipids in UPLC–MS analysis of serum and plasma and the influence these have on sensitivity, specificity and process variation [40]. If the study samples do, and the QC sample do not, contain high concentrations of phospholipids, then variation introduced into the data acquired for study samples will not be represented in the data acquired for QC samples.

The preparation of a QC sample is dependent on the sample type, availability and the size of the scientific study. In all studies, preparation of a pooled QC sample applying small aliquots from each study sample is most appropriate. If not feasible, the purchase of matrix-matched samples that still mimic the qualitative and quantitative composition, but with a lower accuracy than a pooled QC sample, is recommended [23,35].

The majority of large-scale studies of the human population involve the study of biofluids, generally blood plasma, serum or urine. The preparation of a single pooled QC sample applying biofluid samples from a small study ($n < 100$) is relatively simple: small aliquots of each study sample are combined and thoroughly mixed to prepare a pooled QC sample [23,26,27]. Aliquots of this pooled sample can be applied to prepare multiple QC samples to be analyzed in the analytical experiment. Identical SOPs are applied for the preparation of study and QC samples. For example, for the study of serum collected from 100 subjects, pooling of 50 µl aliquots of serum from each subject will be sufficient to prepare a single pooled QC sample appropriate for a single analytical experiment. For small studies, this is easy to perform as all study and QC samples will be prepared on the same day, with QC samples prepared after study samples. This ensures that study and QC samples are passed through a single freeze–thaw cycle.

For studies involving hundreds to thousands of study samples, sample preparation cannot be completed in a single day, but will be separated across multiple days. In these cases, aliquots from each study sample should be collected for each day to prepare a pooled QC sample and then be stored frozen. A separate pooled sample will be collected on each day. Once all study samples are prepared then different QC aliquots collected on each separate day can be pooled and thoroughly mixed to prepare a single pooled QC sample. Although appropriate, this process does provide a discrepancy in the number of freeze–thaw cycles: one cycle for study samples and two cycles for the pooled QC sample. This, although unproven, may provide a separate source of bias between study and QC samples.

In large studies involving thousands of subjects, with samples to be collected and analyses to be performed over many months or years, the preparation of a pooled QC sample for the complete study becomes difficult. In these studies it is common for analyses to start before all the samples have been collected. Therefore, the preparation of a pooled QC sample from all subject samples is not feasible. One solution is to collect aliquots from the first 'n' subjects, pool and thoroughly mix to provide a pooled QC sample that can be used for the complete study. This solution assumes the distribution of subjects in the first **n samples is representative** of the complete sample population. This is acceptable when the number of subject samples applied to construct the pooled QC sample is large ($n > 500$) and the first subjects recruited are randomized and representative of the complete scientific study. An alternative solution is to purchase a commercially available biofluid sample and apply as the pooled QC sample for the complete study. For example, human serum can be purchased from commercial suppliers [23,36]. If this solution is applied, the authors suggest that the large volume pooled QC sample is subaliquoted into suitable fractions at the

start of the study and stored frozen. Sample preparation for each aliquot is then performed at the same time as study sample preparation for a specific analytical experiment. This is preferable to ensure that the pooled QC is passed through two and not many freeze–thaw cycles, and the length of storage of QC samples before extraction is similar to the length of storage of study samples before extraction (i.e., to eliminate the possibility for QC samples to be stored before extraction for less than 2 months, while study samples are stored before extraction for longer than 12 months).

For all of the processes described above the pooled QC sample is processed through the steps of sample preparation, data acquisition and data preprocessing. The variation determined in the QC data is a summed representative of variation introduced by all of these processes; it is difficult to isolate one source of variation from another by applying these QC samples. Also, the processes are only achievable where sufficient supplemental volumes of biofluids are available and can be applied to prepare a pooled QC sample. In studies investigating samples with limited availability (including bile, tears and interstitial fluid as examples), the preparation of a pooled QC sample is extremely difficult and their purchase from commercial suppliers is limited. In these cases the application of a synthetic QC sample or a separate biofluid that mimics its composition is required. For example, for interstitial fluid, blood serum or plasma can be applied as a pooled QC sample as the composition of blood serum/plasma and interstitial fluid are similar.

Although the majority of human-focused studies utilize biofluid samples, there are a minority of typically small studies that assess cell culture or tissue samples. Here, the preparation of a pooled QC sample as applied for biofluids is difficult and a number of options are available. For the investigation of mammalian tissues, 'waste' tissue supplementary to that required for the study, but of similar composition and collected ethically, can be applied for QC sample preparation. Alternatively, commercial sourcing of the same tissue type can be undertaken (e.g., The National Human Tissue Resource Center, PA, USA provides the ethical sourcing of a multitude of human tissues [101]). In extreme cases, where no other options are feasible, the same tissue type from a different species can be applied (e.g., for the study of human liver tissue, animal liver tissue from an appropriate source can be purchased). Caution should always be taken when applying this option because of interspecies differences in tissues and this should always be the last resort. For the investigation of the intracellular metabolome of cells cultured in the laboratory, supplementary samples can be cultured to provide biomass to prepare a pooled QC sample.

These studies applying cells or tissues will generally provide a pooled QC sample after sample preparation and not before; the variation observed in QC data is representative of data acquisition and data preprocessing only, and not sample preparation as well. To determine variability due to sample preparation, the use of technical replicates is probably the best approach. It is technically difficult to pool large masses of tissues and cells, and then subaliquot, without significant damage to or rupture of cell membranes, and the leakage of intracellular metabolites. Therefore, different aliquots of tissue are extracted separately and then each extract is combined to prepare the pooled QC sample. An alternative is to take small aliquots from each of the study sample extracts, combine and thoroughly mix to prepare a pooled QC sample. There are advantages and limitations to each of these options based on the closeness of QC sample composition to study samples and the processes for which technical variability will be determined.

■ The emergence of QC samples in untargeted studies

The incorporation of QC samples into untargeted metabolomic studies of mammals, including the human population, has only recently emerged. The first applications were reported in 2006 from Sangster and colleagues at AstraZeneca in the UK [26] and in 2007 by van der Greef and co-workers in The Netherlands [41]. Although the incorporation of QC samples provides many advantages, as discussed below, surprisingly only a limited number of metabolomic research groups worldwide include QC samples in their studies. These investigations include the study of urine [27,42,43], blood serum and plasma [3,14,36], cell cultures [44] and tissue extracts [45]. As a community, metabolomics researchers need to apply QC samples in their studies and report quantitative measures of precision in their research publications to provide increased confidence in the application of metabolomics by the scientific community.

Why are QC samples applied in untargeted metabolomics studies?

The preparation of QC samples and their incorporation into untargeted metabolomics studies is not a simple process; it is time-consuming and technically demanding. Up to 35% of all injections in an analytical experiment can be QC sample-related, which defines the importance researchers place on their inclusion because of the reduction in subject-sample throughput. So why do we apply QC samples in untargeted metabolomics studies?

■ Equilibration of analytical platforms after routine maintenance

Routine maintenance of chromatographic and MS platforms is required in small- or large-scale studies as continued operation for the analysis of biological samples will inevitably lead to a significant degradation in performance. During the maintenance process, removal of sample residues is performed and this results in reactivation of active sites by removal of absorbed material. Following this routine maintenance, significant variation (of the same level as biological variability) in measured parameters – including response and retention time – can be observed for the first three to ten injections [36,43]. This is a result of components of the sample being absorbed and blocking active sites; following this period of inactivation the performance of the platform 'equilibrates' and analytical variation is observed to be lower than that from biological sources.

The level of drift observed before equilibration has occurred is not acceptable, and in experiments where study samples are analyzed from the first injection onwards, will result in the loss of reproducible data for important study samples that are analyzed in the 'equilibration' period. As a compromise, QC samples can be repeatedly injected prior to the commencement of the analytical run in order to 'condition' or 'equilibrate' the system; these are nonprecious samples for which data acquired can be removed from the dataset before data preprocessing is performed and, therefore, do not impact on the data quality [23]. Generally, five to ten QC injections are performed at the start of each experiment, depending primarily on the analytical platform, but also on the sample type.

■ Correction of small levels of variation within & between analytical experiments

Although every effort is made during the analytical experiment to eliminate sources of variation, these are always present and are represented in the variation observed in acquired QC-response data. Increases and decreases in the measured response of a compositionally identical sample injected multiple times are observed and these changes can be different for diverse metabolites/metabolic features.

As the logical assumption is that the data acquired for all QC samples should be identical, then data preprocessing algorithms can be applied to reduce the analytical variability observed while maintaining the biological variability inherent in the study samples. The first algorithm developed was QC-based robust locally estimated scatterplot smoothing (LOESS) signal correction (QC-RLSC), which was developed for application with blood serum samples as part of the Human Serum Metabolome in Health and Disease project [23]. The QC-RLSC algorithm has since been validated for a number of different sample types, including urine, intracellular metabolomes, cell media (metabolic footprint) and tissue extracts [Dunn WB, Unpublished Data]. QC-RLSC normalizes the response data for subject samples to the response data for QC samples. Simply, a low-order nonlinear LOESS is fitted to the QC data in respect to analysis order, followed by interpolation of the correction curve to the study samples [23]. Other methods have also been developed [42,46–49] or are being developed [Broadhurst D, Unpublished Data].

Internal standards spiked into samples is another method that can be applied to reduce between-sample variability, in particular when an isotopic analogue of the metabolite can be applied (e.g., $^{13}C_6$-glucose for glucose). This is routinely observed in targeted studies. However, it is experimentally difficult to apply an internal standard for each of hundreds or thousands of metabolites detected in untargeted studies because of the cost of purchasing isotopic analogues for each metabolite and the limited qualitative knowledge related to metabolome composition before sample preparation and data acquisition. A single internal standard can be applied to correct analytical variation for a group of metabolites that are either related (e.g., all are present in the same class of metabolites) or unrelated [50]. However, we believe that the use of QC samples to compensate for analytical variation in untargeted studies is most appropriate.

■ Quantitative measurement of technical reproducibility

In targeted analytical methods, QC samples are applied to determine the accuracy and precision

of the analytical method for each metabolite assayed; the exact quantitative composition of the QC sample(s) is known to allow the determination of accuracy (otherwise known as recovery) by comparison to calibration curve data. In untargeted studies, the quantitative composition is not known and no calibration curve data are acquired. Therefore, only analytical precision can be determined through the incorporation of QC samples in metabolomic analyses.

QC samples are analyzed intermittently throughout the analytical experiment after the initial 'equilibration' phase. As these are theoretically the identical sample analyzed multiple times, the data acquired can be used to determine the within-experiment precision (where one analytical experiment has been performed) or the within-study precision (where data from multiple analytical experiments have been integrated). Following signal correction, calculation of the relative SD (RSD) for each separate metabolite/metabolic feature and percentage of QC samples in which the metabolite/metabolic feature was detected is applied for QA.

The percentage detection rate defines whether the metabolite is consistently detected; if not, then this metabolite/feature should be removed from the dataset. The authors typically apply an acceptance criteria of 50%; if the metabolite is detected in fewer than five out of ten QC samples, then the data for that metabolite/metabolic feature is removed [23]. The RSD provides a univariate measure of intra-experiment precision for each metabolite/metabolic feature separately and allows data to be compared with acceptance criteria; specifically, whether that metabolite passes the acceptance criteria. In targeted studies applied in the pharmaceutical industry, the FDA specify the following (RSD is equivalent to coefficient of variation [CV]) [38]: *"The precision of an analytical method describes the closeness of individual measures of an analyte when the procedure is applied repeatedly to multiple aliquots of a single homogeneous volume of biological matrix. Precision should be measured using a minimum of five determinations per concentration. The precision determined at each concentration level should not exceed 15% of the CV except for the LLOQ, where it should not exceed 20% of the CV."*

For biomarkers, the FDA guidelines are slightly more relaxed (20 and 30%, respectively) and these may be more appropriate for metabolomic studies [51]. However, clearly the less variable the QC data are for a particular metabolite the greater the confidence that the analyst can have in it. In untargeted metabolomics studies, hundreds to thousands of metabolites are detected with many present at low concentrations. Therefore, and to take into account these differences between targeted and untargeted methods, the authors apply the following analytical, platform-dependent acceptance criteria: RSD < 20% for UPLC–MS and RSD < 30% for GC–MS [23,36,50]. The higher acceptance limit for GC–MS reflects the greater number of processing steps for GC–MS studies (e.g., chemical derivatization) and the lower injection volume and reproducibility of GC injectors that creates a greater between-sample variation. The authors apply these criteria for all sample types as, in our hands, a number of unpublished studies have shown similar levels of reproducibility for multiple sample types, including urine, cell media (metabolic footprint), cell extracts and tissue extracts. This process provides a univariate filtering step to remove nonreproducible metabolic features before data analysis; only data for nonreproducible metabolites are removed. In the authors experience, 10–30% of all detected metabolic features are removed from the dataset when applying this QA process, dependent on analytical platform, sample type and temporal changes in instrument or researcher [3,45]. The authors recommend that the median and quartile ranges for the RSD for all metabolic features are reported in the publication of results.

The univariate process described above requires that significant data processing and a rapid QA process is not performed following data acquisition. To provide a simple and quick qualitative determination of the data quality for a single analytical experiment, multivariate principal-components analysis can be performed [26]. Principal-components analysis score plots provide a representation of the variability within the data acquired; two data points close together are similar (there is minimal variation between the data acquired for them) and two data points far apart are less similar (there is a larger level of variation between the data acquired for the two samples). The variation between QC samples (the same sample injected multiple times) should be lower than the variation between subject samples (different biological samples). Clustering of QC sample data points should be more compact than the distribution of biological samples (assuming that there is indeed some biological variability). An example is shown in FIGURE 2 for a clinical study composed of two sample classes (cases and controls). If the distribution

of data points for QC and biological samples are similar then the data must be treated with caution. In most applications, the distribution of QC samples is less than for biological samples; however, in cases where the intraclass variation for two or more classes is small, then the variation observed for QC and subject samples will be similar. When similar levels of variation are observed, but are not expected (e.g., the researcher expects large intraclass variation), reanalysis of the samples or their repreparation and analysis is recommended. This provides a rapid check of the quality of the data on the day of analysis.

■ To provide integration of data from multiple analytical experiments

As discussed above, step-based changes in response can be observed between analytical experiments as a result of routine or annual maintenance. This is a significant hurdle to overcome in large-scale studies where the integration of data from multiple analytical experiments is essential to provide statistical robustness in comparison with applying data from single analytical experiments.

In studies where the same pooled QC sample is applied for all analytical experiments, then the data acquired for QC samples can be applied simply to integrate data from multiple analytical batches into a single dataset. The protocols applied have been described previously [23].

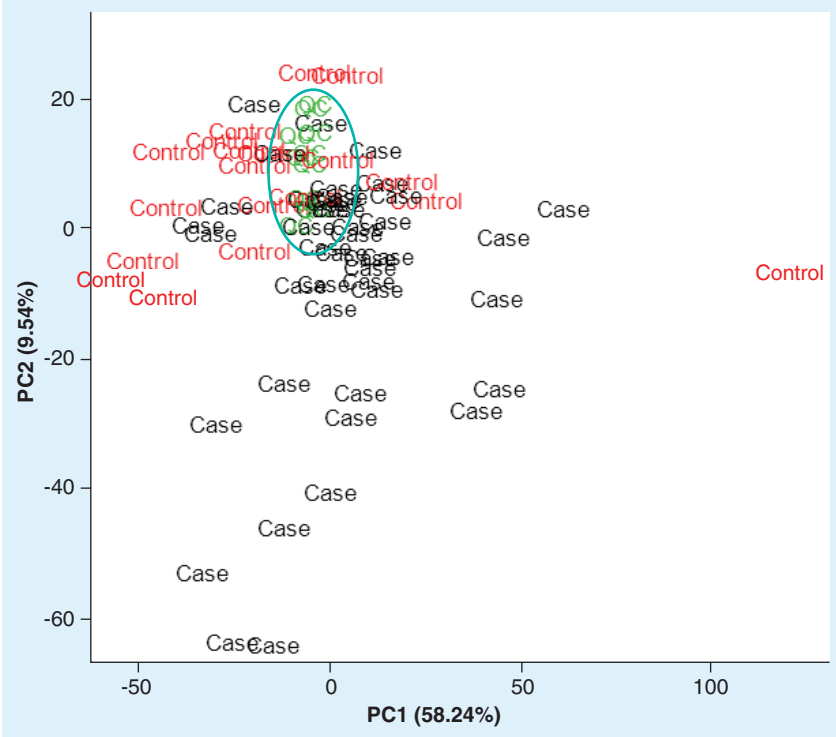


Figure 2. A principal-components analysis scores plot showing an example of the expected variation for QC and study (case and control) samples. The tighter clustering of QC samples (the distribution is defined by the oval) compared with the study samples describes that the biological variation related to the study samples is greater than the technical variation observed for the replicate injection of the same pooled QC sample. Data were collected for a biomarker study of serum applying UPLC–MS. A Waters Acquity UPLC™ system coupled to a Thermo Scientific LTQ Orbitrap Velos™ MS was applied to acquire the data using a previously defined method [23].

QC sample incorporation into analytical experiments

Thus far, we have discussed why it is important to include QC samples in metabolomic studies and how their inclusion can be applied to improve and assess the quality of the data. Now we will describe how QC samples are incorporated into the analytical experiment.

A typical run order applied by the authors is shown in Figure 3 for GC–MS and UPLC–MS. The first nth injections (n = 5 for GC–MS and n = 10 for UPLC–MS) are applied to 'equilibrate' the analytical system following routine maintenance. Data for injections one to three and one to eight are removed from the dataset prior to data processing and data analysis for GC–MS and UPLC–MS, respectively. QC samples are then intermittently injected through the analytical experiment with two QC samples at the end of the analytical experiment. Two, and not one, QC injections are performed at the end of the experiment to eliminate the impact on signal correction if there is one sample-injection or instrument failure; the absence of a QC sample at the end of the experiment significantly impacts on the QC-RLSC algorithm applied.

The frequency at which QC samples are injected is dependent on the data preprocessing steps to be performed. Where QC samples are to be applied to quantify precision within a single analytical experiment only, then analysis of a QC sample every tenth injection is commonly applied and is appropriate [27]. Where QC samples are to be applied for signal correction and to quantify precision, more frequent injections are required, typically every third to seventh injection [23]. The greater frequency is required so as to allow the acquisition of sufficient QC data to ensure the QC-RLSC algorithm operates efficiently and robustly.

Conclusion & future perspective

The variability observed in the environment and human genotype has a significant impact

on the phenotype of humans. In laboratory-based scientific studies, this variation is controlled to minimize its impact on the biological variation observed. In these studies, the sample size is small per class (generally fewer than 20). In the study of human populations outside the laboratory, control of the genotype and environment is minimal. For this reason, the sample size required to acquire statistically valid results is much higher; hundreds or thousands of subjects are required. This provides significant hurdles in large-scale untargeted metabolomic studies of the human population that apply MS platforms.

In this paper, we have described recent innovative advances in experimental design, the inclusion of QC samples and data preprocessing algorithms that have provided the ability to perform these large-scale untargeted MS-focused studies and provide robust data for data analysis and biological interpretation. These methods have been developed as part of the Human Serum Metabolome in Health and Disease project, a study to define the normal serum metabolome of healthy individuals [102]. To date, the project has acquired data on over 3000 different subjects and discussions on data for 1200 subjects have been presented recently [23]. More than 4400 metabolic features were detected applying UPLC–MS, which we estimate to be between 1000 and 2000 metabolites, although current limitations in untargeted metabolomics do not allow accurate identification of all metabolites (see [22] for a review on the limitations of metabolite identification in untargeted metabolomics). 157 metabolites were detected applying GC–MS. Without the innovations described in this paper, this project, and many other projects undertaken by the authors, would not have been successful or would not have provided robust data for biological interpretation. Before these innovations, data for large-scale untargeted discovery studies could only be acquired using NMR spectroscopy [19,25].

All of the processes described would not be possible without the use of QC samples. Signal correction opens up for the first time the ability to acquire and apply reproducible MS data collected in multiple analytical experiments in untargeted metabolomics studies for large-scale studies of the human population. The authors hope that the readers of this article will endeavor to incorporate the experimental design, QC samples and signal correction processes described here in their future metabolomic studies to increase the impact of metabolomics through the assurance of the quality of the data to the scientific community; whether single or multiple analytical experiments are applied as part of a biological study. Only with the community-wide acceptance of QC samples and QA procedures can the comparison of different datasets from across the research-based community be performed. For the first time, this would permit the analysis of metabolomics data on a global scale and enable studies of epidemiological proportions of the human race.

GC–MS

Injection Order	Sample Type
1	QC
2	QC
3	QC
4	QC
5	QC
6	BLANK
7	SUBJECT SAMPLE
8	SUBJECT SAMPLE
9	SUBJECT SAMPLE
10	SUBJECT SAMPLE
11	SUBJECT SAMPLE
12	QC
13	SUBJECT SAMPLE
14	SUBJECT SAMPLE
15	SUBJECT SAMPLE
16	SUBJECT SAMPLE
17	SUBJECT SAMPLE
18	QC
19	SUBJECT SAMPLE
.....	.....
.....	.....
40	SUBJECT SAMPLE
41	SUBJECT SAMPLE
42	QC
43	QC

UPLC–MS

Injection Order	Sample Type
1	QC
2	QC
3	QC
4	QC
5	QC
6	QC
7	QC
8	QC
9	QC
10	QC
11	BLANK
12	SUBJECT SAMPLE
13	SUBJECT SAMPLE
14	SUBJECT SAMPLE
15	SUBJECT SAMPLE
16	SUBJECT SAMPLE
17	QC
18	SUBJECT SAMPLE
19	SUBJECT SAMPLE
20	SUBJECT SAMPLE
21	SUBJECT SAMPLE
22	SUBJECT SAMPLE
23	QC
24	SUBJECT SAMPLE
.....	.....
.....	.....
87	SUBJECT SAMPLE
88	SUBJECT SAMPLE
89	QC
90	QC

Figure 3. Examples of typical injection orders for GC–MS and UPLC–MS.

Financial & competing interests disclosure

This work was supported by the Manchester NIH Research Biomedical Research Centre. The Human Serum Metabolome in Health and Disease project was supported through funding from the BBSRC, MRC, GlaxoSmithKline and AstraZeneca (BBC0082191). The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

Executive summary

Large-scale studies of the human population

- The study of human biofluid metabolomes provides a phenotypic and dynamic picture of intra- and inter-individual variability in the human population.
- Large sample sizes are required to statistically define the inherent variability in the human population.

Experimental design in large-scale studies of the human population

- The determination of global (or untargeted) metabolic profiles of the type required for metabolomic studies, particularly large-scale, require long-term investigations involving hundreds to thousands of samples and requires very careful control if analytical variability is to be monitored and controlled.
- Many experimental considerations are required in designing a large-scale study of the human population.
- These include reproducibility of sample collection across single or multiple sites, randomization during sample preparation and data acquisition, the design of multiple analytical experiments and QC/quality assurance processes.

QC samples in metabolomics

- There are numerous sources of analytical variability in MS-based analytical methods (within and between run), such as changes in analyte response and retention time, which can result in poor data quality.
- A pragmatic approach to monitoring data quality is the use of QC samples prepared from the sample being profiled and analyzed at a regular interval throughout the analysis.
- QC samples can be applied to condition chromatography and MS instruments following maintenance, in order to correct for small levels of variation, to quantitatively measure technical reproducibility and to integrate data from different analytical experiments.
- Data from the QC samples can be used to reject analytical batches where the variability is too high.

References

Papers of special note have been highlighted as:
▪ of interest
▪▪ of considerable interest

1 Shastry BS. SNPs: impact on gene function and phenotype. *Methods Mol. Biol.* 578, 3–22 (2009).

2 Slomko H, Heo HJ, Einstein FH. Minireview: epigenetics of obesity and diabetes in humans. *Endocrinology* 153(3), 1025–1030 (2012).

3 Ugarte M, Brown M, Hollywood KA, Cooper GJ, Bishop PN, Dunn WB. Metabolomic analysis of rat serum in streptozotocin-induced diabetes and after treatment with oral triethylenetetramine (TETA). *Genome Med.* 4(4), 35 (2012).

4 Lewis GD, Farrell L, Wood MJ *et al.* Metabolic signatures of exercise in human plasma. *Sci. Transl. Med.* 2(33), 33–37 (2010).

5 Gavaghan CL, Holmes E, Lenz E, Wilson ID, Nicholson JK. An NMR-based metabonomic approach to investigate the biochemical consequences of genetic strain differences: application to the C57BL10J and Alpk:ApfCD mouse. *FEBS Lett.* 484(3), 169–174 (2000).

6 Dunn WB, Broadhurst DI, Atherton HJ, Goodacre R, Griffin JL. Systems level studies of mammalian metabolomes: the roles of mass spectrometry and nuclear magnetic resonance spectroscopy. *Chem. Soc. Rev.* 40(1), 387–426 (2011).

7 Wishart DS, Knox C, Guo AC *et al.* HMDB: a knowledgebase for the human metabolome. *Nucleic Acids Res.* 37(Database issue), D603–D610 (2009).

8 Goodacre R, Vaidyanathan S, Dunn WB, Harrigan GG, Kell DB. Metabolomics by numbers: acquiring and understanding global metabolite data. *Trends Biotechnol.* 22(5), 245–252 (2004).

9 Bondia-Pons I, Nordlund E, Mattila I *et al.* Postprandial differences in the plasma metabolome of healthy Finnish subjects after intake of a sourdough fermented endosperm rye bread versus white wheat bread. *Nutr. J.* 10, 116 (2011).

10 Pechlivanis A, Kostidis S, Saraslanidis P *et al.* (1)H NMR-based metabonomic investigation of the effect of two different exercise sessions on the metabolic fingerprint of human urine. *J. Proteome Res.* 9(12), 6405–6416 (2010).

11 Lloyd AJ, Beckmann M, Fave G, Mathers JC, Draper J. Proline betaine and its biotransformation products in fasting urine samples are potential biomarkers of habitual citrus fruit consumption. *Br. J. Nutr.* 106(6), 812–824 (2011).

12 Oresic M, Simell S, Sysi-Aho M *et al.* Dysregulation of lipid and amino acid metabolism precedes islet autoimmunity in children who later progress to Type 1 diabetes. *J. Exp. Med.* 205(13), 2975–2984 (2008).

13 Wang TJ, Larson MG, Vasan RS *et al.* Metabolite profiles and the risk of developing diabetes. *Nat. Med.* 17(4), 448–453 (2011).

14 Kenny, LC, Broadhurst DI, Dunn W *et al.* Robust early pregnancy prediction of later preeclampsia using metabolomic biomarkers. *Hypertension* 56(4), 741–749 (2010).

15 Horgan RP, Broadhurst DI, Walsh SK *et al.* Metabolic profiling uncovers a phenotypic signature of small for gestational age in early pregnancy. *J. Proteome Res.* 10(8), 3660–3673 (2011).

16 Vermeer LS, Fruhwirth GO, Pandya P, Ng T, Mason AJ. NMR metabolomics of MTLn3E breast cancer cells identifies a role for CXCR4 in lipid and choline regulation. *J. Proteome Res.* 11(5), 2996–3003 (2012).

17 Catchpole G, Platzer A, Weikert C *et al.* Metabolic profiling reveals key metabolic features of renal cell carcinoma. *J. Cell Mol. Med.* 15(1), 109–118 (2011).

18 Brown MV, McDunn JE, Gunst PR *et al.* Cancer detection and biopsy classification using concurrent histopathological and metabolomic analysis of core biopsies. *Genome Med.* 4(4), 33 (2012).

19 Loo RL, Chan Q, Brown IJ *et al.* A comparison of self-reported analgesic

use and detection of urinary ibuprofen and acetaminophen metabolites by means of metabonomics: the INTERMAP Study. *Am. J. Epidemiol.* 175(4), 348–358 (2012).

20 Harder U, Koletzko B, Peissner W. Quantification of 22 plasma amino acids combining derivatization and ion-pair LC–MS/MS. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.* 879(7–8), 495–504 (2011).

21 Floegel A, Drogan D, Wang-Sattler R *et al.* Reliability of serum metabolite concentrations over a 4-month period using a targeted metabolomic approach. *PLoS ONE* 6(6), e21103 (2011).

▪ **One of the first papers assessing the between- and within-person variation observed in serum metabolite concentrations of humans over a 4-month period.**

22 Dunn WB, Erban A, Weber RJM *et al.* Mass appeal: metabolite identification in mass spectrometry-focused untargeted metabolomics. *Metabolomics* doi:10.1007/s11306–012–0434–0434 (2012) (Epub ahead of print).

23 Dunn WB, Broadhurst D, Begley P *et al.* Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. *Nat. Protoc.* 6(7), 1060–1083 (2011).

▪▪ **Describes all technical aspects related to sample preparation, data acquisition and data preprocessing for the investigation of serum or plasma in relation to large-scale studies of the human population.**

24 Wu H, Xue R, Dong L *et al.* Metabolomic profiling of human urine in hepatocellular carcinoma patients using gas chromatography/mass spectrometry. *Anal. Chim. Acta* 648(1), 98–104 (2009).

25 Lindon JC, Keun HC, Ebbels TM, Pearce JM, Holmes E, Nicholson JK. The Consortium for Metabonomic Toxicology (COMET): aims, activities and achievements. *Pharmacogenomics* 6(7), 691–699 (2005).

▪ **First paper describing the application of NMR spectroscopy to the study of large sample sets, specifically rodent urine and serum in relation to xenobiotic toxicity.**

26 Sangster T, Major H, Plumb R, Wilson AJ, Wilson ID. A pragmatic and readily implemented quality control strategy for HPLC–MS and GC–MS-based metabonomic analysis. *Analyst* 131(10), 1075–1078 (2006).

▪ **One of first papers discussing the pragmatic use of QC samples in untargeted metabolomics.**

27 Want EJ, Wilson ID, Gika H *et al.* Global metabolic profiling procedures for urine using UPLC–MS. *Nat. Protoc.* 5(6), 1005–1018 (2010).

▪▪ **Describes all technical aspects related to sample preparation, data acquisition and data preprocessing for the investigation of urine in relation to large-scale studies of the human population.**

28 Feng Z, Prentice R, Srivastava S. Research issues and strategies for genomic and proteomic biomarker discovery and validation: a statistical perspective. *Pharmacogenomics* 5(6), 709–719 (2004).

29 Broadhurst DI, Kell DB. Statistical strategies for avoiding false discoveries in metabolomics and related experiments. *Metabolomics* 2(4), 171–196 (2006).

▪▪ **Excellent review discussing the significant issues to consider when performing statistical analysis of untargeted metabolomic datasets.**

30 Westerhuis JA, Hoefsloot HCJ, Smit S *et al.* Assessment of PLSDA cross validation. *Metabolomics* 4(1), 81–89 (2008).

31 Arkin CF, Wachtel MS. How many patients are necessary to assess test performance? *JAMA* 263(2), 275–278 (1990).

32 Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology (3rd Edition)*. Lippincott Williams & Wilkins, PA, USA (2008).

33 Teahan O, Gamble S, Holmes E *et al.* Impact of analytical bias in metabonomic studies of human blood serum and plasma. *Anal. Chem.* 78(13), 4307–4318 (2006).

34 Dunn WB, Broadhurst D, Ellis DI *et al.* A GC–TOF–MS study of the stability of serum and urine metabolomes during the UK Biobank sample collection and preparation protocols. *Int. J. Epidemiol.* 37(Suppl. 1), i23–i30 (2008).

35 Barton RH, Nicholson JK, Elliott P, Holmes E. High-throughput ^{1}H NMR-based metabolic analysis of human serum and urine for large-scale epidemiological studies: validation study. *Int. J. Epidemiol.* 37(Suppl. 1), i31–i40 (2008).

36 Zelena E, Dunn WB, Broadhurst D *et al.* Development of a robust and repeatable UPLC–MS method for the long-term metabolomic study of human serum. *Anal. Chem.* 81(4), 1357–1364 (2009).

37 Dunn WB, Broadhurst D, Brown M *et al.* Metabolic profiling of serum using ultra performance liquid chromatography and the LTQ-Orbitrap mass spectrometry system. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.* 871(2), 288–298 (2008).

38 US FDA. *Guidance for Industry, Bioanalytical Method Validation*. US FDA, MD, USA (2001).

39 Michopoulos F, Lai L, Gika H, Theodoridis G, Wilson I. UPLC–MS-based analysis of human plasma for metabonomics using solvent precipitation or solid phase extraction. *J. Proteome Res.* 8(4), 2114–2121 (2009).

40 Guo, X, Lankmayr E. Phospholipid-based matrix effects in LC–MS bioanalysis. *Bioanalysis* 3(4), 349–352 (2011).

41 van der Greef J, Martin S, Juhasz P *et al.* The art and practice of systems biology in medicine: mapping patterns of relationships. *J. Proteome Res.* 6(4), 1540–1559 (2007).

▪ **One of first papers discussing the application of QC samples in metabolomic studies.**

42 Veselkov KA, Vingara LK, Masson P *et al.* Optimized preprocessing of ultra-performance liquid chromatography/mass spectrometry urinary metabolic profiles for improved information recovery. *Anal. Chem.* 83(15), 5864–5872 (2011).

43 Gika HG, Theodoridis GA, Wingate JE, Wilson ID. Within-day reproducibility of an HPLC–MS-based method for metabonomic analysis: application to human urine. *J. Proteome Res.* 6(8), 3291–3303 (2007).

44 Paglia G, Hrafnsdottir S, Magnusdottir M *et al.* Monitoring metabolites consumption and secretion in cultured cells using ultra-performance liquid chromatography quadrupole-time of flight mass spectrometry (UPLC–Q-ToF-MS). *Anal. Bioanal. Chem.* 402(3), 1183–1198 (2012).

45 Dunn WB, Brown M, Worton SA *et al.* The metabolome of human placental tissue: investigation of first trimester tissue and changes related to preeclampsia in late pregnancy. *Metabolomics* 8(4), 579–597 (2012).

46 van der Kloet FM, Bobeldijk I, Verheij ER, Jellema RH. Analytical error reduction using single point calibration for accurate and precise metabolomic phenotyping. *J. Proteome Res.* 8(11), 5132–5141 (2009).

47 Kamleh MA, Ebbels TM, Spagou K, Masson P, Want EJ. Optimizing the use of quality control samples for signal drift correction in large-scale urine metabolic profiling studies. *Anal. Chem.* 84(6), 2670–2677 (2012).

48 Coulier L, Muilwijk B, Bijlsma S *et al.* Metabolite profiling of small cerebrospinal fluid sample volumes with gas

chromatography–mass spectrometry: application to a rat model of multiple sclerosis. *Metabolomics* doi:10.1007/s11306–012–0428–0422 (2012) (Epub ahead of print).

49 Xia J, Mandal, R, Sinelnikov IV, Broadhurst D, Wishart DS. MetaboAnalyst 2.0 – a comprehensive server for metabolomic data analysis. *Nucleic Acids Res.* 40(Web Server issue), W127–W133 (2012).

50 Begley P, Francis-McIntyre S, Dunn WB *et al.* Development and performance of a gas chromatography-time-of-flight mass spectrometry analysis for large-scale nontargeted metabolomic studies of human serum. *Anal. Chem.* 81(16), 7038–7046 (2009).

51 Visawanthan C, Bansal S, Booth B *et al.* Workshop/conference report – quantitative bioanlytical methods validation and implementation: best practices for chromatographic and ligand binding assay. *AAPS J.* 9, E3–E42 (2007).

■ Websites

101 National Human Tissue Resource Center. http://ndriresource.org

102 HUSERMET: Human Serum Metabolome in Health and Disease. www.husermet.org