BMC Bioinformatics

RESEARCH ARTICLE Open Access

# MetaMapp: mapping and visualizing metabolomic data by integrating information from biochemical pathways and chemical and mass spectral similarity

Dinesh K Barupal[1], Pradeep K Haldiya[1], Gert Wohlgemuth[1], Tobias Kind[1], Shanker L Kothari[3], Kent E Pinkerton[2] and Oliver Fiehn[1*]

## Abstract

**Background:** Exposure to environmental tobacco smoke (ETS) leads to higher rates of pulmonary diseases and infections in children. To study the biochemical changes that may precede lung diseases, metabolomic effects on fetal and maternal lungs and plasma from rats exposed to ETS were compared to filtered air control animals. Genome- reconstructed metabolic pathways may be used to map and interpret dysregulation in metabolic networks. However, mass spectrometry-based non-targeted metabolomics datasets often comprise many metabolites for which links to enzymatic reactions have not yet been reported. Hence, network visualizations that rely on current biochemical databases are incomplete and also fail to visualize novel, structurally unidentified metabolites.

**Results:** We present a novel approach to integrate biochemical pathway and chemical relationships to map all detected metabolites in network graphs (MetaMapp) using KEGG reactant pair database, Tanimoto chemical and NIST mass spectral similarity scores. In fetal and maternal lungs, and in maternal blood plasma from pregnant rats exposed to environmental tobacco smoke (ETS), 459 unique metabolites comprising 179 structurally identified compounds were detected by gas chromatography time of flight mass spectrometry (GC-TOF MS) and BinBase data processing. MetaMapp graphs in Cytoscape showed much clearer metabolic modularity and complete content visualization compared to conventional biochemical mapping approaches. Cytoscape visualization of differential statistics results using these graphs showed that overall, fetal lung metabolism was more impaired than lungs and blood metabolism in dams. Fetuses from ETS-exposed dams expressed lower lipid and nucleotide levels and higher amounts of energy metabolism intermediates than control animals, indicating lower biosynthetic rates of metabolites for cell division, structural proteins and lipids that are critical for in lung development.

**Conclusions:** MetaMapp graphs efficiently visualizes mass spectrometry based metabolomics datasets as network graphs in Cytoscape, and highlights metabolic alterations that can be associated with higher rate of pulmonary diseases and infections in children prenatally exposed to ETS. The MetaMapp scripts can be accessed at http://metamapp.fiehnlab.ucdavis.edu.

**Keywords:** Metabolic networks, Enzymatic pathways, Perinatal lung development, Lung surfactants

* Correspondence: ofiehn@ucdavis.edu
[1]UC Davis Genome Center, Metabolomics, Davis 95616, CA, USA
Full list of author information is available at the end of the article

## Background

Exposure to environmental tobacco smoke (ETS) during fetal development can cause serious health consequences in later stages of life to increase the risk of respiratory disease and susceptibility [1-4]. Biochemical studies suggest that ETS can alter cell signaling and metabolic functions that can impair normal cellular growth and morphology in lung tissues [5,6]. Environmental tobacco smoke exposure has also been associated with abnormal fetal development [7,8]. However, the exact biochemical changes in different organs of ETS-exposed animal models are not understood on the systems level.

Among the functional genomics technologies, metabolomics may better assist in understanding physiology on systems level, because the metabolome is the ultimate outcome of biochemical networks and close to disease phenotypes. Metabolic perturbations can be investigated on the levels of genomic topology, gene expression or proteomics, aided by functional ontology interpretation [9] or pathway mapping [10]. However, changes on gene and protein levels may not lead to actual changes in metabolic fluxes and abundance levels, the realm of metabolomic techniques. Data acquisition and statistical analysis of metabolomic data have undergone extensive advancement in the past 10 years [11,12], but interpretation of metabolic data is much less straightforward than that with genomic and proteomic data sets. Unlike in gene and protein expression studies, no single technology platform can completely cover all metabolites present in organisms. Physicochemical properties of complex lipids, volatiles, primary metabolites and exogenous components such as vitamins and food phytochemicals are too different to be analyzed by a single device. Nevertheless, a good default approach is to target the most conserved part of metabolism, called primary metabolic pathways that list well known intermediates such as metabolism of carbohydrates, amino acids, fatty acids, hydroxyl acids, nucleotides, purines and related compounds. Quantitative analysis of common primary metabolites is most useful to understand major physiological consequences, e.g. growth [13] and diseases [14], as many primary metabolic pathways are well studied with respect to regulatory aspects of associated genes and enzymes. Indeed, most primary metabolites are captured in standard biochemical pathway databases such as MetaCyc [15] or the KEGG LIGAND repository [16] while biochemical knowledge repositories for lipid, secondary and volatile metabolism are far less advanced. Most primary metabolites have molecular masses below 550 Da which makes them amenable to data acquisition using gas chromatography (GC) and mass spectrometry (MS) after derivatization [17], albeit with notable exceptions such as di- and triphosphates (e.g. ATP, NADPH, fructose-1,6-bisphosphate) or selected other compounds (e.g. beta-carotene, betaine, S-adenosylmethione).

To cover these compounds, hydrophilic interaction chromatography/electrospray tandem MS [18] and capillary electrophoresis/MS [19] have been used which focus on hydrophilic metabolites and thus complement separations based on lipophilic interactions (reverse phase liquid chromatography/MS) [20]. Each platform faces technical limitations which yet constrain reporting more than 200 identified primary metabolites per data set, as well as additional metabolic signals that refer to unknown and potentially novel metabolic intermediates. GC/MS can be considered as most mature because large mass spectral repositories have been compiled under standard data acquisition procedures to annotate small molecules, most prominently the NIST and Wiley libraries that cover more than 250,000 compounds. Metabolites are distinct in their three-dimensional structure (e.g. glucose, galactose, mannose) and thus need to be referred to by both mass spectra and standardized chromatographic retention which led to the development of small target libraries [21,22]. These libraries support metabolomic databases such as BinBase [23] that automatically process raw data files into input data sets for statistical comparisons, e.g. in cancer biology [24], plant biology [25], microbial studies [26,27] or metabolism of subcellular compartments [28]. Subsequently, the observed differential regulation of metabolites needs to be interpreted based on biochemical and physiological background information, both from pathway repositories [29,30] and literature databases like the human metabolome database HMDB [31]. While many metabolites can be mapped to overall metabolic modules, e.g. using KEGG LIGAND metabolic maps [24], it was noticed that many metabolites could not be mapped to any known metabolic pathways or reactions available in the KEGG database. Beyond the mere extension of genomic reconstruction databases, e.g. by pathway gap analysis [32] or community curation efforts [33], the presence of non-mapped metabolites may be explained by substrate and reaction promiscuity of enzymes [34,35]. In addition, even for the best studied organism like *Escherichia coli*, 40% of the genes are still not annotated with any cellular function [36]. For newly sequenced organisms, the number of non-annotated genes and thus uncertainty about presence of metabolic pathways is certainly even higher. Therefore, to map, to visualize and to interpret altered metabolic levels with respect to biochemical networks remains a formidable bottleneck in metabolomics.

Due to the sparse nature of metabolomic coverage and the presence of unaccounted metabolic signals in metabolomic data sets, efforts have been undertaken to utilize the inherent data structure beyond statistical comparisons. The Pearson's correlation matrix of metabolomic data was used to represent the metabolic relationships in a network context [37]. It was shown that the existence

of such correlation pairs actually reflects biochemical regulation [38,39]. However, translating a correlation link into a biochemical link is not straightforward as correlations may not only be driven by the action of enzymes but also other factors, e.g. transcription regulators. Another approach is to investigate the topological structure of network graphs, for example based on contrasting, comparing and correlating multiple nodes simultaneously [40] using static or dynamic networks [41]. However, while expression-based metabolite networks may shed light on hidden structures in data sets, biochemistry-based network graphs of metabolic reactions can serve as input to develop structural and organization models [42], leading to insights into evolutionary relationships [43-45], establish metabolic routes in a large metabolic networks [46] or predict cellular growth in microorganisms [47].

While the standardized methods for data acquisition and improved data processing have enabled generation of high quality and reliable metabolomics datasets, methods that would assist biological interpretation are confined only to metabolic pathway mapping using biochemical knowledgebases. Among those methods, biochemical visualization of metabolomics datasets using pathway diagrams showed several bottlenecks that should be overcome for biological interpretation of the statistical pattern within data. A genome sequence with its annotation and mRNAs and proteins with their expression values can be overlaid on static pathway diagrams to highlight sequence inferred presence/absence of a pathway or sub pathway in an organism, and to visualize mRNA/protein inferred increase/decrease in metabolic flux in multiple pathways. However, metabolomics datasets do not contain all the metabolites predicted in a genome constrained metabolic network, not all the identified metabolites in a metabolomics datasets can be mapped to pathway diagrams and 2/3 of the detected metabolites are unknowns. Metabolites are all different and cannot be sequenced from a linear code arrangement of building blocks, unlike genes, transcripts and proteins. Furthermore, metabolites can be members of many different reactions, as they reflect the ultimate output phenotype of the underlying complex regulatory and enzymatic network. Therefore, a biochemical visualization approach for metabolomics is required that is independent of genome sequence, and that can visualize all the metabolites in a metabolomics dataset, include all the known biochemical reactions for identified metabolites, yield customizable layout and efficiently visualize differential alteration in metabolite levels to assist biological interpretation. We here present an approach to integrate network graphs based on biochemical reactions with chemistry-based graphs. Differential expression of all the detected metabolite nodes is superimposed onto the graph to aid the biological interpretation of perturbations in metabolic networks.

## Results and discussion

### 41% of all detected metabolites were significantly regulated under exposure to secondhand smoke

GC-TOF MS based metabolomics of maternal and fetal lungs and maternal blood plasma extracts yielded over 700 distinct signals per chromatogram which were automatically deconvoluted and submitted to our open source BinBase mass spectral processing database. BinBase filters out noisy signals that are not positively detected in at least 50% of at least one study design class, excludes known artifacts from data export, adds potentially novel compounds that had never been detected before, annotates spectra by retention index/mass spectral matching to libraries of authentic standards and finally exports a high confidence data matrix for statistical and biochemical analysis, including KEGG and PubChem identifiers for each metabolite. Intensity values for compounds that were absent in some samples but positively detected in others are replaced by target ion signal intensities at the expected peak retention time, minus the lowest noise signal in local neighborhood, ensuring that a complete data matrix was available. Overall, 459 metabolites were detected in a consistent manner over all chromatograms, of which 179 were structurally identified. Between 285–377 metabolites were positively detected per organ (Figure 1A) using the stringent BinBase quality criteria. The far majority of all compounds were detected in at least two organs, verifyin the conserved nature of metabolism and the suitability for comparison of metabolic effects of treatments with environmental tobacco smoke (ETS) between animals and between organs. Consequently, one-way ANOVA comparisons were conducted for each metabolite between ETS-treated and control organs ($p < 0.05$, $n = 8$ per group in dams, $n = 46$ per group for fetuses). With notable exceptions, ETS treatment led to downregulation of metabolite concentrations for most compounds (Additional file 1: Table S1). Interestingly, the largest number of significantly regulated metabolites under ETS treatment was found for fetal lung metabolites (Figure 1B), despite the fact that maternal lungs were much more directly exposed than the embryo itself. Indeed, few compounds were found to be significantly regulated in more than one organ, indicating a highly specific and organ-dependant metabolic response to ETS exposure. Only 7% of the compounds were detected exclusively in blood but not in lung tissues. Moreover, the Venn diagram (Figure 1B) clearly shows that very few metabolic alterations were apparent in blood plasma and even fewer of these were shared with changes in either maternal or fetal lungs. This finding demonstrates that differential expression of lung metabolites was indeed directly associated

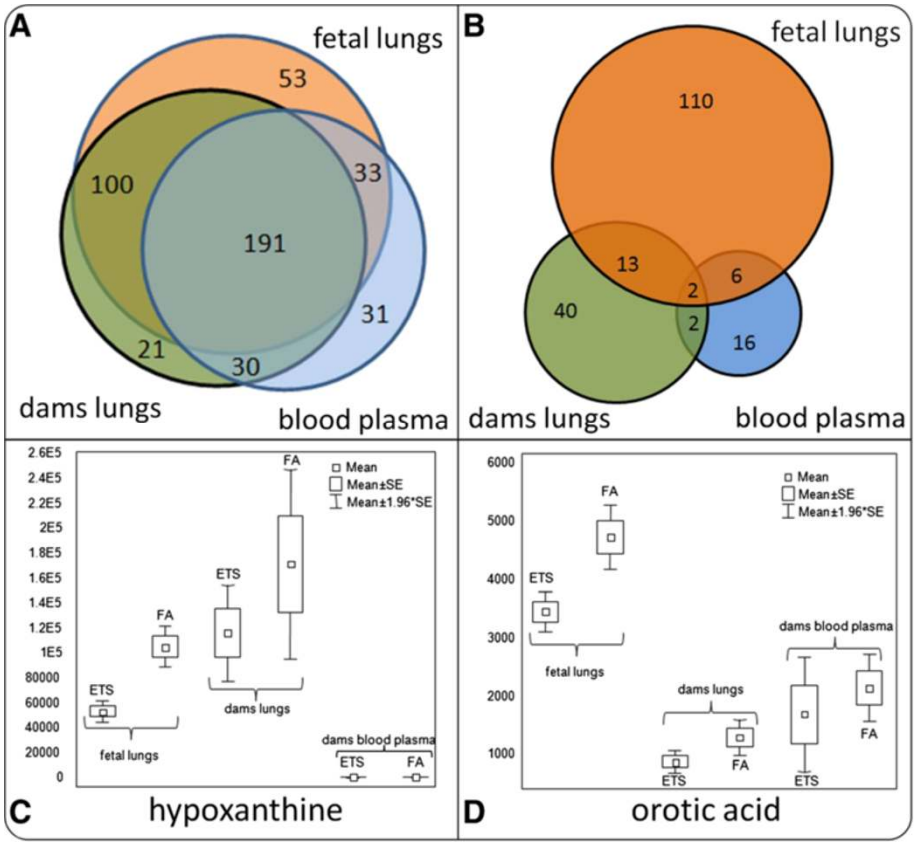


**Figure 1 Overview analysis of metabolomic data and differential metabolic regulation for fetal lungs, and maternal blood plasma and maternal lungs of rats exposed to environmental tobacco smoke (ETS) compared to filtered-air (FA) exposed animals. (A)** High confidence detection (BinBase) and overlap of metabolites among all three tested organs. **(B)** Number of differentially altered metabolites ($p < 0.05$), and overlap of significant differences among three organs. **(C&D)**. Exemplary box and whisker plots of two metabolites that were found significantly altered in three organs.

with tissue-specific changes in cellular regulation in lungs and not conferred by potential contamination with blood metabolites. Overall, 189 of the total of 459 metabolites were significantly different in at least one organ (41%). Each individual compound can be visualized in bar diagrams, e.g. for hypoxanthine and orotic acid which are involved in purine and pyrimidine pathways (Figure 1C, D). Tables or bar charts are hard to navigate with respect to functional relationships, especially when hundreds of variables are investigated. How can we depict all metabolic changes while simultaneously keeping visual clarity of and superimposing biochemical organization?

### Mammalian biochemical databases poorly covered the detected metabolome

Mapping all metabolites to biochemical pathways appears to be a logical approach for further structuring and finally interpreting the observed metabolic changes. A range of databases and tools have been developed over the past 10 years [48-50]. Accordingly, we have matched all 179 identified metabolites against seven biochemical repositories and additional chemical databases (Additional file 2: Table S2) to evaluate how many of these compounds were covered by biochemical knowledgebases. We deemed enzyme reaction databases most relevant that referred to mammalian biochemistry such as HumanCyc [15], Biochemical Genetic and Genomic knowledgebase BiGG [51], Reactome [50] and the Edinburgh Human Metabolic Network (EHMN) [52]. Surprisingly, 30-53% of all identified metabolites could not be mapped this way (Figure 2) although most compounds were supposed to be genuine endogenous lung metabolites and not e.g. derived from gut microbial metabolism (like hippuric acid) or food constituents (like sitosterol), detected solely in blood plasma. This finding indicated that genomic-reconstructed mammalian pathway databases are far from complete. We therefore queried global reaction pathway databases (MetaCyc and KEGG) that would encompass also non-mammalian genomes and reduced the loss biochemical coverage to only 15-24% of the structurally identified metabolites (Figure 2). These

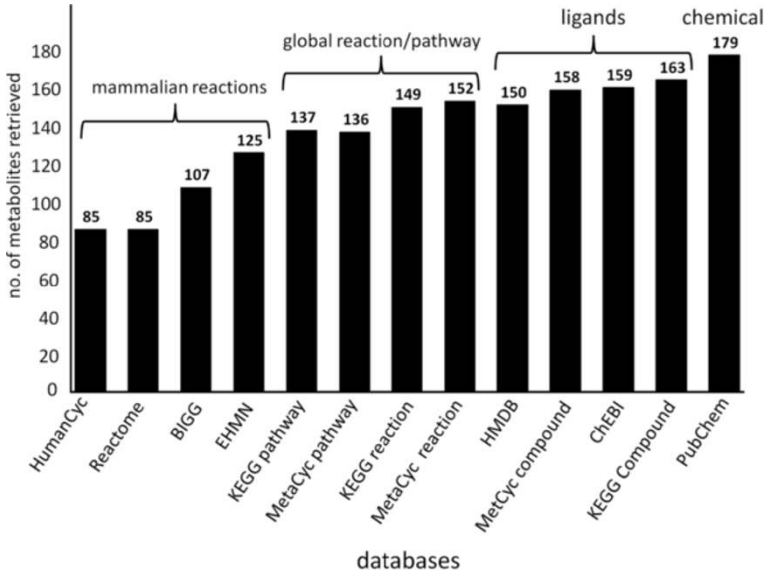


**Figure 2 Data representation of a total of 179 identified metabolites from the rat environmental tobacco smoke metabolomics study by querying various bioinformatics databases.** Databases were queried using KEGG and PubChem identifiers in addition to individual compound names.

compounds may be due to enzyme substrate or reaction promiscuity [53]. In the next step we searched ligand and chemical databases which did not directly refer metabolites to reactions or enzymes: KEGG Compound, MetaCyc Compound, 'Chemicals of Biological Interest' ChEBI [54], Human Metabolome database HMDB [31] and the largest freely available chemical repository, PubChem [55]. Only in PubChem, all compounds were referenced whereas in other databases, 9-16% of the identified metabolites were not catalogued. The majority of missing compounds were composed of lipids, sugars and sugar conjugates, pointing to current lack of knowledge of substrate and reaction specificity for many mammalian enzymes.

### Biochemical mapping leads to loss of structural clarity

Even without those compounds missing from a specific biochemical database, it might still be helpful to display the overall metabolic dysregulation of ETS-impaired lung metabolism on biochemical pathways, either by using available direct visualization tools or by network graphs. Some tools such as MetaCyc are focused on single pathways but do not readily facilitate matching overview results on 'all MetaCyc pathways'. We have first used seven publicly available direct 'global' visualization tools [50,56-61], (Additional file 3: Table S3). Tools were straightforwardly usable and indeed provided the capabilities as referenced. However, results were not satisfying due to several drawbacks: first, all tools had static visualization layouts which were defined by the boundaries of the genes (or proteins, metabolites, respectively) encoded in the tools, but not based on the actual input, here: the 179 structurally identified metabolites. A recently reported tool, MetExplore [60], uses all MetaCyc pathways for global mapping of metabolites. Unfortunately, it does not yield images but only returns lists of associated pathways, and not connections between these. Alternatively, the KEGG Atlas global map of 128 independent pathways [58] can be used. However, global pathway mapping approaches all suffer from lack of visual clarity because typical metabolomic data sets (such as the rat ETS data set used here) are sparsely populated. In cells, metabolic regulation focuses fluxes towards end products and some pools of intermediate branch points, but metabolic regulation does not lead to accumulation of many pathway intermediates which are therefore missing from data sets (Additional file 4: Figure S4), in addition to constraints given by the particular metabolomic platform used for data acquisition. Furthermore, 45% of our identified metabolites could not be mapped onto this KEGG Atlas global map because, only a fraction of the 371 reference pathway maps in the KEGG database are summoned in the Atlas global map (Additional file 4: Figure S4). In order to obtain an overall complete and structurally clear graphical view of mapping metabolites to pathways, alternative strategies need to be taken. Such graphs need to be able to adapt flexibly to the input data while displaying all input metabolites in a biochemically relevant overview. In order to aid biological interpretations, views should facilitate superimposing results of statistical analyses and focus on dysregulated pathway modules while also displaying all encompassed pathways. Additional discussion and comparison of various pathway

mapping tools are given in [62]. Ultimately, graphical pathway mapping visualizations serve biologists to draw conclusions or new hypotheses that would otherwise be difficult to obtain. The purpose of tools like MetaMapp is therefore to map data to biochemical modules to facilitate biological interpretations. Secondly, MetaMapp may be used to integrate metabolomic data with data from other Omics platforms, but it is not aimed at using graphs to predict fluxes or to predict enzymatic reactions for novel metabolites [63]. Instead of using static global maps, we have therefore tested using Cytoscape for visualization of biochemical pathway databases. While this approach certainly enables large overviews and zooming functionality, overall graph structures are determined by the high number of entities in pathway databases. In global KEGG pathway graphs, 4688 metabolites are present (2010 version, excluding the drug-like compounds), which led to densely packed, hairball-type visualizations that are unsuitable if 179 metabolites are matched (Additional file 5: Figure S5). The only exception returning a visually clear graph was by employing the KEGG pathway query that returned a list of 137 of our 179 ETS-study metabolites. Linking those 137 metabolites based on association with pathway maps as edges yielded a customized graph (Additional file 5: Figure S5) that proximately clustered metabolites according to biochemical neighborhood. However, the observed clustering was not very strong, and a range of compounds were found as isolated groups of nodes. In addition, 42 metabolites were not visualized as these were not present in the KEGG PATHWAY database. When using the KEGG reactant pair database, a total of 149 metabolites were mapped on a global view in a unipartite graph (Additional file 5: Figure S5, e). As this visualization was reflecting all metabolites comprised in the KEGG reactant pair database, the network graph still appeared very dense despite some emergence of biochemical modules based on overall reaction distances [64,65] in which the ETS-study metabolites were clustered. A biochemically superior approach is utilizing the information content of substrate/product reactant pairs for which the majority of atoms are shared, and indeed, metabolic pathways are best analyzed using atomic reconstruction mapping [66,67]. Using reactant pairs as founding parameter of metabolic pathways [68,69] is essential to derive biochemically relevant conclusions, unlike efforts that only utilize network topology data. In addition, reconstruction of pathways that are based on reactant pairs assists in identifying pathway gaps that can be filled in by assigning reactions from enzyme paralogs or orthologs of yet not-annotated enzymes [70].

### Chemical network graphs yield cluster resembling biochemical modules

While these initial networks were encouraging, they failed in visual clarity, strength of biochemical clustering and completeness of mapping our detected metabolites. We have therefore explored adding a radically different approach: if biochemistry refers to the conversion of chemically similar compounds by catalytic enzymes, it appears logical to associate all compounds directly by their chemical similarity. Clusters of chemical similar compounds should then resemble biochemical modules. The structures of all 179 identified metabolites can be encoded in molfiles which can be decomposed into substructures (Figure 3, labeled in colors) which are defined by a 881-bit publicly available set of substructures (ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.txt) in PubChem. The presence and count of all of these substructures define a matrix which was subjected to distance calculation utilizing the Tanimoto formula [71], yielding a pair-wise chemical similarity matrix among all studied metabolites. While we have used here only 179 variables, this approach can be easily extended to next generation metabolomics data sets that may encompass many more identified metabolites. These similarity matrixes can be visualized in Cytoscape graphs by applying thresholds of similarity scores to define network edges (Additional file 6: Figure S6). Tanimoto coefficients run from 0 to 1 from 'no similarity' to 'identical structure'. At high threshold settings (0.9), scattered graphs were obtained with many isolated compounds. Even at very low thresholds (0.5), compounds were found in isolation of the network, while clusters begin to disappear into densely packed patterns. At 0.7 Tanimoto coefficient thresholds, clear metabolite clusters resulted (Additional file 7: Figure S7). Isolated compounds were connected to the network by its single closest similar compound (see Methods). Resulting clusters of fatty acids, organic acids, sugars, sugar alcohols, phosphates, amino acids, nucleotides, purines and aromatics indeed were similar to patterns yielded by KEGG RPAIR matching networks. Most importantly, such chemical similarity network graphs can map metabolites that lack reaction annotation in any biochemical database. However, there were compounds that are known to be biochemically closely related (members of the tricarboxylic acid cycle, TCA) that were not found in close proximity in chemical similarity networks (Figure 4A and additional file 7: Figure S7). Succinic, aconitic and fumaric acid had higher chemical similarity to fatty acids than to hydroxyl acid s, and thus were placed in proximity to the fatty acid cluster. Hence, solely relying on chemical similarities fails to generate reactant pair networks that are fully suitable for enzymatic interpretations [63].

### MetaMapp integrated network graphs display all metabolites while maintaining biochemical organization

In order to resolve the shortcomings of both mapping approaches, we therefore combined KEGG reactant pairs and Tanimoto chemical similarity tools into a novel

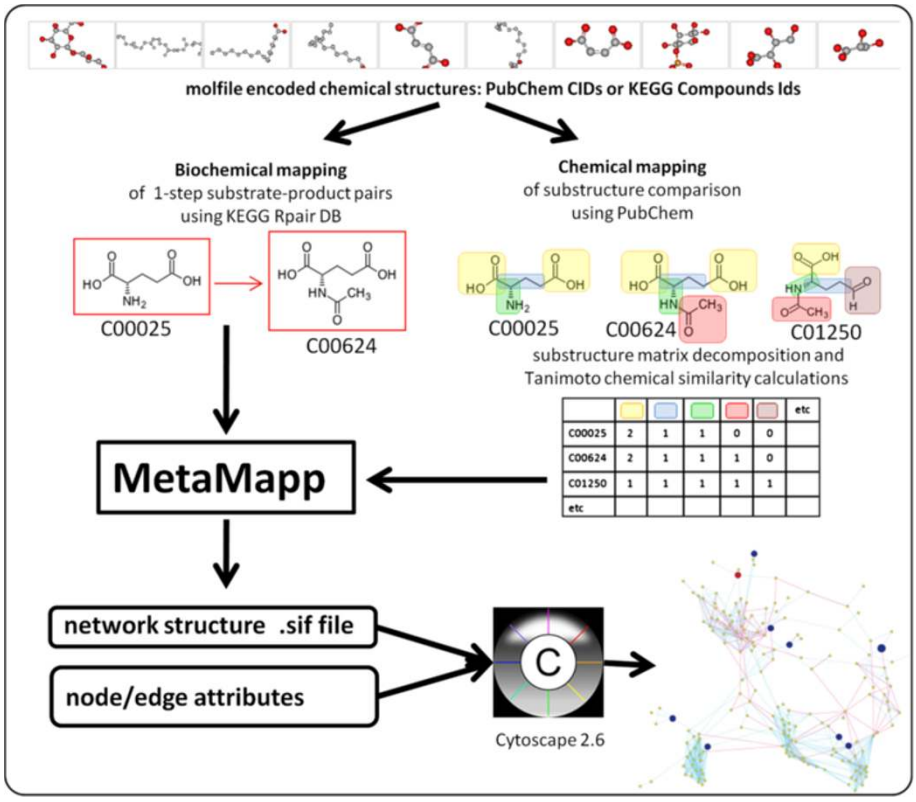


**Figure 3 Schema of network integration and visualization using MetaMapp and Cytoscape.** For biochemical mapping, the KEGG reactant pairs database was used. Chemical similarity mapping was performed using 881- substructure fingerprints within the PubChem database. MetaMapp tools then integrated biochemical and chemical similarity matrix files to visualize the network in Cytoscape. Attribute files such as fold-changes and statistical thresholds were added to inform about metabolic regulation in case/control studies.

method, MetaMapp (Figure 3). We have first mapped all identified metabolites of our ETS study against the presence in the KEGG reactant pair database for single-step conversions. 101 of 179 metabolites were returned in this query. Subsequently, a simple interaction format (.sif) network graph was constructed and visualized in Cytoscape. Cytoscape enables adding further node or edge metadata for visualization purposes, such as statistical significance information or magnitude of regulation. Results for the MetaMapp graphs are given in Figure 4, showing zoom-ins that highlight the improved biochemical interpretability from Tanimoto chemical similarity networks to MetaMapp graphs. Complete Cytoscape session files are given as additional information S8. As demonstrated by Figure 4A, TCA metabolites were scattered into different clusters of nodes using chemical similarity alone. The combination of biochemical reactant pair mapping (red edges) and chemical similarity (blue edges, Figure 4B) into one MetaMapp graph, however, correctly clustered the TCA metabolites into one group, separate from fatty acids, hydroxyl acids and sugar acids.

In order to display our data set in a truly comprehensive manner, we lastly aimed at adding the 280 unknown metabolite signals that could not yet structurally identified using the Fiehnlib or NIST mass spectral libraries. Electron ionization mass spectra of similar structures are known to cluster [72-74]. Hence, mass spectra of unknowns can be mapped against all other compounds, bringing unknown metabolites into proximity of biochemically relevant groups of nodes in networks. Using the NIST mass spectral similarity algorithm [75] at a forward similarity threshold of 700, and integrating sif files from biochemical, chemical and mass spectral similarity networks (yellow edges), all 459 metabolites of the lung and blood metabolome of ETS-treated rats were integrated (Figure 4C). Cytoscape does not provide capability to define one network as host or primary grid and further networks as additions; hence, overall biochemical clarity suffered by adding unknowns using mass spectral similarity. Nevertheless, TCA metabolites were still retained in close proximity (Figure 4C), giving biochemical relevance to three-tiered MetaMapp networks when aiming to classify differentially regulated metabolites of unknown structure into chemical classes and potential biochemical modules. In comparison to other ways of visualization of metabolome data, such as direct

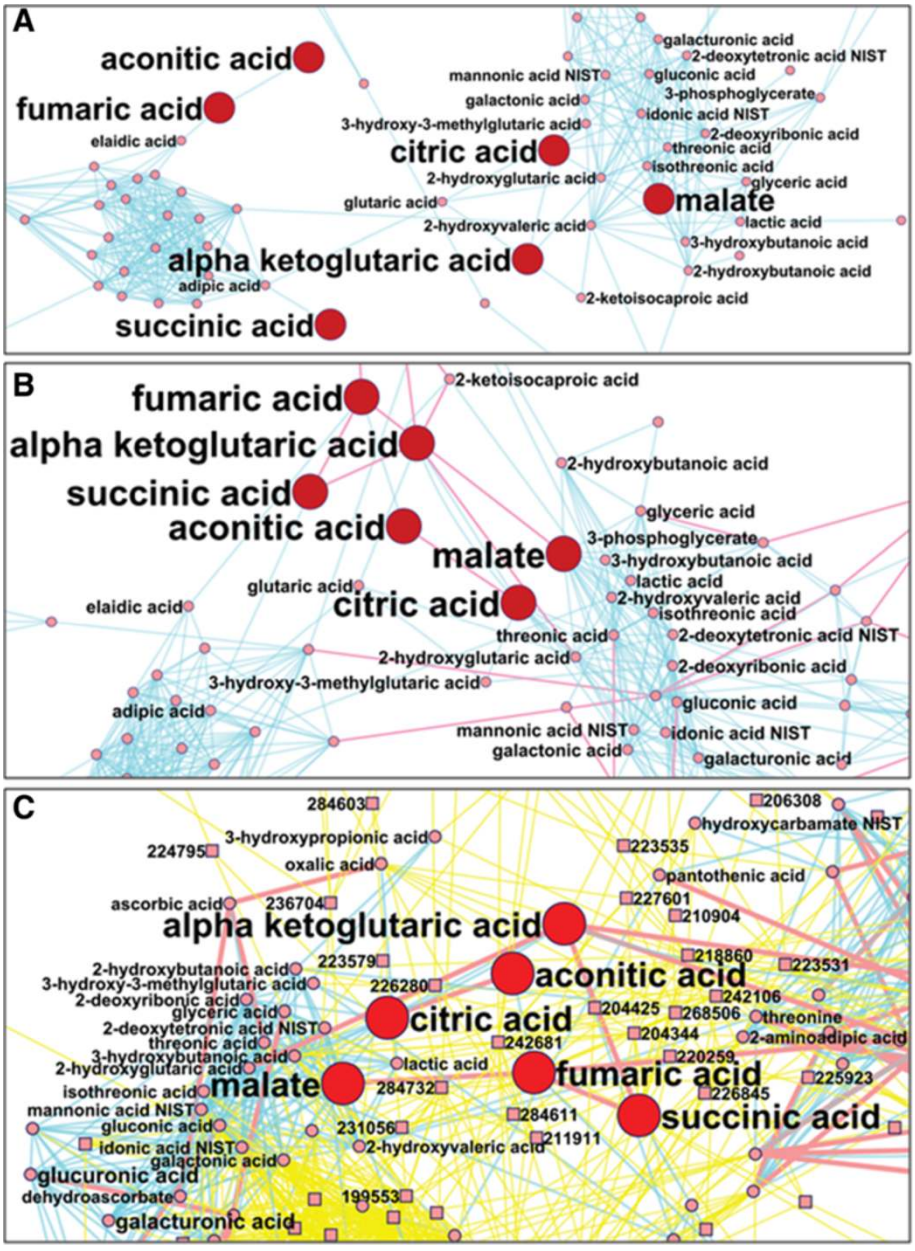


**Figure 4 MetaMapp zoom-ins for results of mapping metabolomic data using three different approaches, focusing on the biochemically strongly related TCA cycle metabolites as example (highlighted with bold labels and red nodes).** Identified metabolites are represented by circle nodes; unknown metabolites by square nodes. Red edges denote KEGG reactant pair links; blue edges symbolize Tanimoto chemical similarity at $T > 700$; yellow edges give mass spectral similarity $> 700$. Cytoscape session files are given as additional information S8, including metabolite names that have been left out of the network graphs for visual clarity. **(A)** Mapping 179 identified metabolites solely using Tanimoto chemical similarity as input data. **(B)** Integration of KEGG reactant pair information with the Tanimoto chemical similarity matrix (threshold $T > 700$). **(C)** Integration of KEGG reactant pair information with the Tanimoto chemical similarity matrix of all 179 identified metabolites and the mass spectral similarity matrix of all 459 compounds, including unknowns (squared nodes, exemplified with BinBase database identifier numbers).

biochemical mapping or mere statistical visualizations (e.g. bar diagrams, heatmaps or multivariate Partial Least Square plots), three-tiered MetaMapp networks appear to structurally organize information in a biochemically relevant way while enabling to overlay the network structure with further metadata, most importantly the significance and magnitude of class-wise statistics.

MetaMapp provides several advantages. First, it is independent of the technology utilized to identify metabolomic profiles, be these mass spectrometry- or NMR-based. This

means that data from different metabolomics platforms can be readily integrated and visualized to infer biological conclusions. The only requirement is that all chemical structures are associated with machine encoded chemical structures. Second, MetaMapp is not constrained by genomics information. All detected metabolites can be straightforwardly mapped across studies or species, enabling mapping of metabolites that originated from diet or gut microbes along with compounds that stem from mammalian enzymes. Third, the MetaMapp layout is not static and is automatically updated according to the input list of compounds; hence, MetaMapp graphs enable high biochemical clarity despite a large number of metabolic nodes. However, MetaMapp also had shortcomings, some of which may be partly resolved in further extensions: (a) MetaMapp cannot be used to compute fluxes, or to compute enzymatic reactions, between metabolites. (b) MetaMapp is scalable to some extent, but any network graph may get blurred when adding large numbers of nodes; in our case, this was observed by adding too many unknowns using MS similarity. (c) MetaMapp is most useful for visualizing case/control comparisons. Visual clarity suffers when statistical results are added from additional comparisons, and we therefore suggest using multiple two-way graphs for displaying data from more complex biological study designs.

### Fetal lung metabolism is more affected by environmental tobacco smoke than metabolism in organs of dams

Changes in the fetus lung metabolic networks in comparison to metabolic impairments in the organs of dams were subsequently visualized in MetaMapp graphs and interpreted in the context of biological functions. As stated above, 189 of the total of 459 metabolites were significantly different in at least one organ (41%) when comparing ETS- challenged rats to rats in control conditions. Using MetaMapp, the results of statistical tests were now superimposed to the biochemical network structure for comparing metabolic alterations in fetal lungs and dams' lung and plasma.

In Figure 5, mapping metabolomic data is highlighted for all three rat organs, displaying only significantly altered metabolites ($p < 0.05$) while not labeling the unchanged compounds. For clarity, metabolites of unknown structure have been excluded. The 179 identified metabolites were clustered into five major network clusters (see, e.g., the fetal lungs metabolomic network, top graph in Figure 5): amino acids and amines in the upper left corner, fatty acids in the lower left corner, purines and pyrimidines in the upper right corner, carbohydrates in the lower left corner and dicarboxylic and hydroxyl acids in the center of the graph. Few intermediate metabolites were interspersed in the graph, including metabolites comprising the TCA cycle that was biochemically correct found to connect the fatty acid cluster to the amino acid cluster.

In all three organs, most of the affected compounds were down-regulated under ETS-stress, as indicated by the blue-colored nodes in the networks. By far the largest number of differences was observed in fetal lung tissues (78 compounds, compared to 29 and 9 metabolites in lungs and plasma of dams). 68% (13/19) of the detected free fatty acids were dysregulated in fetal lungs, compared to only 8% (3/19) of the free fatty acids in the lungs of dams. Fatty acids are important components of pulmonary surface-active lipids and alveolar membranes [76]. Both components are critical in the breathing process and indicate that important metabolic building blocks for lung development were found down regulated. It has been shown that significant de-novo fatty acid biosynthesis is performed in the developing fetal lungs [77], in addition to hepatic metabolism [78]. Decreased amounts of those fatty acids in the developing fetal lung hence may lead to impaired lung function after birth. Moreover, a round 50% of purines/pyrimidines were dysregulated in both fetal and dams lungs (all metabolomic result data are given in Additional file 1: Table S1). Purines and pyrimidines are required for DNA and RNA biosynthesis and are linked through their ribose units to the pentose phosphate cycle, which also generates reducing power by NAPDH production. A decreased level of these metabolites can generally indicate a lower rate of cell division [79]. Conversely, uric acid was among the few compounds found to be increased in fetal lungs. As uric acid is a metabolite known for its antioxidant properties in the respiratory tract [80], its increased concentration might indicate that fetal lungs were already under oxidative stress prior to birth. The only further elevated compounds were metabolites directly related to energy metabolism, i.e. glucose, glucose-6-phosphate and the ketone body 3-hydroxybutyrate. Higher levels of glucose and glucose-6-phosphate can be interpreted by lower fluxes through glycolysis and the pentose phosphate pathway [81], because fewer amounts of structural carbon backbones are needed for biosynthesis in cell division. Antioxidant defense, energy metabolism, nucleotide production and fatty acid metabolism are co-ordinated with the cell cycle [82,83] and hence support the notion of lower rates of cell division in ETS-challenged fetal lungs.

In addition, other critical parts of lung metabolism were impaired as well. The MetaMapp graph for fetal lung dysregulation (Figure 5) shows that several amino acid pools were down-regulated, among them proline levels. Proline is one of the most important building blocks of lung collagen, a structural protein for connective tissues that provides mechanical stability and elasticity to the pulmonary tissues. Similarly, isoleucine was reduced, an amino acid that is found enriched in the lung surfactant protein B. Next, both glycocyamine and its anabolic product creatine/creatinine

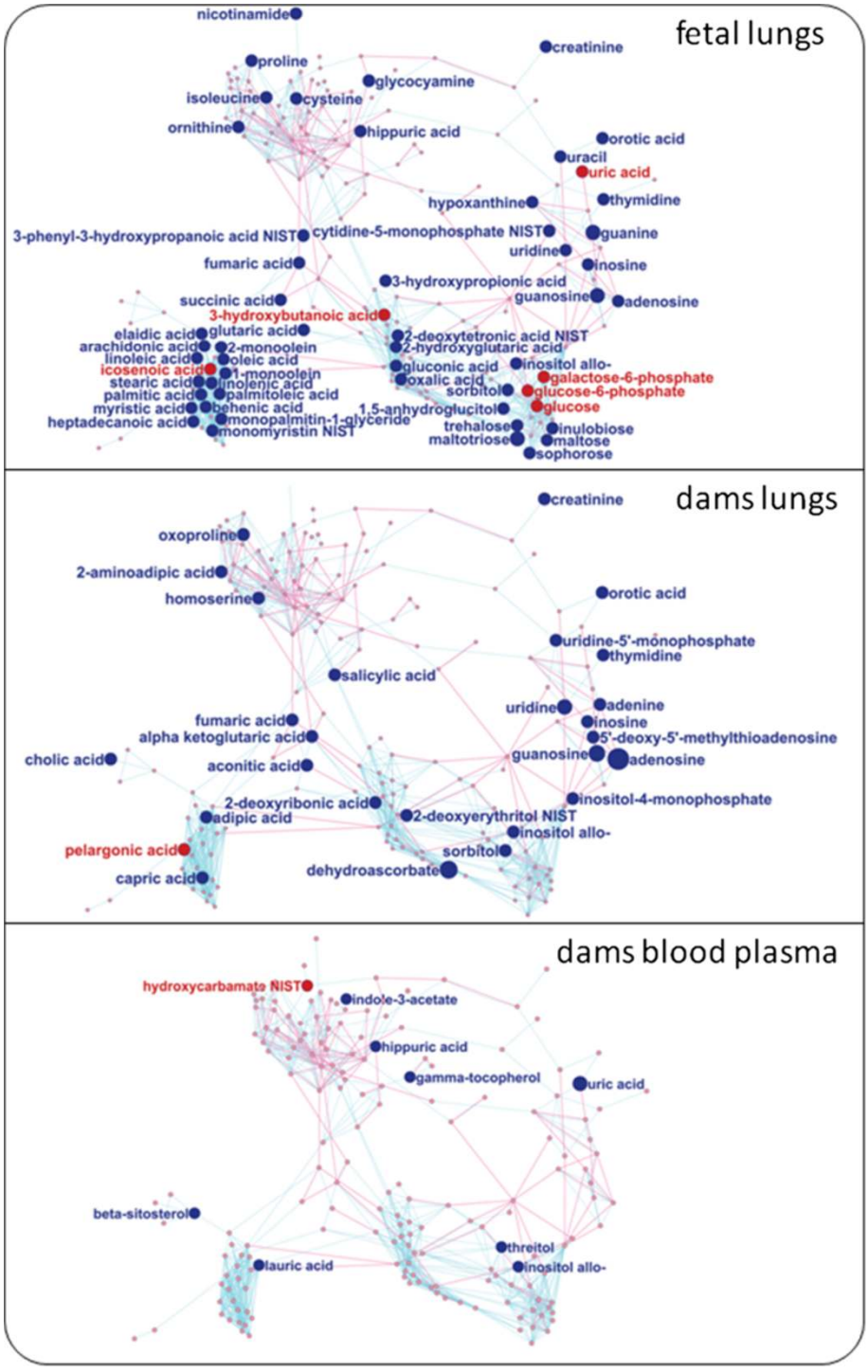


**Figure 5 MetaMapp visualization of metabolomic data highlighting the differential metabolic regulation in fetal lungs, maternal blood plasma and maternal lungs of rats exposed to environmental tobacco smoke compared to filtered-air exposed animals.** Red edges denote KEGG reactant pair links; blue edges symbolize Tanimoto chemical similarity at T > 700; unknowns are left out of these graphs for visual clarity. Metabolites found significantly up- regulated under exposure to environmental tobacco smoke ($p < 0.05$) are given as red nodes and labeled by BinBase names; blue nodes give down-regulated metabolites. Node sizes reflect fold change. Metabolites that were not found to be differentially regulated were left unlabeled for visual clarity. Red edges denote KEGG reactant pair links; blue edges symbolize Tanimoto chemical similarity at T > 700.

were found at lowered levels in fetal lungs. Creatine and creatinine are needed to supply energy in muscles. Decreased amounts may impair lung muscle contractions. Newborns of ETS-exposed dams, thus, appear to be born with compromised lung metabolism that may impact the lung surfactant and membrane fluidity system, lung flexibility and lung muscular strength. In humans, epidemiological studies show similar effects of environmental tobacco smoke. In-utero exposure to cigarette smoking adversily affects tidal flow volume in healthy newborn babies [84]. It was also observed that smoking during pregnancy causes altered height to weight ratio [85] in newborns. These findings support the underlying hypothesis of our study that second hand smoke may impact the development of fetal lungs in a highly critical phase of life, just a day before birth. In comparison to fetal lung metabolism, far fewer changes in blood plasma were observed, excluding the possibility that changes seen in fetal lungs were directly conferred by changes in blood plasma or by contamination of the fetal lung tissues with plasma. Indeed, only allo-inositol and hippuric acid, a product of gut metabolism for detoxifying aromatics were found decreased in both blood and fetal lungs.

## Conclusions

We have developed an improved way to visualize all detected metabolites in metabolomics studies (MetaMapp) that can comprise both identified and unknown compounds while maintaining the modular organization of metabolites in biochemical pathways. As MetaMapp outputs are seamlessly compatible with the open-source platform Cytoscape, visualization of next generation metabolomics datasets with an increased number of identified metabolites and integration with genomics and protemics data sets can be easily achieved. By applying this approach on metabolic responses of ETS in the lungs of dams and their respective unborn offspring (fetuses) as well as in blood plasma we have demonstrated that such network graphs enable rapid overviews on all statistically significant metabolic changes in different organs, including their biochemical context. The down-regulation of critical biochemical substrates in perinatal lung metabolism, most notably purines and pyrimidines, free fatty acids and specific amino acids, may lead to a compromised lung system impaired in a range of vital structural components such as surfactant proteins and lipids, connective tissues and alveolar membranes that are required to provide mechanical stability and elasticity to the pulmonary tissues. Hence, we propose that metabolic changes during this critical phase of development of a life supporting organ may affect lung morphogenesis which ultimately may lead to respiratory compromise and disease in later stages of life.

## Methods

### Environmental tobacco smoke exposure

Timed pregnant Sprague Dawley rats were purchased from Zivic Laboratories (Zeleniople, PA). Viviparous female rats were time-mated over a 12 hour window to insure a narrow gestational time among dams for this study. Conception was confirmed by the presence of a visible vaginal plug. Dams were shipped to the Center for Health and the Environment (UC Davis) on gestation day 3. Exposure to aged and diluted sidestream cigarette smoke as a surrogate to ETS was begun on gestation day 5. Dams were housed two per plastic cage on TEK-chip pelleted paper bedding using a 12 hours light/12 hours dark cycle. Animals during non-exposure hours had access to water and laboratory rodent diet 5001 (ad libitum). In compliance to Reporting *In Vivo* Experiments" (ARRIVE) guidelines, all animals were handled according to the U.S. Animal Welfare Acts, and all procedures were performed under the supervision of the University Animal Care and Use Committee (University of California, Davis). Dams were randomly divided into groups exposed to filtered air or to ETS in Hinners-type inhalation chambers. Humidified 3R4F research cigarettes (Lexington, KY) were used. An automatic metered puffer was used to smoke cigarettes under Federal Trade Commission conditions (35 ml puff, 2 seconds duration, 1 puff per minute). The smoke was collected in a chimney, diluted with fresh air and delivered to whole body exposure chambers. Exposure to smoke was for 6 hours/day, 7 days/week from gestation day 5 to gestation day 20 at a target concentration of total suspended particulate (TSP) of1mg/m^3Dams and their respective unborn offspring (fetuses) were studied at gestation day 20 of pregnancy.

### Metabolomic data acquisition and statistics

Rats were sacrificed one day before term (gestation day 19). Dams lungs were perfused using PBS (Phosphate buffer saline) while fetal lungs were too small for this procedure and contained residues of blood plasma. Lung tissues were prepared from liquid-nitrogen frozen status by grinding in 2 ml Eppendorf tubes for 2 minutes at 25 s-1 using 20 mm i.d. metal balls in a MM300 ball mill (RETSCH, Germany). Subsequent extraction was carried out using 1 ml of an one phase mixture of degassed isopropanol/acetonitrile/water (3:3:2) at −20°C for 5 min. Tubes were centrifuged for 30 s at 14,000 g and the supernatant was collected and concentrated to complete dryness. Samples were derivatized for GC-TOF-MS analysis as previously published [21]. BinBase database processing results and metabolite annotation matrices [21] were downloaded from the SetupX database [86], compliant to the recommendations by the metabolomics standards initiative (MSI); experiment id SX-394122. Reports contained deconvoluted mass spectra, retention indices, unique ions, standard compound identifiers and

compound names, class annotations and KEGG and PubChem compound identifiers (CIDs); see Additional file 1: Table S1. Data were used without further normalization. Student t-tests between ETS treated groups and filter aired groups were calculated in MS Excel 2007 using $p < 0.05$ as significance threshold. Fold changes were calculated by dividing the median of metabolites in ETS group by the median of metabolites in filter aired group. No multiple testing correction was applied as the focus here was not to identify biomarkers but to have a large set of potential changes to be highlighted in pathways. Directions of alteration were determined including fold changes. An output file was saved as a Cytoscape node attribute file for up-regulated, down-regulated and unchanged metabolites.

### Bioinformatics database queries

Nine small-molecule and pathway databases were queried in order to find the maximum substance coverage in each database and to obtain functional information. KEGG and PubChem identifiers of the identified metabolites were mapped against bioinformatics databases in batch modes using various web tools available as KEGG, PubChem, HMDB [31], MetaCyc [15] and ChEBI [54] websites. Additional databases such as Reactome, EHMN, and BIGG- UCSD were downloaded as BioPax or SBML file format. Mapping was performed in Cytoscape using the advance search option. Biochemical and chemical metabolic relationships between metabolites were utilized to construct metabolomics network graphs. Species-specific reaction networks were downloaded from their respective database, i.e. Reactome [50] (http://Reactome.org/download/index.html), HumanCyc (www.biocyc.org) and EHMN [52] (http://wwwtest.bioinformatics.ed.ac.uk/wiki/PublicCSB/EHMN) for mammalian reactions which were downloaded from their websites as systems biological markup language files (SBML). BIGG-UCSD [51] was received as SBML format from Dr. Palsson's laboratory, UCSD (http://systemsbiology.ucsd.edu/In_Silico_Organisms/Other_Organisms). Global reaction networks were constructed by parsing the reaction information from a text file downloaded from KEGG (ftp://ftp.genome.jp/pub/KEGG/ligand/reaction/reaction.lst) and MetaCyc databases (www.metacyc.org). Atomic mapping of reaction network was constructed by parsing the KEGG RPAIR text file (ftp://ftp.genome.jp/pub/KEGG/ligand/rpair/rpair). Parsed information was converted into Cytoscape SIF (simple interaction format) network file format and visualized in Cytoscape version 2.6. Metabolites-pathway relationships were extracted from a text file downloaded from the KEGG database (ftp://ftp.genome.jp/pub/kegg/pathway/map/cpd_map.tab). The information was converted into Cytoscape SIF file format. Results of KEGG pathway mapping for a given list of KEGG ids were converted into SIF file format using text pad, which is a useful text editor for windows operating systems.

### MetaMapp graph construction and cytoscape visualization

PubChem CIDs were utilized to obtain molfile encoded structures from PubChem using batch entrez online utility http://www.ncbi.nlm.nih.gov/sites/batchentrez. A 881 bit long substructure fingerprint is pre-calculated and stored for each compound entry in PubChem. Pair-wise Tanimoto chemical similarity co-efficients [71] among metabolites were calculated using the substructure fingerprints of input metabolites. The similarity co-efficient ranges between 0.0 and 1.0; high score reflects high similarity between two metabolites. Using online structure clustering tools of PubChem, pair wise matrices were subjected to a single linkage clustering algorithm that clustered the chemical compounds according to their chemical similarities. The similarity matrix was then downloaded from the website and converted into SIF formatted networks (Additional file 8: S8) using MetaMapp scripts using thresholds of 0.5, 0.6, 0.7, 0.8 and 0.9. A pair wise mass spectral similarity matrix was calculated by the BinBase database using the NIST similarity co-efficient. The matrix was subjected to a hierarchical clustering algorithm in the TMEV software. The mass spectral similarity network was constructed by MetaMapp scripts using 500, 600, 700, 800 and 900 similarity thresholds. Cytoscape was utilized to visualize the differential statistics output on network graphs. All the network graphs were imported into Cytoscape [87], and visualized using the 'organic layout' algorithm. Organic layout computes the node position in a graph on the basis of node degree and clustering co-efficient. An increase in clustering co-efficients means that the nodes are highly similar to each other, placing those nodes into a single cluster with short edges. As our objective was also to retrieve clusters of structurally similar metabolites, we have chosen to use the organic layout. Node and edge attributes were imported and mapped to nodes and edges. Statistical results were mapped as node color; fold changes were mapped as node size. All MetaMapp tools have been automated and can be accessed from http://metamapp.fiehnlab.ucdavis.edu.

## Additional Files

**Additional file 1: Table S1.** Results of Analysis of variance (ANOVA) for all 459 metabolites detected in the rat environmental tobacco smoke exposure study, compliant to MSI-recommendations (Metabolomics standard including international chemical identifier keys (InChI), PubChem and KEGG database identifiers and retention index and quantification ion information.
ETS = Environmental Tobacco smoke exposed, FA = filtered air exposed.

**Additional file 2: Table S2.** Bioinformatics databases that were queried for identified metabolites.

**Additional file 3: Table S3.** A list of web tools for pathway mapping analysis of a list of metabolites associated with KEGG or PubChem Identifiers.

**Additional file 4: Figure S4.** KEGG Atlas Global Map visualization. Mapped

metabolites are highlighted as black nodes yielding an overall sparse coverage of the graph. 45% of the identified metabolites in the rat ETS study were not covered by the KEGG Atlas Global Map.

**Additional file 5: Figure S5.** Mapping the 179 identified metabolites of the rat ETS study on biochemical network graphs using various publicly available tools and databases. See method section for details on construction of these graphs. (a) Edinburgh human metabolic network; (b) HumanCyc; (c) Reactome; (d) BIGG-UCSD; (e) KEGG RPAIR network; (f) MetaCyc reaction DB; (g) Cytoscape network using only the 137 metabolites retrieved from the KEGG pathway repository; (g) Cytoscape network of the 137 metabolites retrieved from the KEGG pathway on all metabolites comprised in the KEGG pathway repository.

**Additional file 6: Figure S6.** Impact of Tanimoto chemical similarity thresholds on visual appearance and clarity of metabolomic networks in Cytoscape without addition of KEGG reactant pair information. Nodes are metabolites and edges are chemical similarity links. (a) Network without any similarity threshold; (b) using a Tanimoto threshold of 0.9; (c) using a Tanimoto threshold of 0.8; (d) using a Tanimoto threshold of 0.7; (e) using a Tanimoto threshold of 0.6; (f) using a Tanimoto threshold of 0.5; (g) linking all metabolites to the two most Tanimoto-similar compounds; (h) combining data matrices from networks (d) and (g).

**Additional file 7: Figure S7.** A MetaMapp network graph displaying labels for all the identified metabolites. Nodes are metabolites, red edges are KEGG RPAIR links and blue edges denote chemical similarity links.

**Additional file 8: S8.** A zip file containing the input Tanimoto chemical similarity matrix, ANOVA output, KEGG Ids, CID pairs and Cytoscape session files for all 179 identified metabolites. The session files can be opened directly into Cytoscape.

### Competing interest

The authors declare that they have no conflict of interest.

### Authors' contributions

DKB performed all database queries and visualizations, analyzed and interpreted data, conceptualized and coded MetaMapp in R and wrote manuscript drafts. PDH programmed MetaMapp in JavaScript under supervision by GW who also programmed BinBase extensions to yield mass spectral similarity matrices. TK helped in all database- and substructure related queries and software tools. SLK aided in study design and interpretation. KEP designed the study in coordination with OF, specifically for smoke exposure regimes, supervised all animal work and wrote part of the manuscript. OF designed all aspects of computational studies in coordination with DKB, interpreted data, performed statistical analyses, and wrote and edited manuscript drafts. All authors read and approved the final manuscript.

### Acknowledgements

We acknowledge technical assistance for metabolomic data acquisition by Sevini Shahbaz. This research project was mainly supported by the National Institute of Health grant NIH R01 ES013932, and partially supported by U.S. Department of Defense (DOD) grant W81XWH-10-1-0635 and National Sciences Foundation grant MCB 1139644.

### Author details

[1]UC Davis Genome Center, Metabolomics, Davis 95616, CA, USA. [2]UC Davis Center for Health and the Environmental, Davis 95616, CA, USA. [3]DBT-Bioinformatics Infrastructure Facility, University of Rajasthan, Jaipur, India.

Received: 24 December 2011 Accepted: 25 April 2012
Published: 16 May 2012

### References

1. Mukhopadhyay P, Horn KH, Greene RM, Michele Pisano M: **Prenatal exposure to environmental tobacco smoke alters gene expression in the developing murine hippocampus.** *Reprod Toxicol* 2010, **29:**164–175.
2. Gilmour MI, Jaakkola MS, London SJ, Nel AE, Rogers CA: **How exposure to environmental tobacco smoke, outdoor air pollutants, and increased pollen burdens influences the incidence of asthma.** *Environ Health Perspect* 2006, **114:**627–633.
3. Gilliland FD, Berhane K, McConnell R, Gauderman WJ, Vora H, Rappaport EB, Avol E, Peters JM: **Maternal smoking during pregnancy, environmental tobacco smoke exposure and childhood lung function.** *Thorax* 2000, **55:**271–276.
4. Zhong CY, Zhou YM, Joad JP, Pinkerton KE: **Environmental tobacco smoke suppresses nuclear factor-kappaB signaling to increase apoptosis in infant monkey lungs.** *Am J Respir Crit Care Med* 2006, **174:**428–436.
5. Gairola CG, Wu H, Gupta RC, Diana JN: **Mainstream and sidestream cigarette smoke-induced DNA adducts in C7Bl and DBA mice.** *Environ Health Perspect* 1993, **99:**253–255.
6. Flouris AD, Metsios GS, Carrillo AE, Jamurtas AZ, Gourgoulianis K, Kiropoulos T, Tzatzarakis MN, Tsatsakis AM, Koutedakis Y: **Acute and short-term effects of secondhand smoke on lung function and cytokine production.** *Am J Respir Crit Care Med* 2009, **179:**1029–1033.
7. DiFranza JR, Aligne CA, Weitzman M: **Prenatal and postnatal environmental tobacco smoke exposure and children's health.** *Pediatrics* 2004, **113:**1007–1015.
8. Rehan VK, Asotra K, Torday JS: **The effects of smoking on the developing lung: insights from a biologic model for lung development, homeostasis, and repair.** *Lung* 2009, **187:**281–289.
9. Majeti R, Becker MW, Tian Q, Lee TL, Yan X, Liu R, Chiang JH, Hood L, Clarke MF, Weissman IL: **Dysregulated gene expression networks in human acute myelogenous leukemia stem cells.** *Proc Natl Acad Sci U S A* 2009, **106:**3396–3401.
10. Perez-Plasencia C, Vazquez-Ortiz G, Lopez-Romero R, Pina-Sanchez P, Moreno J, Salcedo M: **Genome wide expression analysis in HPV16 cervical cancer: identification of altered metabolic pathways.** *Infect Agent Cancer* 2007, **2:**16.
11. Fiehn O, Kopka J, Dormann P, Altmann T, Trethewey RN, Willmitzer L: **Metabolite profiling for plant functional genomics.** *Nat Biotechnol* 2000, **18:**1157–1161.
12. Xia J, Psychogios N, Young N, Wishart DS: **MetaboAnalyst: a web server for metabolomic data analysis and interpretation.** *Nucleic Acids Res* 2009, **37:**W652–660.
13. Kleijn RJ, Buescher JM, Le Chat L, Jules M, Aymerich S, Sauer U: **Metabolic fluxes during strong carbon catabolite repression by malate in Bacillus subtilis.** *J Biol Chem* 2010, **285:**1587–1596.
14. Lu X, Bennet B, Mu E, Rabinowitz J, Kang Y: **Metabolomic changes accompanying transformation and acquisition of metastatic potential in a syngeneic mouse mammary tumor model.** *J Biol Chem* 2010, **285:**9317–9321.
15. Caspi R, Altman T, Dale JM, Dreher K, Fulcher CA, Gilham F, Kaipa P, Karthikeyan AS, Kothari A, Krummenacker M, *et al*: **The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases.** *Nucleic Acids Res* 2010, **38:**D473–479.
16. Kanehisa M: **KEGG for linking genomes to life and the environment.** *Nucleic Acids Res* 2008, **36:**D480–D484.
17. Fiehn O: **Extending the breadth of metabolite profiling by gas chromatography coupled to mass spectrometry.** *Trends Analyt Chem* 2008, **27:**261–269.
18. Brauer MJ, Yuan J, Bennett BD, Lu W, Kimball E, Botstein D, Rabinowitz JD: **Conservation of the metabolomic response to starvation across two divergent microbes.** *Proc Natl Acad Sci U S A* 2006, **103:**19302–19307.
19. Ohashi Y, Hirayama A, Ishikawa T, Nakamura S, Shimizu K, Ueno Y, Tomita M, Soga T: **Depiction of metabolome changes in histidine-starved Escherichia coli by CE-TOFMS.** *Mol Biosyst* 2008, **4:**135–147.
20. Tolstikov VV, Fiehn O: **Analysis of highly polar compounds of plant origin: Combination of hydrophilic interaction chromatography and electrospray ion trap mass spectrometry.** *Anal Biochem* 2002, **301:**298–307.
21. Kind T, Wohlgemuth G, Lee do Y, Lu Y, Palazoglu M, Shahbaz S, Fiehn O: **FiehnLib: mass spectral and retention index libraries for metabolomics based on quadrupole and time-of-flight gas chromatography/mass spectrometry.** *Anal Chem* 2009, **81:**10038–10048.
22. Adams RP: *Identification of Essential Oil Components by Gas Chromatography/ Mass Spectroscopy*. IL, USA: Allured Publishing; 2007.
23. Fiehn O, Wohlgemuth G, Scholz M, Kind T, Lee do Y, Lu Y, Moon S, Nikolau B: **Quality control for plant metabolomics: reporting MSI-compliant studies.** *Plant J* 2008, **53:**691–704.
24. Denkert C, Budczies J, Weichert W, Wohlgemuth G, Scholz M, Kind T, Niesporek S, Noske A, Buckendahl A, Dietel M, Fiehn O: **Metabolite profiling**

of human colon carcinoma–deregulation of TCA cycle and amino acid turnover. *Mol Cancer* 2008, **7:**72.
25. Zhang B, Tolstikov V, Turnbull C, Hicks LM, Fiehn O: **Divergent metabolome and proteome suggest functional independence of dual phloem transport systems in cucurbits.** *Proc Natl Acad Sci U S A* 2010, **107:**13532–13537.
26. Hartman AL, Lough DM, Barupal DK, Fiehn O, Fishbein T, Zasloff M, Eisen JA: **Human gut microbiome adopts an alternative state following small bowel transplantation.** *Proc Natl Acad Sci U S A* 2009, **106:**17187–17192.
27. Shin MH, Lee do Y, Wohlgemuth G, Choi IG, Fiehn O, Kim KH: **Global metabolite profiling of agarose degradation by Saccharophagus degradans 2–40.** *N Biotechnol* 2010, **27:**156–168.
28. Seifert EL, Fiehn O, Bezaire V, Bickel DR, Wohlgemuth G, Adams SH, Harper ME: **Long-chain fatty acid combustion rate is associated with unique metabolite profiles in skeletal muscle mitochondria.** *PLoS One* 2010, **5:**e9834.
29. Adriaens ME, Jaillard M, Waagmeester A, Coort SL, Pico AR, Evelo CT: **The public road to high- quality curated biological pathways.** *Drug Discov Today* 2008, **13:**856–862.
30. Bader GD, Cary MP, Sander C: **Pathguide: a pathway resource list.** *Nucleic Acids Res* 2006, **34:**D504–506.
31. Wishart DS, Tzur D, Knox C, Eisner R, Guo AC, Young N, Cheng D, Jewell K, Arndt D, Sawhney S, Fung C, Nikolai L, Lewis M, Coutouly MA, Forsythe I, Tang P, Shrivastava S, Jeroncic K, Stothard P, Amegbey G, Block D, Hau DD, Wagner J, Miniaci J, Clements M, Gebremedhin M, Guo N, Zhang Y, Duggan GE, Macinnis GD, Weljie AM, Dowlatabadi R, Bamforth F, Clive D, Greiner R, Li L, Marrie T, Sykes BD, Vogel HJ, Querengesser L: **HMDB: The human metabolome database.** *Nucleic Acids Res* 2007, **35:**D521–526.
32. Green ML, Karp PD: **A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases.** *BMC Bioinforma* 2004, **5:**76.
33. Pico AR, Kelder T, van Iersel MP, Hanspers K, Conklin BR, Evelo C: **WikiPathways: pathway editing for the people.** *PLoS Biol* 2008, **6:**e184.
34. Khersonsky O, Tawfik DS: **Enzyme promiscuity: a mechanistic and evolutionary perspective.** *Annu Rev Biochem* 2010, **79:**471–505.
35. Babtie A, Tokuriki N, Hollfelder F: **What makes an enzyme promiscuous?** *Curr Opin Chem Biol* 2010, **14:**200–207.
36. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson BØ: **A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information.** *Mol Syst Biol* 2007, **3:**121.
37. Steuer R, Kurths J, Fiehn O, Weckwerth W: **Observing and interpreting correlations in metabolomic networks.** *Bioinformatics* 2003, **19:**1019–1026.
38. de la Fuente A, Bing N, Hoeschele I, Mendes P: **Discovery of meaningful associations in genomic data using partial correlation coefficients.** *Bioinformatics* 2004, **20:**3565–3574.
39. Camacho D, de la Fuente A, Mendes P: **The origin of correlations in metabolomics data.** *Metabolomics* 2005, **1:**53–63.
40. Atkinson HJ, Morris JH, Ferrin TE, Babbitt PC: **Using sequence similarity networks for visualization of relationships across diverse protein superfamilies.** *PLoS One* 2009, **4:**e4345.
41. Theocharidis A, van Dongen S, Enright AJ, Freeman TC: **Network visualization and analysis of gene expression data using BioLayout Express(3D).** *Nat Protoc* 2009, **4:**1535–1550.
42. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL: **The large-scale organization of metabolic networks.** *Nature* 2000, **407:**651–654.
43. Spirin V, Gelfand MS, Mironov AA, Mirny LA: **A metabolic network in the evolutionary context: multiscale structure and modularity.** *Proc Natl Acad Sci U S A* 2006, **103:**8774–8779.
44. Zhang Y, Li S, Skogerbo G, Zhang Z, Zhu X, Sun S, Lu H, Shi B, Chen R: **Phylophenetic properties of metabolic pathway topologies as revealed by global analysis.** *BMC Bioinforma* 2006, **7:**252.
45. Peregrin-Alvarez JM, Sanford C, Parkinson J: **The conservation and evolutionary modularity of metabolism.** *Genome Biol* 2009, **10:**R63.
46. Croes D, Couche F, Wodak SJ, van Helden J: **Metabolic PathFinding: inferring relevant pathways in biochemical networks.** *Nucleic Acids Res* 2005, **33:**W326–330.
47. Blank LM, Kuepfer L, Sauer U: **Large-scale 13C-flux analysis reveals mechanistic principles of metabolic network robustness to null mutations in yeast.** *Genome Biol* 2005, **6:**R49.
48. Kanehisa M, Goto S, Kawashima S, Nakaya A: **Thed KEGG databases at GenomeNet.** *Nucleic Acids Res* 2002, **30:**42–46.
49. Karp PD, Riley M, Saier M, Paulsen IT, Paley SM, Pellegrini-Toole A: **The EcoCyc and MetaCyc databases.** *Nucleic Acids Res* 2000, **28:**56–59.
50. Vastrik I, D'Eustachio P, Schmidt E, Gopinath G, Croft D, de Bono B, Gillespie M, Jassal B, Lewis S, Matthews L, *et al*: **Reactome: a knowledge base of biologic pathways and processes.** *Genome Biol* 2007, **8:**R39.
51. Schellenberger J, Park JO, Conrad TM, Palsson BO: **BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions.** *BMC Bioinforma* 2010, **11:**213.
52. Ma H, Sorokin A, Mazein A, Selkov A, Selkov E, Demin O, Goryanin I: **The Edinburgh human metabolic network reconstruction and its functional analysis.** *Mol Syst Biol* 2007, **3:**135.
53. Khersonsky O, Roodveldt C, Tawfik DS: **Enzyme promiscuity: evolutionary and mechanistic aspects.** *Curr Opin Chem Biol* 2006, **10:**498–508.
54. Degtyarenko K, de Matos P, Ennis M, Hastings J, Zbinden M, McNaught A, Alcantara R, Darsow M, Guedj M, Ashburner M: **ChEBI: a database and ontology for chemical entities of biological interest.** *Nucleic Acids Res* 2008, **36:**D344–350.
55. Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Bryant SH: **PubChem: a public information system for analyzing bioactivities of small molecules.** *Nucleic Acids Res* 2009, **37:**W623–633.
56. Kono N, Arakawa K, Ogawa R, Kido N, Oshita K, Ikegami K, Tamaki S, Tomita M: **Pathway projector: web-based zoomable pathway browser using KEGG atlas and Google Maps API.** *PLoS One* 2009, **4:**e7710.
57. Letunic I, Yamada T, Kanehisa M, Bork P: **iPath: interactive exploration of biochemical pathways and networks.** *Trends Biochem Sci* 2008, **33:**101–103.
58. Okuda S, Yamada T, Hamajima M, Itoh M, Katayama T, Bork P, Goto S, Kanehisa M: **KEGG Atlas mapping for global analysis of metabolic pathways.** *Nucleic Acids Res* 2008, **36:**W423–426.
59. Geer LY, Marchler-Bauer A, Geer RC, Han L, He J, He S, Liu C, Shi W, Bryant SH: **The NCBI BioSystems database.** *Nucleic Acids Res* 2010, **38:**D492–496.
60. Cottret L, Wildridge D, Vinson F, Barrett MP, Charles H, Sagot MF, Jourdan F: **MetExplore: a web server to link metabolomic experiments and genome-scale metabolic networks.** *Nucleic Acids Res* 2010, **38**(Suppl):W132–137.
61. Frolkis A, Knox C, Lim E, Jewison T, Law V, Hau DD, Liu P, Gautam B, Ly S, Guo AC, Xia J, Liang Y, Shrivastava S, Wishart DS: **SMPDB: The Small Molecule Pathway Database.** *Nucleic Acids Res* 2010, **38:**D480–D487.
62. Fiehn O, Kind T, Barupal DK: **Data Processing, Metabolomic Databases and Pathway Analysis**. In *Annual Plant Reviews Volume 43*. Edited by. : Wiley-Blackwell; 2011:367–406.
63. Kotera M, McDonald AG, Boyce S, Tipton KF: **Eliciting possible reaction equations and metabolic pathways involving orphan metabolites.** *J Chem Inf Model* 2008, **48:**2335–2349.
64. Pitkanen E, Jouhten P, Rousu J: **Inferring branching pathways in genome-scale metabolic networks.** *BMC Syst Biol* 2009, **3:**103.
65. Heath AP, Bennett GN, Kavraki LE: **Finding metabolic pathways using atom tracking.** *Bioinformatics* 2010, **26:**1548–1555.
66. Arita M: **In silico atomic tracing by substrate-product relationships in Escherichia coli intermediary metabolism.** *Genome Res* 2003, **13:**2455–2466.
67. Arita M: **The metabolic world of Escherichia coli is not small.** *Proc Natl Acad Sci* 2004, **101:**1543.
68. Mu F, Williams RF, Unkefer CJ, Unkefer PJ, Faeder JR, Hlavacek WS: **Carbon-fate maps for metabolic reactions.** *Bioinformatics* 2007, **23:**3193–3199.
69. Faust K, Croes D, van Helden J: **Metabolic pathfinding using RPAIR annotation.** *J Mol Biol* 2009, **388:**390–414.
70. Moriya Y, Shigemizu D, Hattori M, Tokimatsu T, Kotera M, Goto S, Kanehisa M: **PathPred: an enzyme-catalyzed metabolic pathway prediction server.** *Nucleic Acids Res* 2010, **38**(Suppl):W138–143.
71. Willett P, Barnard J, Downs G: **Chemical similarity searching.** *J Chem Inf Comput Sci* 1998, **38:**983–996.
72. Hummel J, Strehmel N, Selbig J, Walther D, Kopka J: **Decision tree supported substructure prediction of metabolites from GC-MS profiles.** *Metabolomics* 2010, **6:**322–333.
73. Stein SE: **Chemical substructure identification by mass spectral library searching.** *Journal of the American Society for Mass Spectrometry* 1995, **6:**644–655.
74. Varmuza K, Werther W: **Mass Spectral Classifiers for Supporting Systematic Structure Elucidation†.** *J Chem Inf Comput Sci* 1996, **36:**323–333.

75. Stein SE, Scott DR: **Optimization and testing of mass spectral library search algorithms for compound identification.** *Journal of the American Society for Mass Spectrometry* 1994, **5:**859–866.
76. Viscardi RM: **Role of fatty acids in lung development.** *J Nutr* 1995, **125:**1645S–1651S.
77. Maniscalco WM, Finkelstein JN, Parkhurst AB: **De novo fatty acid synthesis in developing rat lung.** *Biochimica et Biophysica Acta (BBA)-Lipids and Lipid Metabolism* 1982, **711:**49–58.
78. Gross I, Warshaw JB: **Enzyme activities related to fatty acid synthesis in developing mammalian lung.** *Pediatr Res* 1974, **8:**193–199.
79. Smith PMC, Atkins CA: **Purine biosynthesis. Big in cell division, even bigger in nitrogen assimilation.** *Plant Physiol* 2002, **128:**793–802.
80. Peden DB, Hohman R, Brown ME, Mason RT, Berkebile C, Fales HM, Kaliner MA: **Uric acid is a major antioxidant in human nasal airway secretions.** *Proc Natl Acad Sci U S A* 1990, **87:**7638–7642.
81. Fisher AB: **Intermediary metabolism of the lung.** *Environ Health Perspect* 1984, **55:**149–158.
82. Havens CG, Ho A, Yoshioka N, Dowdy SF: **Regulation of late G1/S phase transition and APC Cdh1 by reactive oxygen species.** *Mol Cell Biol* 2006, **26:**4701–4711.
83. Vizan P, Alcarraz-Vizan G, Diaz-Moralli S, Solovjeva ON, Frederiks WM, Cascante M: **Modulation of pentose phosphate pathway during cell cycle progression in human colon adenocarcinoma cell line HT29.** *Int J Cancer* 2009, **124:**2789–2796.
84. Lodrup Carlsen KC, Jaakkola JJ, Nafstad P, Carlsen KH: **In utero exposure to cigarette smoking influences lung function at birth.** *Eur Respir J* 1997, **10:**1774–1779.
85. Naeye RL: **Influence of maternal cigarette smoking during pregnancy on fetal and childhood growth.** *Obstet Gynecol* 1981, **57:**18–21.
86. Scholz M, Fiehn O: **SetupX–a public study design database for metabolomic projects.** *Pac Symp Biocomput* 2007, **12:**169–180.
87. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** *Genome Res* 2003, **13:**2498–2504.

doi:10.1186/1471-2105-13-99
**Cite this article as:** Barupal *et al.*: MetaMapp: mapping and visualizing metabolomic data by integrating information from biochemical pathways and chemical and mass spectral similarity. *BMC Bioinformatics* 2012 **13**:99.