Integration of genomic and metabonomic data in systems biology - are we 'there' yet?
Craig E Thomas¹ & Gopinath Ganji²

Address
¹Toxicology and Drug Disposition
Lilly Research Laboratories
A Division of Eli Lilly & Co
PO Box 708
Greenfield
IN 46140
USA
Email: thomas_craig_e@lilly.com

²Lilly Systems Biology
1 Science Park Road #04-01
The Capricorn
Science Park II
Singapore 117528
Email: gopi@lilly.com

Correspondence can be addressed to either author

Current Opinion in Drug Discovery & Development 2006 9(1):92-100
© The Thomson Corporation ISSN 1367-6733

The measurement of genes, proteins and metabolites has gained increasing acceptance as a means by which to study the response of an organism to stimuli, whether they are environmental, genetic, pharmacological, toxicological, etc. Typically referred to as genomics, proteomics, and metabonomics or metabolomics, respectively, these methods as independent entities have undoubtedly provided new biological insight that was not attainable a decade ago. Not surprisingly, scientists continue to push the boundaries to extract knowledge from data, and it is currently recognized that the full realization of these technologies is limited by a lack of tools to enable data integration. Integration of these 'omic datasets, or integromics, is desirable as it links the individual biological elements together to provide a more complete understanding of dynamic biological processes. Accordingly, in addition to developing new data analysis methods to extract further details from each of the high-content datasets individually, effort is also being expended to create or improve statistical methods, databases, annotations and pathway mapping to maximize our learning. There are several recent examples, in both mammalian and non-mammalian systems, in which genes, proteins and/or metabolites have been integrated using either biology- or data-driven strategies. Herein, key findings are reviewed, gaps in our current tools and technologies are identified and illustrated, and perspective is provided on the potential of integromics in biological research.

Keywords Genomics, metabonomics, microarray, pathways, proteomics, toxicology

Introduction
Harnessing the knowledge embedded within genomes has been touted as a key driver for the future discovery and development of new pharmaceutical therapeutics. Undoubtedly, a significant amount has been learned from the field of functional genomics, and targeted therapies in cancer, such as Gleevec®, Avastin®, Iressa®, Velcade® and others, are emerging [1]. It is readily apparent, however, that biology is not always ordered, nor is it linear. Genomics and proteomics allow for an understanding of what may happen, but cannot ascertain alone the full biological impact of gene expression and protein level alterations as they measure proximal responses to stimuli. It can be argued that one approach to understanding the impact, or consequences, of genomic and proteomic alterations in a biological system is to monitor metabolites, since they are one of the most distal elements in a dynamic process. The measurement of metabolite composition within a given system is termed metabonomics or metabolomics, and this field has seen a recent emergence in the scientific literature. As is highlighted in a review by Pognan, prior to 2001, the number of publications on metabolomics was miniscule, relative to genomics and proteomics [2••]. However, since 2002, while the total number of publications remains low, there has been a constant rise, reflecting both an increase in scientific curiosity and, most probably, advances in analytical methodology and data analysis tools. The primary areas of interest for metabonomic investigations are apparent in a review focused on metabolomics and systems biology by Kell, who used a text mining tool to demonstrate that publications through 2003, which had 'metabolome' or 'metabolomics' in their title, could be clustered based on their themes into: (i) technology development; (ii) integration with other 'omics'; or (iii) predicting higher order properties, such as disease [3••].

Over the past ten years, the study of gene expression, or functional genomics, has unleashed a tremendous amount of information across multiple disciplines encompassing both the plant and animal kingdoms. The impact of this has been felt in many areas, including plant genetics and breeding, pharmaceutical target identification, toxicology, clinical drug development, etc. The study of the proteome lags somewhat behind the study of the genome, largely due to technological challenges. Protein structure and function do not lend themselves to the high-throughput methods used to study gene expression on microarrays. The dynamic range for protein concentrations varies greatly, and separation and identification of proteins is difficult, tedious and time-consuming. Despite the fact that protein folding, protein-protein interaction and post-translational modifications (PTM; eg, phosphorylation) are all critical determinants in ultimate outcome, proteomics has been used extensively as a follow-up to or confirmation of gene expression studies. Developments in methods that do not require gel electrophoresis [4,5] suggest that the directed study of protein alterations will play an increasingly important role in many arenas; for example, there are now protein-based biomarkers that are used to predict certain types of cancer [6].

Ideally, a complete understanding of the response of a system to environmental, genetic or pharmacological stimuli would include a determination of alterations in gene expression, proteins (including PTM) and metabolites, coupled with more traditional biological data, such as clinical chemistry and histopathology (Figure 1). This
'omic data can lead to a more comprehensive understanding of dynamic processes activated in response to many environmental, biological and chemical stimuli. Measurement of changes in gene expression, proteins and metabolites can increase knowledge of dynamic biological processes. In drug development, for example, integrating such high-content datasets with currently accepted endpoints, such as clinical and morphological pathology, can provide added value in deducing mechanisms of toxicity and biomarker identification.

systems biology knowledge represents a noble goal, but one that is challenged by the complexity of biology. This review will not delve deeply into all of the available literature on genomics, proteomics or metabonomics as independent entities. Rather, this article will consider the current state of integration of the 'omic data streams, with emphasis placed on the development of methods for extracting knowledge from metabolite identification and quantitation, since the immaturity of this field represents one of the prime hurdles to systems biology applications. Recent articles describing efforts to model biological responses through 'omics will be discussed, new developments in data analysis and integration tools will be highlighted, and challenges yet to be overcome in systems biology will be touched upon.

**Correlating changes in gene expression and metabolites**

*Plant kingdom*

The maturity of genetics in plant research is evidenced, for example, by the sequencing of whole genomes for *Arabidopsis thaliana* and rice, and the widespread use of microarrays to study their gene expression [7•]. Similarly, the use of metabolites to define phenotypes has been widely employed in plant research for a number of years, primarily using mass spectrometry (MS) [8]. Not surprisingly then, these disciplines have been linked together in an attempt to define the function of gene products. Fridman and Pichersky nicely reviewed the steps necessary to correlate gene expression, proteins and metabolites with a particular focus on the elucidation of novel enzymatic pathways [9]. In particular, these authors review a number of published reports, the majority of which describe studies in which specific metabolites were directly related to enzymatic activity, and in some cases the metabolites were used to assist in the identification of genes contained within expressed sequence tag (EST) databases. Fridman and Pichersky conclude that, even with significant improvements in the measurement of genes and metabolites, identifying the enzymatic function of newly discovered proteins requires some *a priori* knowledge of related enzyme families, and that the metabolic profiling of substrates and products has yet to identify an enzyme unrelated to a known protein family.
A good example of comparing and correlating metabolite and gene profiles, to discriminate between different potato tuber systems, was published by Urbanczyk-Wochniak et al [7•]. The advantage of using the potato tuber system is that it has clearly defined developmental stages that allow for assessing the impact of transgenic alterations. Using a custom array, these investigators found that transcriptional variation during development obfuscated the ability to detect genetic perturbations. On the other hand, when metabolites were profiled using gas chromatography-mass spectrometry (GC-MS), the transgenic events clustered independently by principle components analysis (PCA). A comparison of the two indices was achieved using a pairwise correlation analysis. While expected correlations (eg, between sucrose and the sucrose transporter) were found, the majority of the correlations were novel findings (eg, between the genes associated with lysine correlating with vitamins C and E). Clearly, follow-up studies would be required to determine the biological validity of such statistical correlations.

Nuclear magnetic resonance (NMR) spectroscopy has been used to determine whether maize seeds from transgenic plants can be identified and classified solely on the basis of their metabolite profiles [10•]. NMR was chosen for this study, in part, because it minimized introduction of bias into the identified metabolite set, since no derivatization or chromatographic separation was required. Multivariate methods of PCA and partial least squares (PLS) discriminant analysis readily identified discriminant metabolites between transgenic samples and classified the various seeds.

Animal kingdom

Within the past one to two years, there has been an increased effort to use multiple 'omics datasets to study chemically induced toxicity. Coen et al investigated acetaminophen toxicity in mice, wherein liver tissue was analyzed for gene expression using microarrays, and plasma and tissue extracts were evaluated by high-resolution magic angle spinning (HR-MAS) NMR for metabolite changes [11]. A major challenge posed by these datasets was how best to compare or correlate them to provide the most meaningful biological context. Even when studying genes or metabolites individually, distinguishing pharmacological effects from toxicity is difficult, particularly when attempting to correlate endpoints since there may be varying ‘threshold levels’ that separate pharmacology from toxicity for genes and metabolites. In this study, acetaminophen was dosed at 50, 150 and 500 mg/kg to allow such a distinction of drug activity, and plasma and liver samples were obtained at multiple time points over 15 to 240 min post-dose. Gene changes were deemed significant if they met the specified 2-fold criteria in at least two of the five time points, and detailed annotation and information related to pathways and molecular function was gleaned using public and proprietary databases. NMR spectra indicated changes in lipid metabolism with a compensatory increase in glycogenolysis and glycolysis. Largely, the correlation between genes and metabolites in this study was limited to linking specific genes with defined metabolic events. For example, insulin-like growth factor (IGF) binding protein-1 was upregulated from 120 min onward. This protein is known to stabilize IGFs on IGF receptors, which is consistent with an attempt to minimize hepatic glycolysis and leverage lipid catabolism. Several other biochemically tenable links between metabolites and gene function were highlighted. Overall, while the data demonstrated that there is some additional confidence in the toxicity mechanism that was proposed by comparing genes and metabolites, the analysis appeared to be heavily reliant upon an empirical search for correlations after independently determining the most significantly changed genes or metabolites. Although valuable, this deductive approach is less likely to elucidate entirely novel correlations and highlights the currently limited options for developing meaningful genomic and metabolomic correlations. This concept will be further discussed in the section on biology-driven versus data-driven strategies for data integration.

An interesting study employing a ‘reverse functional genomic and metabolic approach’ to study orotic acid-induced fatty liver was reported by Griffin et al [12••]. An extremely detailed analysis of metabolic changes was conducted in hepatic tissue extracts, intact liver tissue and plasma using 1H-NMR, and in liver tissue extract using GC-MS. The gene and metabolite datasets were interrogated using multivariate statistics, and hierarchical clustering was used to parse and visualize the transcriptional data. Importantly, attempts were made using PLS to model time progression since a temporal shift in changes at the gene and metabolite level is expected [13•]. PLS regression with the two matrices of differentially expressed genes and metabolite changes was used to correlate the two indices, and 60 genes were identified as being altered over time by orotic acid. Hierarchical clustering of the 60 genes did not readily separate them into clearly regulated subgroups, but did highlight the major areas of affected metabolism. Mapping these transcripts to the Kyoto Encyclopedia of Genes and Genomes (KEGG) [14], as well as to the metabolites, demonstrated that measuring the two endpoints strengthened the biological deduction that would have been made from either dataset independently. A particularly interesting observation was made when statistical bootstrapping was used to search for correlations between the 60 most differentially regulated transcripts and trimethylamine-N-oxide, which was the detected metabolite with the largest change in aqueous tissue extracts. The most significant negative correlation was found with stearyl-CoA desaturase 1, which also correlated negatively with other metabolites in lipogenic pathways. The authors also surmised that orotic acid decreased transcriptional activation of sterol regulatory element binding protein-1c, although some genes under its control were not altered as expected, which could be interpreted as a disturbance of the healthy co-ordinate regulation of the genes possibly contributing to the formation of a fatty liver [12••]. This research nicely illustrates the use of currently available in silico analysis tools to search for gene and metabolite correlations that may not be intuitively obvious.

In a detailed study of hydrazine-induced hepatic injury, proteomics was used in addition to metabonomics and
genomics to study the mechanism of toxicity, and to uncover biomarkers of drug toxicity [15]. This study had an extremely complex design owing to the need to invasively obtain tissue for gene and protein analysis, and serum was used for 1H-NMR detection of metabolites. The study used ten rats per group, but it clearly was not possible to search for correlations between the three endpoints at the individual animal level; this is often one of the most significant limitations in attempting to integrate such datasets, particularly since at this largely exploratory stage there is usually a desire to link the changes back to the phenotype as defined by clinical and morphological pathology. The datasets were first analyzed independently and then inspected for correlations. As a 699-gene subarray designed for toxicity studies was used, a relatively small number of genes were identified as being altered. Similarly, two-dimensional (2D) gel analysis of proteins demonstrated just 60 proteins as being statistically significantly altered. With these small numbers, it was possible to manually search for direct gene to gene product (protein) correlations. Fifteen of the proteins were represented in the array, 13 of which corresponded qualitatively. Interestingly, a general observation was that transcripts were up- or downregulated by a factor of two or more as compared to the corresponding protein. At the level of the protein, transitional endoplasmic reticulum-adenosine triphosphatase represented a protein that was upregulated at both high and low dose, with no changes observed transcriptionally. The authors of this study speculate that this may be a classic example in which early transcriptional regulation was coupled to a short half-life messenger RNA (mRNA) such that expression changes were missed. Conversely, catechol-O-methyl transferase was induced at the message level, with no corresponding change in protein. However, there was considerable variation in the gene data from studies using three rats per group, while protein analysis used five rats per group, which emphasizes an important need to have sufficient numbers of animals per treatment when attempting to integrate "omics data. Serum metabolites were analyzed using PCA and the overall conclusion was that lipid metabolism was affected in a manner consistent with hydrazine inducing hepatic steatosis. No attempts to statistically correlate metabolites to proteins or genes were described, but impact at several points within pathways relating to lipid, glucose and energy metabolism was observable.

The results of the aforementioned study demonstrate that it is simpler to deductively identify associated proteins with genes, than genes with metabolites, simply on the basis of the protein being encoded by the mRNA. In an interesting study in a protein kinase C δ knockout mouse, Mayr et al were able to demonstrate a clear correlation in heart between altered proteins involved in energy metabolism and the metabolite profile as determined by 1H-NMR [16]. However, it is noteworthy that metabolites and genes were both measured in heart tissue rather than in separate biological compartments. For gene and metabolite integration, comparisons are mostly made between tissue gene expression and urinary metabolite profiles. Nonetheless, while gene to protein correlation is generally easier, there is research which suggests that novel biology may be revealed when comparing genes with metabolites. In two related papers, Heijne et al described their findings on bromobenzene toxicity, firstly using a genomics and proteomics approach [17], and then with a genomics and metabonomics approach [18•]. Proteins and genes were analyzed using PCA and both endpoints revealed changes in protein synthesis/degradation, but only modest overlap between functions of protein with gene changes. Comparing genes with metabolites, on the other hand, provided new knowledge around several pathways apparently involved in bromobenzene-induced hepatic necrosis. Such findings raise the expectation that, with advances in analytical and statistical methods, previously unrecognized biological interactions will be revealed when these high-content datasets are integrated.

Ringeissen et al used genomic and metabonomic analyses to identify novel, non-invasive biomarkers of peroxisomal proliferators [19]. 1H-NMR evaluation of urine from treated rats demonstrated an increased excretion of N-methyl-Nicotinamide and N-methyl-4-pyridone-3-carboxamide, which are end products of the tryptophan-nicotinamide adenine dinucleotide pathway. Global gene expression profiling using a toxicology-specific array detected no changes, but selected genes in the putative metabolic pathway were quantitated using reverse transcription-polymerase chain reaction. It was demonstrated that peroxisome proliferator-activated receptor (PPAR) α or δ agonists downregulated the hepatic expression of aminocarboxymuconate-semialdehyde dehydrogenase by up to 11-fold. This finding was consistent with that of other researchers who had demonstrated decreased activity of the enzyme following PPAR agonism. Mechanistically, it is speculated that the enzyme is a control point for the formation of nicotinamide adenine dinucleotide (NAD+) from tryptophan and that the liver is responding to a need for NAD+ to permit β-oxidation of long-chain fatty acids. Therefore, in this case, transcriptional data were employed to support the use of novel biomarkers identified by metabonomics which might not necessarily have been expected to be urinary measures of PPAR agonism.

Methods of data integration

As described above, functional genomics and metabonomics approaches are sources of rich information in terms of the biological plasticity of cellular responses, with each method independently providing a somewhat different view. In light of this it is of special importance that supplementary innovative and integrative analyses be employed to provide insights into the mechanisms that underlie the molecular events, hence allowing genomic, metabonomic and proteomic results to be linked in cause-effect relationships to explain phenotypic responses. Some biology- and data-driven approaches addressing this gap that have emerged in recent times are discussed below.

Biology-driven strategies

It is clear from the preceding sections that many investigators are currently attempting to qualitatively integrate orthogonal data from transcript profiling, 2D gel electrophoresis, 1H-NMR spectroscopy and/or MS [11,17,20,21•,22]. In the previously described acetaminophen
study in mice, it was concluded that genomics and metabolomics data together explained the global failure in energy metabolism resulting from acetaminophen intoxication [11]. Hirai et al, on the other hand, demonstrated co-ordinate regulation of gene-to-metabolite networks in primary and secondary metabolism of Arabidopsis thaliana [20]. These researchers were able to map significant changes in gene expression and metabolite levels under long-term nutritional stresses to nitrate assimilation, amino acid metabolism, carbon metabolism and sulfate assimilation. An interesting approach to integrating gene expression and metabonomics is illustrated in a recent report on biomarker identification of human neuroendocrine (NE) cancers, a clinically aggressive type of cancer with poor prognosis [23••]. Here, the authors combined data from a variety of samples (transgenic mice with a metastatic cancer arising from their prostatic NE cell lineage, NE cell lines established from this cancer and primary human NE tumors with varying degrees of aggressiveness) subjected to microarray and MS-based metabolomics for in silico metabolic reconstructions of NE cell metabolism. Predictions of cellular metabolism were made from transcript data, using METAVIEW, an in-house-developed software [24], and these were used to direct GC-MS and liquid chromatography (LC)-MS analysis of metabolites in specific metabolic pathways present in KEGG. Enriched γ-aminobutyric acid levels, through decarboxylation of glutamate by GAD1 (glutamate metabolism pathway) as well as an alternative pathway involving ornithine decarboxylase which yielded additional cancer-relevant metabolites, were confirmed by MS assays. Aside from combining the various molecular readouts such as genomics, proteomics and metabolomics data, integration of phenotypic data is key to understanding the system, as has been illustrated in some highly interesting studies [25-28,29•].

Biological, hypothesis-driven strategies for the integration of high-content datasets are mostly driven by bioinformatics-based pathway analysis. By borrowing content from a variety of curated commercially and publicly available pathway resources (a comprehensive listing can be found in reference [30]), gene and metabolite lists can be mapped to a multitude of high-quality canonical and assembled pathways. Arita has provided a useful review with tables detailing links to computational resources for metabolomics, including pathway resources [31].

Goodacre et al have summarized technologies and approaches to translating individual metabolites into a contextual understanding of the biology taking place in the system under study [32•]. Text-mining methods have been increasingly used to extract relationships between various biological entities. In particular, microarray data have benefited from Gene Ontology whereby genes/proteins have been classified into a hierarchical structure of curated biological processes, cellular components or molecular functions. Curtis et al have commented on the various tools available to determine the significance of such 'functional grouping' of genes [33•]. It must be noted, however, that the coverage of pathway annotation varies from organism to organism. For example, there is relatively lower coverage for rat than for human or mouse.

Unfortunately, with metabonomics being a nascent field, there are few resources that offer curated pathway content, such as KEGG and Reactome [34]. An exciting possibility is that text mining applied to metabolites is likely to yield previously unknown connections in biological networks or overlooked associations with different genes. For example, currently most pathway databases do not have information on molecules such as vitamin D or ceramide that have roles in regulatory and signaling pathways, which may therefore result in key regulatory elements being missed. A point to consider is that the results of pathway mapping are highly dependent upon the coverage of pathway information within any one given resource. One potential way to overcome this problem is to incorporate multiple resources to improve coverage, helping to reconcile inconsistencies wherever possible.

Data-driven strategies

A complementary approach to integrate large-scale quantifications can be to utilize an unsupervised, data-driven methodology. Correlations are often performed to search for groups of genes/metabolites that may be involved in a shared biological process or are co-regulated [35,36]. Several pattern-recognition algorithms have been employed to determine trends in individual data, but have also been used to combine different analyte measurements. Previously in this review, it was mentioned that PCA, a multivariate statistical technique which reduces high-dimensional data into fewer dimensions by preserving the overall characteristics of the data, is a popular method for viewing high-content datasets [12••,18•]. Linear combinations of the variables in the form of orthogonal principal components (PCs) are calculated, where consecutive PCs capture decreasing amounts of variance and allow visualization of clusters or groupings in the data. PC-discriminant analysis (PC-DA) is a 'supervised' derivation of this method and involves discriminant analysis of pre-assigned groups that are assumed to be more similar than other groups, to identify the largest contributors to the difference. PC-DA was used in the study conducted by Verhoeckx et al to integrate mRNA, protein and lipid measurements in U937 macrophage cells [37•]. It was found that the resolution of transcriptomics was sufficient to separate three classes of anti-inflammatory compounds, β2 agonists, corticosteroid and a proteasome inhibitor, although separation was much improved after integration with protein and lipid data.

Self-organizing map (SOM) is another technique that has been employed for global integration [38,39]. SOM is a neural network algorithm based on unsupervised learning and allows visualization of correlations in the data on a grid. While it is an unsupervised technique, it often involves dimension reduction as a prior step where loss of information is possible, and the technique can be computationally expensive and is limited by the dimensions of the lattice. Using this approach, regulatory metabolites
and transcriptional factor genes in glucosinolate biosynthesis and anthocyanin biosynthesis were identified by Hirai et al [39].

PLS is a regression-based method and has been employed by many groups to integrate gene expression with metabolite levels [12••,40••]. Li et al extended this approach by incorporating a genetic algorithm (GA)-based selection of genes prior to building a PLS framework for hepatocellular metabolism [40••]. The advantage of GA/PLS is that a gene can be assigned to more than one group and takes into account metabolic flux profiles. A priori information can be incorporated into the framework to build models robust to noise. Bayesian networks are an excellent choice for adding prior information. In some of the authors' own research, it was concluded that correlation-based methods such as cluster correlations [41] and relevance networks [42] have potential in integrating various data types.

Truly dynamic metabolic network reconstructions, based on reciprocal experimentation and modeling are few and far between. Such approaches not only integrate transcript and metabolite changes, but aim to link them causally. Ordinary differential equations-based modeling is yet to be applied in mammalian systems, but has demonstrated interesting results in bacteria [43••].

Technical and conceptual challenges for integromics

Undoubtedly, many researchers are providing key tools and technologies to facilitate the integration of high-content datasets. Nonetheless, there are areas for which significant gaps remain, but the authors foresee many of these being addressed as a result of collaboration among scientists of many disciplines. Table 1 highlights the major gaps and the factors that contribute to them. In the final section, these gaps and the ways in which they pose challenges to integromics are discussed.

Data standards and databases

The integration of multiple data types has the potential to yield many valuable insights into a biological system under investigation. However, there are several issues that ought to be considered in such analyses. High-throughput technologies are likely to yield false positives, and integrating error-prone data points from one such assay with another could result in noisy correlations. While great strides have been made in statistical methods used to analyze individual transcriptomic, proteomic, metabolomic or phenotypic data, there is an urgent need for sophisticated algorithms to combine the readouts in different ways to yield significant, high-confidence hypotheses. In this regard, a significant challenge facing the community is to make various kinds of results comparable, and the lack of standards poses a huge problem. Microarray standards are becoming more widely accepted with the advent of large repositories of publicly accessible databases. Proteomics and metabolomics standards have lagged behind, however, in part due to the relative adoption and state of the technology. A few data exchange formats have been emerging in recent times. Bino and his distinguished co-authors have proposed a standard for performing metabolomics experiments, known as MIAMET (Minimum Information About a METabolomics experiment) [44]. Similar proposals have been made by others, including those heavily invested in NMR technology [45]. Kell also provides an excellent treatise on database standards [46••]. The absence of a clear winner among data standards helps to explain the fact that, while a few databases with sparsely populated proteomics datasets exist, metabolomics data warehouses are almost non-existent.

Additionally, genes/proteins have universal identifiers (IDs) (eg, GeneID, Refseq ID, Genbank Account, etc) that make meta-analyses of various microarray platforms, as well as integration of protein and transcript profiles, possible in an objective manner. Metabolites do not have universally accepted ID assignments, and hence custom IDs (eg, KEGG compound IDs, CAS numbers, etc) or faulty string/structure matching can cripple data integration efforts.

Time dependencies and dynamic modeling

A caveat in combining various analyte readouts is that biological responses are complex, multi-faceted, context-

---

Table 1. Technical and conceptual challenges for integromics.

<table>
<thead>
<tr>
<th>Identified gaps</th>
<th>Contributing factors</th>
</tr>
</thead>
</table>
| Data standards and databases | • No universally accepted identifier assignments for metabolites  
• Lack of universally accepted standards for proteomic and metabolomic data and hence, limited data exchange formats  
• Repositories of proteomics and metabolomics data are sparse |
| Time dependencies and dynamic modeling | • Readouts from steady-state systems do not reflect time course dynamics  
• Lack of sampling of informative time points results in time lags between various data types (ie, gene, protein, metabolite) being ignored  
• Need for invasive tissue collection limits time points  
• Attempts at correlating endpoints occurs across subjects (eg, transcripts and proteins in one animal and metabolites in another)  
• One to many relationships, such as a metabolite being a product of an enzymatic reaction but also serving as a feedback loop inhibitor, or inhibitor of an unrelated enzyme |
| Analytics and annotation | • Multiple analytical methodologies for metabolites such as MS, 1H-NMR and Fourier transform-infrared spectroscopy  
• Generally poor annotation for metabolites  
• Lack of reference metabolite databases |
dependent and inherently dynamic. Therefore, data collected in steady-state systems may not account for dynamic trends in levels of transcript, protein, metabolite and phenotype. This may help to explain poor correlations between the transcriptome and proteome or metabolome [2••,13•,46•,47,48], as was described previously for the study on hydrazine-induced hepatotoxicity. In simple terms, the process can be linearized as genes, then proteins and finally, metabolites. However, it is readily apparent that with multiple control mechanisms, feedback loops, PTM of proteins, etc, this is a gross oversimplification of the situation. Metabonomics readily lends itself to time course studies when performed in non-invasive biofluids. In most cases transcriptomics and proteomics requires collection of tissue, which greatly limits the number of samples and time points available. Therefore, for data integration, there is often a discordance between the times at which the various endpoints are measured, and extrapolations are made across animals; for example, gene expression in one animal is linked to metabolite changes in a different animal, or in other cases, group mean values are used, rather than values from individual animals, to make inferences on gene to protein to metabolite inter-relationships. In either case, there is a loss of statistical power and it is not possible to link genes and metabolites to a defined phenotype, especially if the goal, for example, is to understand a pathological lesion in a toxicology study. Currently, it is clear that the strongest relationship between a gene and metabolic consequences, as judged by metabolite profiling, occurs when there is a specific gene alteration such as in a knockout or transgenic species, as has been demonstrated repeatedly in non-mammalian systems. Increasingly, this is being observed in mammalian models, such as the ndnl mouse model of neuronal ceroid lipofuscinosis, reviewed by Griffin, who also describes the use of metabolites to define the action of dystrophin and related proteins in a mouse model of Duchenne muscular dystrophy [49].

Intriguingly, the integrated data approach has also served to exemplify the importance of looking not just at individual metabolites linked to particular genes or proteins, but in understanding metabolic fluxes as described in metabolic control analysis [50]. A recently published article describes how it can be important to understand the dynamics of signaling pathways modulated by metabolites, rather than simply the instantaneous concentrations of their components [46•]. This is similar in concept to considering large-scale gene expression changes through transcription factor linkages, as opposed to clustering genes strictly on the basis of expression values. It is thus apparent that sampling of the systems at informative time points, and applying an appropriate time series analysis which reflects lag times between the various analytes in a context-dependent manner, is essential to functionally integrate these dynamic biological processes.

Analytics and annotation
Keeping in mind the pitfalls and resolution of data generated by various high-throughput technologies, such as gene chips, protein arrays, LC-MS, GC-MS, NMR, etc, an integrative analysis should ideally leverage the information provided by each data type. Transcript profiling is extremely powerful in understanding transcriptional regulatory networks. Due to post-transcriptional and translational events, as a technology, microarrays yield lesser insight into the activation of various signaling pathways compared with other technologies, unless the effects are manifested at the mRNA level. Differential proteomics data (eg, isotopically coded affinity tags), albeit of relatively lower resolution, are able to reveal signaling pathways in function. Similarly, metabolomics data, with an even lower resolution, is suited to understanding the metabolic pathways being modulated. Increased resolution can be afforded by providing more comprehensive coverage of the metabolome [44]. One approach to this problem is to improve the analytical techniques, both MS and NMR. Krishnan et al provide a good review of the improvements in NMR for metabolite fingerprinting in plants, and compare the platform to the more traditional MS methods [51]. Clearly, NMR continues to improve in sensitivity, but annotation remains limiting. Coupled MS methods have increased the coverage of the metabolome, but global profiling is challenged by the need for volatalization for GC-MS, and chromatographic separation for LC-MS. Improvements in NMR technology, such as cryogenic probes and hyphenated techniques (LC-NMR-MS) in which samples are analyzed by both NMR and MS, may help to address the shortcomings imposed by either technique used in isolation. An important technical note is that it becomes imperative to have well-controlled experimental conditions when measuring greater numbers of metabolites with greater sensitivity, to prevent the detection of minor metabolites influenced by factors not germane to testing the hypothesis [52,53,54••]. Technological advances in instrumentation have increased our ability to visualize the metabolome, and while it is evident that greater numbers of metabolites can be detected in many cases, annotation remains a major gap, as a lack of annotations renders it difficult to apply biological context to the data. One of the challenges presented by the rapid evolution of multiple analytical platforms is that the creation of reference metabolite databases can be unique to each platform and the instrumental conditions. Therefore, this field would benefit from collaboration among researchers to begin to create reference libraries that could be leveraged to increase annotation of complex spectra.

Summary
The potential power in harnessing both genomics and metabonomics together is evident in the burgeoning databases in which genes and metabolites are overlaid (eg, MetaCyc and MapMan) [44]. Taken together, intersections between transcriptional regulatory, signaling and metabolic pathways will reveal holistic insights about the system under investigation. As part of the critical path initiative, the US Food and Drug Administration has formalized voluntary genomics data submissions, whereby the relevance of genomic data to decision-making is being assessed. It may be envisioned that in the future, with technological advances and a multitude of high-dimensional data available, integration of different data types will provide a
comprehensive understanding of drug treatment in terms of efficacy and safety, which far surpasses our current level of knowledge [48]. Undoubtedly, the recent surge in metabolite analyses has brought integromics closer to reality; nonetheless, it is clear that we are not ‘there’ yet.

Acknowledgments
The authors thank Axel Meissner, Brad Ackermann, Jean-Marie Colet, Tim Ryan and Jim Stevens for critical review of this manuscript. The authors also acknowledge Ketan Patel for insightful comments and for generating Figure 1.

References

29. Jones GL, Sang E, Goddard C, Mortishire-Smith RJ, Sweatman BC, Jones GL, Sang E, Goddard C, Mortishire-Smith RJ, Sweatman BC. Additional paper: Computational resources for the analysis of pathways to the analysis of models of cardiac disease in the mouse. Multivariate statistics allowed the profiling is an effective means to phenotypically describe four separate •
32. Goodacre R, Vaidyanathan S, Dunn WB, Harrigan GG, Kell DB. Metabolomics by numbers: Acquiring and understanding global metabolic data. Trends Biotechnol (2004) 22(5):245-252. • Provides a highly readable review chapter that overviews the current status of metabolomics, with an emphasis on analytical methods and data analysis tools that are used to turn information into knowledge.
37. Verhoeckx KC, Biljsma S, Jensersen S, Ramaker R, Verheij ER, Witters LA, van der Greel J, Rodenburg RJ. Characterization of anti-inflammatory compounds using transcriptionrics, proteomics and metabolomics in combination with multivariate data analysis. Int Immunopharmacol (2004) 4(12):1499-1514. • Provides an excellent example demonstrating the power of integrating microarrays, 2D gel electrophoresis and LC-MS-based lipid profiling in U937 macrophage cells (stimulated with lipopolysaccharide with or without compound) to characterize the effects of six anti-inflammatory compounds. PC-DA was used to integrate the various data types.
46. Kell DB: Metabolomics, machine learning and modelling: Towards an understanding of the language of cells. Biochem Soc Trans (2005) 33(07):520-524. • Review offering provocative thoughts regarding philosophical approaches to integrating biological endpoints into systems biology. An emphasis is placed on the concept that understanding of biological signaling pathways is not fully realized when measuring/focusing solely on absolute concentrations of pathway intermediates, but rather on how downstream elements are affected by dynamic changes.
54. Nicholson JK, Holmes E, Wilson JD: Gut microorganisms, mammalian metabolism and personalized health care. Nat Rev Microbiol (2005) 3(5):431-436. • Describes advances in analytical techniques and data analysis that permitted the detection and identification of greater numbers of metabolites than ever before. Accordingly, it becomes even more imperative to understand the implications or relevance of the findings. How gut microflora can profoundly affect the response of an organism to nutrition and drugs, and hence the metabolic profile, are also discussed.