CrossMark

IntLIM: integration using linear models of metabolomics and gene expression data

Jalal K. Siddiqui[1], Elizabeth Baskin[1], Mingrui Liu[1], Carmen Z. Cantemir-Stone[1], Bofei Zhang[1,5], Russell Bonneville[2,3], Joseph P. McElroy[4], Kevin R. Coombes[1] and Ewy A. Mathé[1*]

Abstract

Background: Integration of transcriptomic and metabolomic data improves functional interpretation of disease-related metabolomic phenotypes, and facilitates discovery of putative metabolite biomarkers and gene targets. For this reason, these data are increasingly collected in large (> 100 participants) cohorts, thereby driving a need for the development of user-friendly and open-source methods/tools for their integration. Of note, clinical/translational studies typically provide snapshot (e.g. one time point) gene and metabolite profiles and, oftentimes, most metabolites measured are not identified. Thus, in these types of studies, pathway/network approaches that take into account the complexity of transcript-metabolite relationships may neither be applicable nor readily uncover novel relationships. With this in mind, we propose a simple linear modeling approach to capture disease-(or other phenotype) specific gene-metabolite associations, with the assumption that co-regulation patterns reflect functionally related genes and metabolites.

Results: The proposed linear model, metabolite ~ gene + phenotype + gene:phenotype, specifically evaluates whether gene-metabolite relationships differ by phenotype, by testing whether the relationship in one phenotype is significantly different from the relationship in another phenotype (via a statistical interaction gene:phenotype *p*-value). Statistical interaction *p*-values for all possible gene-metabolite pairs are computed and significant pairs are then clustered by the directionality of associations (e.g. strong positive association in one phenotype, strong negative association in another phenotype). We implemented our approach as an R package, IntLIM, which includes a user-friendly R Shiny web interface, thereby making the integrative analyses accessible to non-computational experts. We applied IntLIM to two previously published datasets, collected in the NCI-60 cancer cell lines and in human breast tumor and non-tumor tissue, for which transcriptomic and metabolomic data are available. We demonstrate that IntLIM captures relevant tumor-specific gene-metabolite associations involved in known cancer-related pathways, including glutamine metabolism. Using IntLIM, we also uncover biologically relevant novel relationships that could be further tested experimentally.

Conclusions: IntLIM provides a user-friendly, reproducible framework to integrate transcriptomic and metabolomic data and help interpret metabolomic data and uncover novel gene-metabolite relationships. The IntLIM R package is publicly available in GitHub (https://github.com/mathelab/IntLIM) and includes a user-friendly web application, vignettes, sample data and data/code to reproduce results.

Keywords: Metabolomics, Transcriptomics, Linear Modeling, Integration

* Correspondence: ewy.mathe@osumc.edu
[1]Department of Biomedical Informatics, College of Medicine, The Ohio State University, Columbus, OH, USA
Full list of author information is available at the end of the article

Siddiqui *et al. BMC Bioinformatics* (2018) 19:81

Page 2 of 12

Background

Metabolomics data is increasingly collected in human biospecimens to identify putative biomarkers in diseases such as cancer [1–6]. Metabolites (small molecules < 1500 Da) are ideal candidates for biomarker discovery because they directly reflect disease phenotype and downstream effects of post-translational modifications [6]. However, interpretation of metabolomics data, including understanding how metabolite levels are modulated, is challenging. Reasons for this challenge include the presence of many (hundreds) of unidentified metabolites when untargeted approaches are applied [7, 8], and the fact that metabolomics profiles generated in human biospecimens are typically 'snapshots' or time-averaged representations of a disease state. Despite these difficulties, analyzing metabolomics data in light of other omics information, such as the transcriptome, can help to functionally interpret metabolomics phenotypes [9–15]. *Data integration,* or the use of multiple sources of information or data to provide a better model and understand a biological system [16], offers the opportunity to combine metabolomics data with other omics data-sets (e.g. transcriptome). Measurement and integration of the transcriptome and metabolome in the same cells, samples, or individuals, are thus increasingly applied to elucidate mechanisms that drive diseases, and to uncover putative biomarkers (metabolites) and targets (genes).

Current approaches that integrate transcriptomic and metabolomic data can be broadly categorized as numerical or pathway/network based. Numerical approaches include multivariate analyses (e.g. logistic regression, principal component analysis, partial least squares) and correlation-based approaches (e.g. canonical correlations) [17–19]. Differential correlation or coexpression methods have also been developed to capture changes in relationships between conditions [20]. Open-source tools, including MixOmics [21, 22] and DiffCorr [23], are available for integrating data but generally require in-depth statistical knowledge for their use and may not be as accessible to non-computational experts. Of note, such numerical approaches typically do not capture the complex and indirect relationships between transcripts and metabolites. For example, non-linear reaction kinetics mechanisms, metabolite-metabolite connections that regulate metabolite levels, and post-translational modifications all contribute to the complexity of gene-metabolite relationships [24, 25]. To better capture these complex relationships, pathway or network based approaches can be applied. Open-source tools such as Metaboanalyst [26], INMEX [27], XCMS Online [28], Metabox [29], and IMPALA [30] integrate transcriptomic and metabolomics data at a pathway level. One caveat of these approaches is that they rely on curated pathways or reaction-level information (knowledge of which enzymes produce a given metabolite) [18]. Pathway approaches are thus limited to metabolites that are identified and that can be mapped to pathways, which represents a fraction of what can be measured. In fact, of the 114,100 metabolites in the Human Metabolome Database [31–33], only 18,558 are detected and quantified, and of those, only 3115 (17%) map to KEGG pathways. Further, network approaches that attempt to study the complex many to many associations between genes and metabolites may not scale well when tens of thousands of gene-metabolite pairs are evaluated.

Importantly, previous studies have shown that functionally related genes and metabolites show coherent co-regulation patterns [20, 34, 35]. We make this functionality assumption here and propose a linear modeling approach for integrating metabolomics and transcriptomics data to identify phenotype-specific gene-metabolite relationships. Of note, typical numerical integration approaches uncover patterns of molecular features that are globally correlated or aim to predict phenotype [20]. However, these methods do not directly and statistically test whether associations between metabolites and gene expression differ by phenotype. This distinction is important because global associations between genes and metabolites may not only reflect one phenotype of interest, but could reflect other features (e.g., environment, histology). As for methods that uncover differentially correlated pairs between conditions [35], they either do not capture pairs of features that are correlated in one group and not correlated in another group, or they bin relationships into different types (e.g. positive correlation in one group, negative correlation in another group), thereby making it difficult to compare more than 2 phenotypes [20, 34, 35]. Further, these approaches are not implemented into user-friendly frameworks. Our approach is thus advantageous because it directly evaluates the relationship between genes and metabolites in the context of phenotype, it can easily incorporate potential covariates, and is applicable to categorical (> = 2 groups) or continuous phenotypes. Further, our approach is implemented as a publicly available R package IntLIM (Integration through Linear Modeling), available at our GitHub repository [36], which includes an R Shiny web interface making it user-friendly to non-computational experts. In the wake of increasing amounts of metabolomics and transcriptomic data generated, availability of open-source, user-friendly, and streamlined approaches is key for reproducibility. Using IntLIM, we evaluated phenotype-specific relationships between gene and metabolite levels measured in the NCI-60 cancer cell lines [10], and in tumor and adjacent non-tumor tissue of breast cancer patients [9]. We demonstrate that IntLIM is useful for uncovering known and novel gene-metabolite

Siddiqui *et al. BMC Bioinformatics* (2018) 19:81

Page 3 of 12

relationships (which would require further experimental validation).

Methods

NCI-60 cell line data pre-processing

The NCI-60 cancer cell line metabolomics (Metabolon platform) and gene expression data (Affymetrix U133 microarray) were downloaded from the Developmental Therapeutics Program (National Cancer Institute) website [10, 37]. Metabolomics and gene expression data, available in 57 cell lines, were pre-processed and normalized according to the Metabolon and Affymetrix MAS5 algorithms [38, 39], respectively. The metabolomics data contains 353 metabolites, of which 198 are unidentified. Each cell line is measured in triplicates (technical replicates), except for A498 and A549/ATCC, which had 4 and 2 technical replicates, respectively. The median of coefficients of variation (CVs) within technical replicate samples was calculated for each metabolite to assess consistency of abundance measurements. Metabolites with CVs < 0.3 were removed (280 metabolites remaining), abundances were log2 transformed, and the average technical replicate value was calculated for each metabolite. Next, the number of imputed values was estimated for each metabolite. The standard imputation method used by Metabolon is to impute missing values for a given metabolite by the minimum value of that metabolite across all samples. Thus, for each metabolite, the number of samples with a value equal to the minimum value (for that metabolite across all samples) minus "1" (one of those values is the true minimum value and should be subtracted) was used as an estimate of the number of missing values per metabolite. Metabolites with more than 80% imputed values were filtered out resulting in 220 metabolites, 111 of which are unidentified. Probes from the Chiron Affymetrix U133 microarrays were mapped to genes using the Bioconductor Ensembl database hgu133.plus.db [40]. In cases where more than one probe was matched to a given gene, the probe with the highest mean expression across all samples was retained for analysis, resulting in 17,987 genes with available expression. Lastly, we removed the 10% (arbitrary cutoff) of the lowest expressing genes, resulting in a total of 16,188 genes. For the linear modeling analyses, 220 metabolites and 16,188 genes were input.

For the NCI-60 cell line data, the phenotypes compared were leukemia cell lines vs. breast/prostate/ovarian (BPO) cell lines. Because this dataset was used to develop our approach, we purposefully chose cells from cancers that are known to be highly different in terms of their molecular profiles (e.g. blood cancer vs. solid tumor). The breast, prostate, and ovarian cancer cell lines were grouped together because they share susceptibility loci [41] and our aim was to increase sample size.

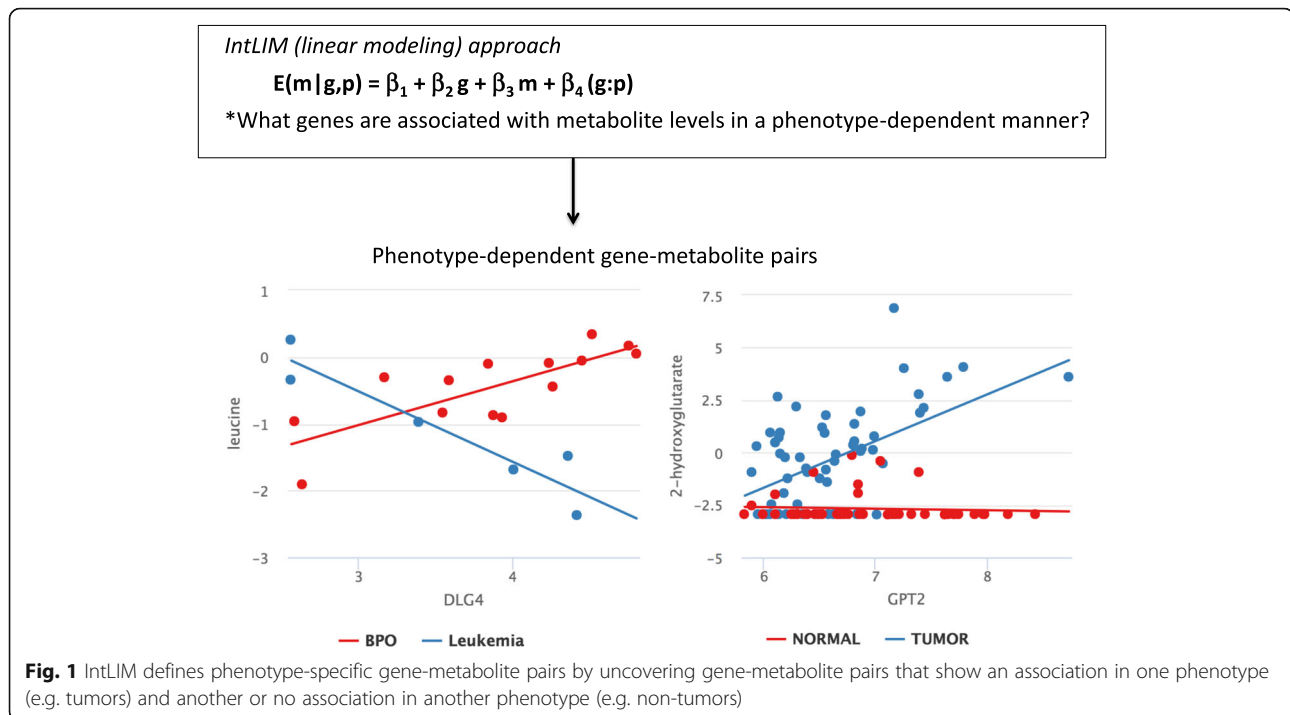### Breast cancer data pre-processing

Normalized gene expression (Affymetrix Gene Chip Human Gene 1.0 ST Arrays) and metabolomics (Metabolon) data in tumor and adjacent non-tumor tissue of breast cancer patients are publicly available through the Gene Expression Omnibus (GSE37751) and the supplementary data of the original publication, respectively [9, 42]. The data was normalized using the Metabolon algorithm (metabolites) and RMA algorithm [43] (genes), as previously described [9]. Both gene and metabolite levels are available for 61 tumor and 47 adjacent non-tumor breast tissue. The metabolomics data consists of 536 metabolites (203 of which are unidentified) in tumor and non-tumor tissue. Metabolites with more than 80% imputed values were removed, resulting in 379 metabolites, 119 of which are unidentified. Probes from the gene expression data not mapping to a gene symbol (Human Gene 1.0 ST Arrays) were removed. Similar to the NCI-60 data pre-processing, the probe with the highest mean expression was used for analysis when multiple probes mapped to a single gene. This resulted in 20,254 genes measured in tumor and non-tumor tissue. After removing the 10% lowest expression genes, we analyzed 18,228 genes. With this breast cancer data, our aim was to compare gene-metabolite associations between tumor and non-tumor tissue. A total of 379 metabolites and 18,228 genes were used for this analysis.

IntLIM: Integration through linear modeling approach

The linear model applied to integrate transcriptomic and metabolomic data is:

$$m = \beta_1 + \beta_2 g + \beta_3 p + \beta_4 (g : p) + \varepsilon \qquad (1)$$

where **"m"** and **"g"** are normalized (see data pre-processing above) and log2-transformed metabolite abundances and gene levels respectively, **"p"** is phenotype (e.g. cancer type, tumor vs. normal), **"(g:p)"** is the statistical interaction [44] between gene expression and phenotype, and **"ε"** is the error term that is assumed to be independent and normally distributed ($\varepsilon = N(0, \sigma)$) . A statistically significant two-tailed *p*-value of the **"(g:p)"** interaction term indicates that the slope relating gene expression and metabolite abundance is different from one phenotype compared to the other. Through this model, we can identify gene-metabolite associations that are specific to a particular phenotype (Fig. 1). This model has been applied to all possible gene-metabolite pairs including those involving unidentified metabolites in the publicly available NCI-60 cancer cell line data [10] as well as previously published data from a breast cancer study [9]. Two-tailed *p*-values are subsequently corrected for multiple comparisons using the method by

Siddiqui *et al. BMC Bioinformatics* (2018) 19:81

Page 4 of 12

Fig. 1 IntLIM defines phenotype-specific gene-metabolite pairs by uncovering gene-metabolite pairs that show an association in one phenotype (e.g. tumors) and another or no association in another phenotype (e.g. non-tumors)

Benjamini and Hochberg to control the false discovery rate (FDR) [45]. Gene-metabolite pairs with an FDR-adjusted interaction *p*-value less than 0.10 or 0.05 in the NCI-60 cell line and breast cancer data, respectively, were used to determine statistical significance. (Due to the larger sample size in the breast cancer data set and the much larger amount of significant gene-metabolite pairs, our threshold for significance was more stringent).

To filter and cluster the list of statistically significant gene-metabolite pairs, the difference in Spearman correlations between the two phenotypic groups being compared (leukemia vs. BPO for NCI-60 cells and tumor vs. non-tumor for breast cancer tissue) was used as an effect size. Volcano plots of the difference in Spearman correlations vs. the −log10 (FDR-adjusted *p*-values) are depicted to visualize the distributions and help determine appropriate *p*-value and effect size cutoffs (Additional file 1: Figure S3). For both datasets, a minimum absolute difference in correlations of 0.5 was used as an effect size cutoff.

The results can be visualized via a hierarchically clustered heatmap of gene-metabolite Spearman correlations calculated for each phenotypic group. Hierarchical clustering is performed with the *hclust* function. The Euclidean distance is used as the distance metric and the complete linkage method is used for agglomeration. The resulting dendrogram is used to create a heatmap that helps visualize how relevant gene-metabolite pairs cluster by their effect size (e.g. differences in Spearman correlation between the two phenotypic groups).

IntLIM R package

A pipeline has been developed in the form of an R package to streamline integration of metabolomics and gene expression data using IntLIM. The package has been optimized and can solve a high number of linear models (3–7 million gene-metabolite pairs) in 2 to 6 min on a laptop with 2.7GHz quad-core Intel Core i7 processor and 16 GB, 2133 MHz memory. Of note, IntLIM requires less than 3% of the time to solve all possible linear models compared to iterating through each model using the *lm* function in R for performing linear regression analysis as it contains a matrix algebra implementation of that function [46]. Extensive documentation is available in the package, including a vignette, and formatted NCI-60 and breast cancer datasets are linked and available in the IntLIM GitHub repository [36]. The steps for analysis are:

1) *Load data*: input CSV files containing normalized and log2-transformed gene expression data, normalized and log2-transformed metabolite abundance data, metadata for the samples (e.g. cancer status), and optionally metadata information on the genes and metabolites

2) *Filter data*: gene expression and metabolomics data are optionally filtered by gene and metabolite abundances and missing values

3) *Run IntLIM*: run linear models for all possible gene-metabolite pairs and extract FDR-adjusted interaction *p*-values and effect sizes (e.g. differences in slope/correlations between the groups)

Siddiqui *et al. BMC Bioinformatics* (2018) 19:81

Page 5 of 12

4) *Filter gene-metabolite pairs*: filter results by user-input cutoffs of FDR-adjusted p-values and effect size. A volcano plot (absolute difference in correlation vs. −log10(FDR-adjusted p-values) is shown to help users determine appropriate adjusted *p*-value and effect size cutoffs. Resulting pairs are then clustered with hierarchical clustering, based on correlations within each, and visualized through heatmaps.

5) *Visualize relevant gene-metabolite pairs*: user-selected gene-metabolite pairs can be visualized through scatterplots, color-coded by phenotypic groups of interest (e.g. leukemia vs. BPO, tumor vs. non-tumor).

The IntLIM package also includes an RShiny web interface, a powerful tool that transforms complex analysis pipelines into interactive, user-friendly web applications [47]. The App guides users through all steps available in the package, as mentioned above. Of note, most plots are coded in highcharter [48] or plotly [49, 50] so users can promptly assess the effect of changing parameters on analysis results (e.g. immediate updates of tables and plots resulting from user changes of effect size and p-value cutoffs). We believe this interactivity accelerates data analysis and hence discovery of phenotype-specific gene-metabolite pairs. Further the app makes the analysis accessible to non-computational researchers. More information can be found in the Additional file 2: IntLIM documentation.

Pathway analysis

Pathway and upstream regulator analyses were performed using the Ingenuity Pathway Analysis (IPA) software. The list of genes or identified metabolites from each cluster (e.g. highly correlated in one group but no correlation in the other) of statistically significant gene-metabolite pairs were input to conduct pathway analysis to analyze input genes or metabolites in the context of biological pathways or functions [51]. IPA also includes an upstream regulator analysis to determine whether those molecules were associated with a particular upstream regulator. *P*-values, calculated from the Right-tailed Fisher's Exact Test, reflect whether the number of overlapping molecules associated with a particular pathway or upstream regulator is greater than expected by chance [52]. For upstream regulator analysis, both direct and indirect relationships between molecules and their targets were considered (confidence = Experimentally observed). [53].

Results

IntLIM (integration through LInear modeling)

Our goal is to find gene-metabolite pairs that have a strong association in one phenotype (e.g. leukemia vs. breast/prostate/ovarian cancers (BPO), tumor vs. non-

tumor) and an inverse or no association in another phenotype. We anticipate that gene-metabolite relationships that are phenotype dependent will help interpret metabolomics phenotypes and will highlight molecular functions and pathways worth evaluating further. With accumulating transcriptomic and metabolomics data generated in the same samples, uncovering phenotype-specific relationships could elucidate novel co-regulation patterns. Because commonly leveraged untargeted metabolomics approaches produce large amounts of unidentified metabolites, approaches that rely on reaction-level or pathway annotations may not be sufficient to capture all or novel relationships. To accomplish our goal, we thus rely on numerical data integration and developed a linear modeling approach that predicts metabolite levels from gene expression in a phenotype-dependent manner (Fig. 1) (see Methods). Unlike correlation-based and logistic regression approaches, our approach specifically evaluates whether the association between gene and metabolite levels is related to a phenotype. Furthermore, it is important to keep in mind that metabolite abundances can be modulated by a group of enzymes, which in turn are regulated by a myriad of regulatory processes (e.g. transcription, post-translational modifications). Thus, gene expression, protein abundances, and metabolite levels do not always have linear relationships. While these more complex relationships will not be readily detected using our approach [14], co-regulated gene-metabolite relationships tend to share biological functions [34] and we make this assumption here. Our approach is implemented as an R package, which is publicly available through GitHub (See Methods and IntLIM Documentation in Additional file 2) [36].
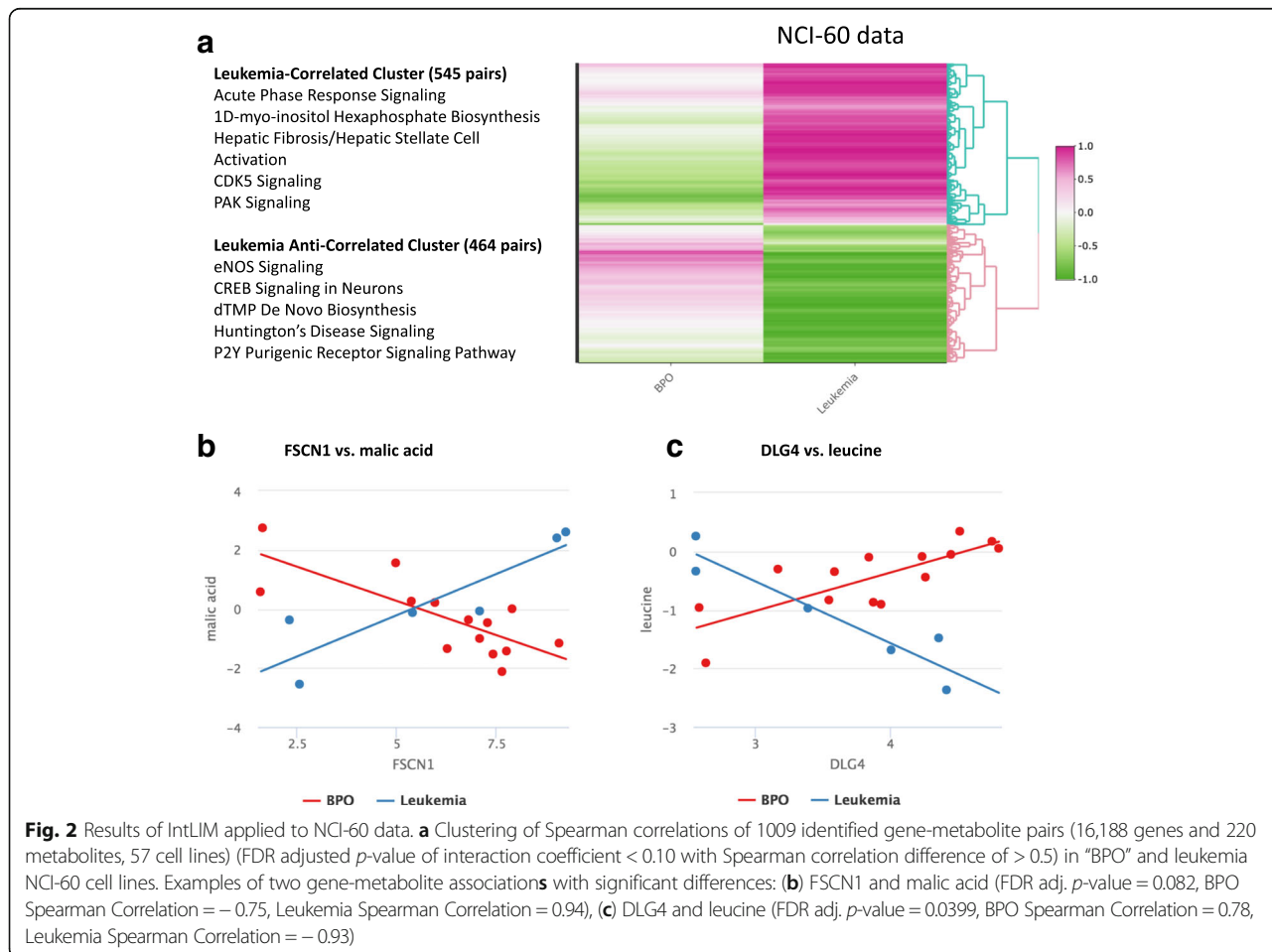
Application to NCI-60 data

The NCI-60 cell lines [10] were developed as a drug-screening tool focusing on a range of cancer types, including renal, colon, prostate, breast, ovarian, leukemia, and non-small cell lung cancer [54]. Transcriptomic (Affymetrix) and metabolomic (Metabolon) data are available for 57 of those cell lines [10] We applied IntLIM to identify cancer-type specific gene-metabolite associations. The two major subgroups compared were leukemia (6 cell lines: CCRF-CEM, HL-60 (TB), K-562, MOLT-4, RPMI-8226, SR) vs. the breast/prostate/ovarian (BPO) cancer cell lines (14 total cell lines: BT-549, DU-145, HS 578 T, IGROV1, MCF7, MDA-MB-231/ATCC, NCI/ADR-RES, OVCAR-3, OVCAR-4, OVCAR-5, OVCAR-8, PC-3, SK-OV-3, T-47D) consisting of 16,188 genes and 220 metabolites (see Methods). The latter cancers were grouped together as they share common susceptibility loci [41]. Unsupervised clustering using principal components analysis (PCA) on the log2-transformed and filtered metabolomics and gene expression data (Additional file 3: Figure S1A

Siddiqui *et al. BMC Bioinformatics* (2018) 19:81

Page 6 of 12

and B) clearly delineates the two major subgroups (Additional file 3: Figure S1C and D).

All possible combinations of gene-metabolite pairs (3,561,360 models run) were evaluated, using "BPO" and "leukemia" as cancer type. We identified 1009 cancer-type dependent gene-metabolite associations (FDR-adjusted *p*-value < 0.1 and correlation difference effect size > 0.5, Additional file 4: Data S1, Additional file 1: Figure S3A) involving 785 genes and 68 metabolites, of which 37 are unidentified. Clustering of these pairs by the direction of association (e.g. positive or negative correlation) within each cancer type subgroup revealed two major clusters (Fig. 3). First, the "leukemia correlated cluster" consists of 545 gene-metabolite pairs (429 unique genes and 54 unique metabolites of which 31 are unidentified) with relatively high positive correlations in leukemia cell lines and low or negative correlations in BPO cell lines (Fig. 2a). Second, the "leukemia anti-correlated cluster" consists of 464 gene-metabolite pairs (356 unique genes and 45 unique metabolites of which 24 are unidentified) with relatively high negative correlations in leukemia cell lines and positive or low negative correlations in BPO cell lines. Two of the top ranked gene-metabolite pairs (ranked in descending order of absolute value of Spearman correlation differences between BPO and leukemia) in the leukemia correlated and leukemia anti-correlated clusters are FSCN1-malic acid (Fig. 2b) and DLG4-leucine (Fig. 2c), respectively. FSCN1 and malic acid (Fig. 2b) are positively correlated in leukemia (*r* = 0.94) but negatively correlated in BPO cancers (*r* = − 0.75) (Fig. 2b). FSCN1 is associated with the progression of prostate cancer [55], while malic acid (or ionized malate) is an intermediate involved in glutamine metabolism pathways that play major roles in cancer metastasis [56, 57]. DLG4 and leucine (Fig. 2c) are negatively correlated in leukemia (*r* = − 0.92) but positively correlated (*r* = 0.78) in BPO cancers (Fig. 2c). DLG4 is downregulated in human cervical cancer cell lines infected with human papillomavirus and may act as a tumor suppressor [58], while leucine deprivation inhibits cell proliferation and induces apoptosis in breast cancer cells [59]. Interestingly, leucine supplementation has been shown to enhance pancreatic cancer growth in mouse models [60]. These opposing correlations of DLG4-leucine and FSCN1-malic acid between leukemia and BPO suggest possible tissue-specific relationships that can be differentially targeted.
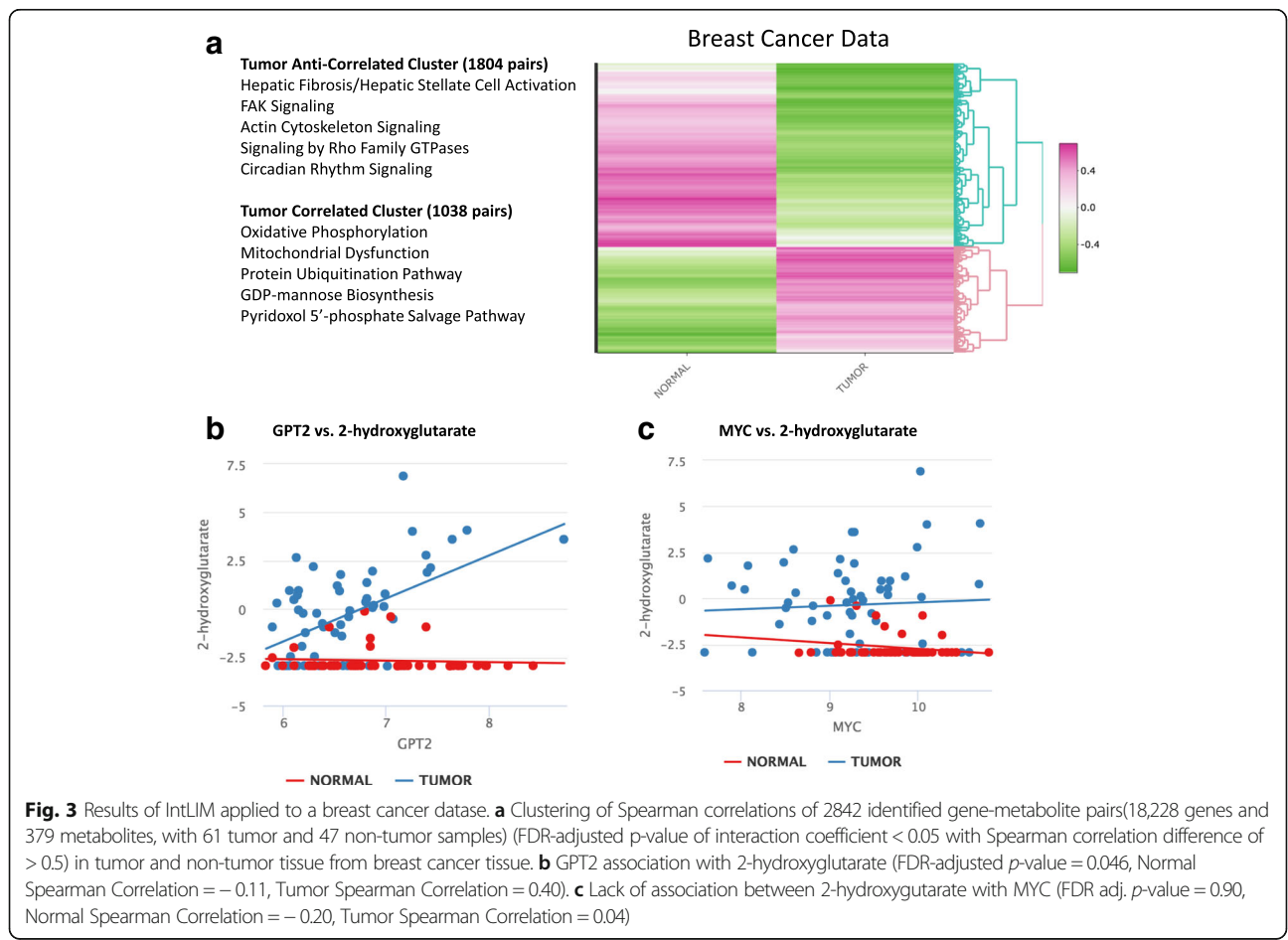

Fig. 2 Results of IntLIM applied to NCI-60 data. **a** Clustering of Spearman correlations of 1009 identified gene-metabolite pairs (16,188 genes and 220 metabolites, 57 cell lines) (FDR adjusted *p*-value of interaction coefficient < 0.10 with Spearman correlation difference of > 0.5) in "BPO" and leukemia NCI-60 cell lines. Examples of two gene-metabolite association**s** with significant differences: (**b**) FSCN1 and malic acid (FDR adj. *p*-value = 0.082, BPO Spearman Correlation = − 0.75, Leukemia Spearman Correlation = 0.94), (**c**) DLG4 and leucine (FDR adj. *p*-value = 0.0399, BPO Spearman Correlation = 0.78, Leukemia Spearman Correlation = − 0.93)

Siddiqui *et al. BMC Bioinformatics* (2018) 19:81

Page 7 of 12

Pathway analysis on 419 unique and mappable genes in the "leukemia correlated cluster" showed enrichment of the following pathways: acute phase response signaling, 1D–myo-inositol hexakisphosphate biosynthesis, hepatic fibrosis/hepatic stellate cell activation, CDK5 signaling, and PAK signaling (Additional file 5: Table S1). The "leukemia anti-correlated cluster" genes ($N = 351$) were enriched for endothelial NOS signaling, CREB signaling in neurons, dTMP de novo biosynthesis, Huntington's Disease signaling, and the P2Y purigenic receptor signaling pathway (Additional file 5: Table S1). Most of these pathways are relevant to cancer biology. For example, nitric oxide has been found to have both tumor suppressive (e.g. promoting apoptosis, inhibition of cancer growth) and tumor promoting properties (promotion of angiogenesis, DNA repair mechanisms) [61]. cAMP-regulator element binding protein (CREB) has been shown to be over-expressed and phosphorylated in several cancers (including acute myeloid leukemia) and might play a role in cancer pathogenesis [62]. These preliminary results demonstrate how different pathways may be differentially regulated in a cancer-type dependent manner. Since only 9 of 54 and 10 of 45

metabolites in the leukemia correlated and leukemia anti-correlated clusters, respectively, could be mapped to Human Metabolome Database (HMDB) IDs [31–33], pathway analyses were not possible for the metabolites.

Application to breast cancer data

We further applied IntLIM to a previously published breast cancer study [9]. Gene expression and metabolomics profiling of tumor ($n = 61$) and adjacent non-tumor tissue samples ($n = 47$) was measured in tissue from breast cancer patients [9]. Importantly, gene expression and metabolomics were measured in the same tissue biospecimens. The original study identified a relationship between MYC activation and 2-hydroxyglutarate (2-HG) accumulation as associated with poor prognosis in breast cancer [9]. Studies involving MYC overexpression and knockdown in human mammary epithelial and breast cancer cells further corroborated this relationship [9]. When assessing the relationship between MYC gene expression and 2-HG though, we did not observe this association at the transcription level (Fig. 3c). Our goal was thus to identify other potential regulators of 2-hydroxyglutarate accumulation in breast cancer tissue,

Fig. 3 Results of IntLIM applied to a breast cancer datase. **a** Clustering of Spearman correlations of 2842 identified gene-metabolite pairs(18,228 genes and 379 metabolites, with 61 tumor and 47 non-tumor samples) (FDR-adjusted p-value of interaction coefficient < 0.05 with Spearman correlation difference of > 0.5) in tumor and non-tumor tissue from breast cancer tissue. **b** GPT2 association with 2-hydroxyglutarate (FDR-adjusted *p*-value = 0.046, Normal Spearman Correlation = − 0.11, Tumor Spearman Correlation = 0.40). **c** Lack of association between 2-hydroxygutarate with MYC (FDR adj. *p*-value = 0.90, Normal Spearman Correlation = − 0.20, Tumor Spearman Correlation = 0.04)

Siddiqui *et al. BMC Bioinformatics* (2018) 19:81

Page 8 of 12

and to assess whether other gene-metabolite associations were specific to either tumor or non-tumor tissue. The data consists of 18,228 genes and 379 metabolites (119 unidentified) measured in 61 tumor samples and 47 adjacent non-tumor samples (Additional file 6: Figure S2A and B). Unsupervised clustering of gene and metabolite abundances separated tumor from non-tumor tissue (Additional file 6: Figure S2C and D).

IntLIM was applied to all possible combinations of gene-metabolite pairs (6,908,412 models), with tumor and non-tumor as the phenotype. Our approach identified 2842 tumor-dependent gene-metabolite correlations (FDR-adjusted interaction p-value < 0.05, and a Spearman correlation difference > 0.5) involving 761 genes and 212 metabolites of which 48 are unidentified. (Additional file 7: Data S2, Additional file 1: Figure S3B). The resulting heatmap of gene-metabolite Spearman correlations for tumor and non-tumor groups is divided into two major clusters (Fig. 3a). The first is a "tumor-correlated cluster" of 1038 gene-metabolite pairs (288 unique genes and 155 metabolites of which 35 are unidentified) with relatively high correlations in tumor samples and mostly negative correlations in non-tumor samples. The second major cluster, "tumor anti-correlated cluster", comprises 1804 gene-metabolite pairs (479 unique genes and 188 metabolites of which 39 are unidentified) with high negative correlations for tumor samples and mostly negative correlations for non-tumor samples.

Upstream analysis of the genes involved in the tumor-correlated cluster ($N = 283$) did identify MYC as an upstream transcriptional regulator (Right-tailed Fisher's Exact Test p-value = 6×10^{-3}), even though MYC and 2-HG are not differentially associated (Fig. 3c). 2-HG was, however, found to be associated with GPT2 (FDR adj p-value = 0.046, $r = 0.40$ in tumors, and $r = -0.11$ in non-tumors) (Fig. 3b, Additional file 7: Data S2). GPT2 plays a role in glutamine metabolism and encodes a glutamic-pyruvic transaminase that catalyzes reverse transamination between alanine and 2-oxoglutarate to generate pyruvate and glutamate [63]. Cancer cells exhibit a metabolic reprogramming that results in increased lactate acid production in the Warburg effect and the use of glutamine to replenish the tricarboxylic acid cycle (TCA) [64, 65]. The role of GPT2 serves to drive the utilization of glutamine as a carbon source for TCA analplerosis [63, 65]. While the exact mechanisms underlying increased levels of 2-hydroxyglutarate in breast cancer cells are not all known, our results suggest that metabolic reprogramming changes the relationship between GPT2 and 2-hydroxyglutarate. Furthermore, GPT2 is found to be in 18 (FDR adjusted p-value < 0.05 and correlation difference > 0.5) other tumor-specific gene-metabolite associations (Additional file 7: Data S2).

In addition to GPT2 and 2-HG, we identified 15 other gene-metabolite pairs involving metabolites linked to glutamine metabolism. Of those genes paired with glutamine, ASNS, which encodes asparagine synthetase, is directly involved in metabolizing glutamine [66] and SLC7A1 is involved in glutamine transport [64] (Additional file 7: Data S2). Furthermore, there are 65 gene-metabolite pairs with glutamate and 25 pairs involving alanine (Additional file 7: Data S2), and 5 gene-metabolite pairs involving the WIF gene, which is part of the Wnt signaling pathway [9] (Additional file 7: Data S2).

Pathway analysis revealed that genes in the "tumor-correlated cluster" (283 mapped into IPA out of 288 genes) were enriched for oxidative phosphorylation, mitochondrial dysfunction, protein ubiquitination pathway, GDP-mannose biosynthesis, and the pyridoxal 5′-phosphate salvage pathway (Additional file 8: Table S2). Genes in the "tumor anti-correlated cluster" (468 mapped onto IPA out of 479 genes) were enriched for hepatic fibrosis/hepatic stellate cell activation, FAK signaling, actin cytoskeleton signaling, signaling by Rho family GTPases, and circadian rhythm signaling (Additional file 8: Table S2). Expectedly, we find that pathways such as FAK signaling, actin cytoskeleton, the protein ubiquitination pathway, and circadian rhythm signaling have strong links to breast cancer pathogenesis [67–71]. Of note, the top two pathways in the tumor-correlated cluster (oxidative phosphorylation and mitochondrial dysfunction) play roles in cellular energetics [72].

Pathway analysis of the metabolites in the "tumor-correlated cluster" (100 mapped onto IPA out of 155 metabolites) resulted in enrichment of pathways related to tRNA charging and nucleotide degradation (Additional file 9: Table S3). The "tumor anti-correlated cluster" (115 mapped onto IPA out of 188 metabolites) was also enriched for tRNA charging, citrulline metabolism, urea cycle, purine nucleotide degradation, and purine ribonucleosides degradation to ribose-1-phosphate (Additional file 9: Table S3). Pathways related to tRNA and the urea cycle have been implicated in cancer [73–75]. Citrulline metabolism and the urea cycle have also been linked to glutamine metabolism [57, 76, 77]. These findings are consistent with previous studies [9, 57, 63, 64] that highlight the role of glutamine metabolism in cancer cell proliferation and maintenance, especially with regards to breast cancer [9]. Further, the urea cycle has been shown to be implicated in breast cancer and is linked to glutamine metabolism [77]. Notably, our IntLIM results identify 2 gene-metabolite pairs with urea and 5 gene-metabolite pairs with arginine (FDR-adjusted p-value of 0.05 or less, absolute Spearman Correlation difference > 0.5), a major metabolite in the urea cycle (Additional file 7: Data S2) [77].

Siddiqui *et al. BMC Bioinformatics* (2018) 19:81

Page 9 of 12

Discussion

As more and more transcriptomic and metabolomic data are collected in the same samples or individuals, there is a need for streamlined methods and associated user-friendly tools that integrate these data. We implemented a novel linear modeling approach into an IntLIM R package that includes a user-friendly web interface, to statistically test whether gene and metabolite associations differ by phenotype. Formally testing this dependency on phenotype differentiates our approach from other numerical integration approaches such as logistic regression and canonical correlations. Compared to other existing methods that take into account phenotype dependency [20, 34], IntLIM is user-friendly, it uses a well-developed methodology (linear model interactions), can easily account for other covariables (e.g. gender, BMI, etc.), and can be applied to phenotypes that have more than two categories or are continuous. Ultimately, uncovering phenotype-specific relationships can provide insight into how metabolites are being regulated by genes and on which pathways may be involved in these phenotype-specific changes.

While knowledge of relevant pathways is powerful in developing potential disease interventions and treatments, pathway enrichment analyses are hampered by the large fraction of metabolites that are identified or cannot be mapped to pathways. Importantly, IntLIM uncovers phenotype-dependent gene-metabolite associations without a priori curated information on pathways and networks, allowing discovery of potentially novel associations (that would require further experimental validation). Because untargeted metabolomics data produces many unidentified features, phenotype-specific associations with IntLIM could help further characterize these unidentified molecules. These data-driven discoveries would require further experimental validation and could generate new hypothesis to be tested. When pathway annotations are available though, pathway enrichment analysis of genes and metabolites that show similar patterns (e.g. positive correlation in tumors but no correlation in non-tumors) can offer greater insight onto pathways that are altered between phenotypes. With this in mind, IntLIM produces a list of relevant genes and metabolites that could be input into pathway integration approaches and software [26, 28–30].

To demonstrate the utility of IntLIM to uncover cancer-relevant gene-metabolite relationships, we evaluated transcriptomic and metabolomics data measured in the NCI-60 cell lines [10] and breast tumor/adjacent non-tumor tissue [9] (Figs. 2 and 3). In both these data sets, we uncovered biologically relevant gene-metabolite relationships and pathways. For example, glutamine metabolism clearly stood out as an altered pathway in the breast cancer data, in line with previous published results [9]. Interestingly, we also uncovered novel putative associations, such as the possible modulation by GPT2 of 2-hydroxyglutarate accumulation in breast cancer tissue (validation of this relationships would require further experimentation).

While this first iteration of IntLIM uncovers phenotype-specific gene-metabolite pairs, the approach can easily be extended to other omics data (e.g., metabolomics/microbiome data, metabolomics/proteome, proteome/transcriptome). Of note, because IntLIM makes use of a linear model, we assume that the independent variables (e.g. metabolite levels) are normally distributed to meet the normality assumption. We have verified the normality assumption in the NCI-60 and breast cancer datasets and leave it up to the user to appropriately transform and check the normality of their data prior to using IntLIM. Furthermore, our current linear model does not make use of the fact that some of the samples may be paired. In our breast cancer data [9], only a subset of the patients ($N = 41$) have both tumor and adjacent non-tumor available. It would be feasible to take into consideration the paired nature of the samples using a mixed model methodology, and thereby increase our power to detect significant relationships. Finally, future developments of IntLIM will accommodate greater flexibility in defining models. For example, we will include the capability of testing whether phenotype-specific gene-metabolite associations are independent of other putative confounders (e.g. age, gender, race, etc). Further, while IntLIM currently only supports a binary phenotype, it is readily generalizable to multicategorical phenotypes.

Like most approaches, IntLIM and the studies conducted are not without limitations. The biochemical pathways that drive gene expression to protein production to post-translational modifications to metabolite production/consumption are complex [24]. The abundance of a given metabolite typically depend on a group of enzymes that produce/consume that metabolite. Additionally, those enzymes have distinct kinetic parameters, and their activity depends on a range of posttranslational modifications and regulatory processes. As a result, transcript levels are not the only factors that modulate metabolite abundance, and the gene-metabolite relationship may not be linear. In this regard, IntLIM may not adequately capture these complex relationships. Nonetheless, linear-based approaches are well-developed, have successfully been applied when integrating omics data, and co-regulated genes and metabolites tend to be associated with functional roles [10, 20, 34]. Further, we demonstrate that this simple approach can identify biologically meaningful, putative phenotype-dependent gene-metabolite relationships that can be investigated with further experiments. Another limitation is that IntLIM does not take into consideration time-dependency of biochemical reaction steps, especially

Siddiqui *et al. BMC Bioinformatics* (2018) 19:81

Page 10 of 12

given the time delay between gene expression and protein production and further on metabolite production/consumption. However, in clinical and translational applications, metabolomic and transcriptomic data is typically collected at a "snapshot" in time, where time-dependent analyses are not possible [78]. Lastly, our approach, along with other numerical and pathway based integration approaches, does not take into account cellular heterogeneity in specimens analyzed, even though this heterogeneity could impact gene-metabolite correlations in different regions of cells or tissues [79]. Because IntLIM remains agnostic to the input, especially with regards to cell/tissue heterogeneity, it is the user's responsibility to interpret the data as well as design future experiments to test findings from results. Despite these limitations, IntLIM provides a user-friendly, reproducible framework to integrate metabolomics and transcriptomics data, or other omics data and provides a readily implementable first step in integration.

Conclusions

Metabolomics and transcriptomic data are increasingly collected in the same samples to uncover putative metabolite biomarkers and gene therapeutic targets. User-friendly approaches that integrate these data types will thus facilitate data interpretation in these studies, and could generate data-driven hypothesis. With this in mind, we developed a novel linear modeling approach that statistically tests whether gene-metabolite associations are specific to particular phenotypes (tumor vs. non-tumor, cancer-type, etc.). Our approach is available as a publicly available R package, IntLIM, with an associated user-friendly web application. We applied IntLIM to two cancer datasets and uncovered known and novel gene-metabolite pairs and pathways that were associated with cancer phenotypes. It is our hope that IntLIM will assist researchers, with or without computational expertise, in formulating novel hypothesis and proposing new studies especially with regards to the gene-metabolite pairs identified. Integrating the results with pathway analysis tools will provide further insight. The IntLIM R package and App are available for download via GitHub and a sample data-set and vignette are provided for users.

Additional files

Additional file 1: Figure S3. "Volcano plots" of Spearman correlation differences vs. FDR- adjusted *p*-values (of interaction term in linear model, see Methods) for A) NCI-60 cell line analysis and B) breast cancer data analysis. (PDF 604 kb)

Additional file 2: Documentation on IntLIM. (DOCX 1444 kb)

Additional file 3: Figure S1. Preliminary analysis of filtered NCI-60 data involving 14 breast/prostrate/ovarian cancer (BPO) lines and 6 leukemia cell lines with 220 filtered metabolites and 16,188 genes. A) Distribution

of normalized (Metabolon method) metabolite abundances among NCI-60 cell lines. B) Distribution of normalized (MAS5 algorithm) gene expression data. C, D) Principal component analysis of metabolomics and gene expression data, respectively. In the IntLIM package Rshiny app, these plots are interactive and hovering over points will provide information on those points (e.g. sample names). (PDF 555 kb)

Additional file 4: Data S1. NCI-60 Results with FDR Adjusted *p*-value < 0.10 and Correlation Difference > 0.50. (XLSX 104 kb)

Additional file 5: Table S1. NCI-60 Data pathway analysis results of genes. Ingenuity Pathway Analysis Canonical Pathways from Genes involved in Gene-Metabolite Pairs of the Leukemia Correlated Cluster and Leukemia Anti-Correlated Cluster. *P*-values are all calculated from right-tailed Fisher's Exact Test. (PDF 56 kb)

Additional file 6: Figure S2. Preliminary analysis of filtered breast cancer data involving 108 samples (61 tumor and 47 non-tumor) with 379 metabolites and 18,228 genes. A, B) Distribution of normalized metabolite levels (Metabolon method) and RMA-normalized gene expression levels for all samples, respectively. C,D) Principal component analysis of metabolomics and gene expression data, respectively. In the IntLIM package Rshiny app, these plots are interactive and hovering over points will provide information on those points (e.g. sample names). (PDF 543 kb)

Additional file 7: Data S2. Breast Cancer Results. FDR Adjusted *p*-value < 0.05 and Spearman Correlation > 0.5. (XLSX 220 kb)

Additional file 8: Table S2. Breast Cancer Data pathway analysis results of genes. Ingenuity Pathway Analysis Canonical Pathways from Genes involved in Gene-Metabolite Pairs of the Tumor Correlated Cluster and Tumor Anti-Correlated Cluster. *P*-values are all calculated from right-tailed Fisher's Exact Test. (PDF 55 kb)

Additional file 9: Table S3. Breast Cancer Data pathway analysis results of metabolites. Ingenuity Pathway Analysis Canonical Pathways from Metabolites involved in Gene-Metabolite Pairs of the Tumor Correlated Cluster. *P*-values are all calculated from right-tailed Fisher's Exact Test. (PDF 53 kb)

Authors' contributions
JKS helped design study, conducted analyses, and developed the software. EB assisted with conducting analyses and design of software. ML helped with design and development of the software. CZC assisted with analyses of results. BZ assisted with developing the software. RB assisted with developing software and with conducting analyses. JPM and KRC helped analyze and interpret results and offered suggestions for manuscript. EM designed study, conducted analyses, and developed the software. All authors read and approved the final manuscript.

Ethics approval and consent to participate
Not applicable

Siddiqui *et al. BMC Bioinformatics* (2018) 19:81

Page 11 of 12

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Author details

[1]Department of Biomedical Informatics, College of Medicine, The Ohio State University, Columbus, OH, USA. [2]Biomedical Sciences Graduate Program, The Ohio State University, Columbus, OH, USA. [3]Comprehensive Cancer Center, Department of Internal Medicine, The Ohio State University, Columbus, OH, USA. [4]Center for Biostatistics, The Ohio State University, Columbus, OH, USA. [5]Biomedical Engineering Undegraduate Program, The Ohio State University, Columbus, OH 43210, USA.

References

1. Mathé EA, Patterson AD, Haznadar M, Manna SK, Krausz KW, Bowman ED, Shields PG, Idle JR, Smith PB, Anami K, Kazandjian DG, Hatzakis E, Gonzalez FJ, Harris CC. Noninvasive urinary metabolomic profiling identifies diagnostic and prognostic markers in lung cancer. Cancer Res. 2014;74(12):3259-70
2. Halama A, Riesen N, Möller G, Hrabě de Angelis M, Adamski J. Identification of biomarkers for apoptosis in cancer cell lines using metabolomics: tools for individualized medicine. J Intern Med. 2013;274(5):425–39.
3. Van Ravenzwaay B, Cunha GC-P, Leibold E, Looser R, Mellert W, Prokoudine A, Walk T, Wiemer J. The use of metabolomics for the discovery of new biomarkers of effect. Toxicol Lett. 2007;172(1):21–8.
4. Haznadar M, Cai Q, Krausz KW, Bowman ED, Margono E, Noro R, Thompson MD, Mathé EA, Munro HM, Steinwandel MD, Gonzalez FJ, Blot WJ, Harris CC. Urinary metabolite risk biomarkers of lung cancer: a prospective cohort study. Cancer Epidemiol Biomarkers Prev. 2016;25(6):978-86.
5. Gummer J, Banazis M, Maker G, Solomon P, Oliver R, Trengove R. Use of mass spectrometry for metabolite profiling and metabolomics. Aust. Biochemist. 2009;40(3):5–8.
6. Zhang A, Sun H, Yan G, Wang P, Wang X. Metabolomics for biomarker discovery: moving to the clinic. Biomed Res Int. 2015;2015:354671.
7. Goveia J, Pircher A, Conradi LC, Kalucka J, Lagani V, Dewerchin M, Eelen G, DeBerardinis RJ, Wilson ID, Carmeliet P. Meta-analysis of clinical metabolic profiling studies in cancer: challenges and opportunities. EMBO Mol. Med. 2016;8(10):1134–42.
8. Schrimpe-Rutledge AC, Codreanu SG, Sherrod SD, McLean JA. Untargeted metabolomics strategies—challenges and emerging directions. J Am Soc Mass Spectrom. 2016;27(12):1897–905.
9. Terunuma A, Putluri N, Mishra P, Mathé EA, Dorsey TH, Yi M, Wallace TA, Issaq HJ, Zhou M, Killian JK, et al. MYC-driven accumulation of 2-hydroxyglutarate is associated with breast cancer prognosis. J Clin Invest. 2014;124(1):398–412.
10. Su G, Burant CF, Beecher CW, Athey BD, Meng F. Integrated metabolome and transcriptome analysis of the NCI60 dataset. BMC bioinformatics. 2011;12(1):S36.
11. Stempler S, Yizhak K, Ruppin E. Integrating transcriptomics with metabolic modeling predicts biomarkers and drug targets for Alzheimer's disease. PLoS One. 2014;9(8):e105383.
12. Budhu A, Roessler S, Zhao X, Yu Z, Forgues M, Ji J, Karoly E, Qin LX, Ye QH, Jia HL. Integrated metabolite and gene expression profiles identify lipid biomarkers associated with progression of hepatocellular carcinoma and patient outcomes. Gastroenterology. 2013;144(5):1066–75. e1061
13. Zhang G, He P, Tan H, Budhu A, Gaedcke J, Ghadimi BM, Ried T, Yfantis HG, Lee DH, Maitra A. Integration of metabolomics and transcriptomics revealed a fatty acid network exerting growth inhibitory effects in human pancreatic cancer. Clin Cancer Res. 2013;19(18):4983–93.
14. Hakimi AA, Reznik E, Lee C-H, Creighton CJ, Brannon AR, Luna A, Aksoy BA, Liu EM, Shen R, Lee W. An integrated metabolic atlas of clear cell renal cell carcinoma. Cancer Cell. 2016;29(1):104–16.
15. Ren S, Shao Y, Zhao X, Hong CS, Wang F, Lu X, Li J, Ye G, Yan M, Zhuang Z. Integration of metabolomics and transcriptomics reveals major metabolic pathways and potential biomarker involved in prostate cancer. Mol Cell Proteomics. 2016;15(1):154–63.
16. Gomez-Cabrero D, Abugessaisa I, Maier D, Teschendorff A, Merkenschlager M, Gisel A, Ballestar E, Bongcam-Rudloff E, Conesa A, Tegnér J. Data integration in the era of omics: current and future challenges. BMC. 2014; 8(Suppl 2):I1.
17. Rui W, Feng Y, Jiang M, Liang Wang X, Shi ZZ. Pattern recognition of Glycyrrhiza uralensis Metabonomics on rats with MixOmics package of R software. Procedia Eng. 2011;24:510–4.
18. Wanichthanarak K, Fahrmann JF, Grapov D. Genomic, proteomic, and metabolomic data integration strategies. Biomark Insights. 2015;10(Suppl 4):1.
19. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype-phenotype interactions. Nat Rev Genet. 2015;16(2):85.
20. Siska C, Bowler R, Kechris K. The discordant method: a novel approach for differential correlation. Bioinform. 2015;32(5):690–6.
21. Lê Cao K-A, González I, Déjean S. integrOmics: an R package to unravel relationships between two omics datasets. Bioinformatics. 2009;25(21):2855–6.
22. Rohart F, Gautier B, Singh A, Le Cao K-A. mixOmics: an R package for 'omics feature selection and multiple data integration. BioRxiv. 2017; https://doi.org/10.1101/108597.
23. Fukushima A. DiffCorr: an R package to analyze and visualize differential correlations in biological networks. Gene. 2013;518(1):209–14.
24. Zelezniak A, Sheridan S, Patil KR. Contribution of network connectivity in determining the relationship between gene expression and metabolite concentration changes. PLoS Comput Biol. 2014;10(4):e1003572.
25. Buescher JM, Driggers EM. Integration of omics: more than the sum of its parts. Cancer Metab. 2016;4:4.
26. Xia J, Mandal R, Sinelnikov IV, Broadhurst D, Wishart DS. MetaboAnalyst 2.0—a comprehensive server for metabolomic data analysis. Nucleic Acids Res. 2012;40(W1):W127–33.
27. Xia J, Fjell CD, Mayer ML, Pena OM, Wishart DS, Hancock RE. INMEX—a web-based tool for integrative meta-analysis of expression data. Nucleic Acids Res. 2013;41(W1):W63–70.
28. Huan T, Forsberg EM, Rinehart D, Johnson CH, Ivanisevic J, Benton HP, Fang M, Aisporna A, Hilmers B, Poole FL. Systems biology guided by XCMS online metabolomics. Nat Methods. 2017;14(5):461–2.
29. Wanichthanarak K, Fan S, Grapov D, Barupal DK, Fiehn O. Metabox: a toolbox for metabolomic data analysis, interpretation and integrative exploration. PLoS One. 2017;12(1):e0171046.
30. Kamburov A, Cavill R, Ebbels TM, Herwig R, Keun HC. Integrated pathway-level analysis of transcriptomics and metabolomics data with IMPaLA. Bioinform. 2011;27(20):2917–8.
31. Wishart DS, Jewison T, Guo AC, Wilson M, Knox C, Liu Y, Djoumbou Y, Mandal R, Aziat F, Dong E. HMDB 3.0—the human metabolome database in 2013. Nucleic Acids Res. 2013;41(D1):D801–7.
32. Wishart DS, Knox C, Guo AC, Eisner R, Young N, Gautam B, Hau DD, Psychogios N, Dong E, Bouatra S. HMDB: a knowledgebase for the human metabolome. Nucleic Acids Res. 2008;37(suppl_1):D603–10.
33. Wishart DS, Tzur D, Knox C, Eisner R, Guo AC, Young N, Cheng D, Jewell K, Arndt D, Sawhney S. HMDB: the human metabolome database. Nucleic Acids Res. 2007;35(suppl_1):D521–6.
34. Bradley PH, Brauer MJ, Rabinowitz JD, Troyanskaya OG. Coordinated concentration changes of transcripts and metabolites in Saccharomyces cerevisiae. PLoS Comput Biol. 2009;5(1):e1000270.
35. Kayano M, Shiga M, Mamitsuka H. Detecting differentially coexpressed genes from labeled expression data: a brief review. IEEE/ACM Trans. Comput. Biol. Bioinform. 2014;11(1):154–67.
36. IntLIM: Integration through LInear Modeling. https://github.com/Mathelab/IntLIM/. Accessed 20 Nov 2017.
37. Molecular Target Data - NCI DTP Data - National Cancer Institute - Confluence Wiki. https://wiki.nci.nih.gov/display/ncidtpdata/molecular+target+data. Accessed 1 June 2016.
38. Affymetrix I: Statistical algorithms description document. Technical paper 2002.
39. Hubbell E, Liu W-M, Mei R. Robust estimators for expression analysis. Bioinform. 2002;18(12):1585–92.
40. hgu133plus2.db. http://bioconductor.org/packages/hgu133plus2.db/. Accessed 1 June 2016.
41. Kar SP, Beesley J, Al Olama AA, Michailidou K, Tyrer J, Kote-Jarai Z, Lawrenson K, Lindstrom S, Ramus SJ, Thompson DJ. Genome-wide meta-

Siddiqui *et al. BMC Bioinformatics* (2018) 19:81

Page 12 of 12

analyses of breast, ovarian, and prostate cancer association studies identify multiple new susceptibility loci shared by at least two cancer types. Cancer Discov. 2016;6(9):1052–67.

42. Supplemental Table. https://www.jci.org/articles/view/71180/sd/2. Accessed 15 Feb 2017.

43. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res. 2003;31(4):e15.

44. Boulesteix A-L, Janitza S, Hapfelmeier A, Van Steen K, Strobl C. Letter to the editor: on the term 'interaction'and related phrases in the literature on random forests. Brief Bioinform. 2014;16(2):338–45.

45. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Methodol. 1995;57(1):289–300.

46. Object-Oriented Microarray and Proteomics Analysis. http://oompa.r-forge.r-project.org. Accessed 1 July 2017.

47. Shiny by RStudio. https://shiny.rstudio.com. Accessed 10 Jan 2017.

48. Highcharter. http://jkunst.com/highcharter/. Accessed 5 June 2017.

49. Sievert C, Parmer C, Hocking T, Chamberlain S, Ram K, Corvellec M, Despouy P. Plotly: create interactive web graphics via'plotly. Js'. R package version. 2016;3:6.

50. Plotly: Modern Visualization for the Data Era. https://plot.ly. Accessed 5 June 2017.

51. Ingenuity Pathway Analysis - QIAGEN Bioinformatics. https://www.qiagenbioinformatics.com/?qia-storyline=products/ingenuity-pathway-analysis. Accessed 15 May 2017.

52. Calculating and interpreting the p-values for functions, pathways, and lists in IPA. Ingenuity Systems White Paper. 2009. http://qiagen.force.com/KnowledgeBase/servlet/fileField?entityId=ka1D00000008j11IAA&field=Article_Attachment__Body__s.

53. Krämer A, Green J, Pollard Jr J, Tugendreich S. Causal analysis approaches in ingenuity pathway analysis. Bioinform. 2013;30(4):523–30.

54. Shoemaker RH. The NCI60 human tumour cell line anticancer drug screen. Nat Rev Cancer. 2006;6(10):813.

55. Fuse M, Nohata N, Kojima S, Sakamoto S, Chiyomaru T, Kawakami K, Enokida H, Nakagawa M, Naya Y, Ichikawa T. Restoration of miR-145 expression suppresses cell proliferation, migration and invasion in prostate cancer by targeting FSCN1. Int J Oncol. 2011;38(4):1093–101.

56. Gaude E, Frezza C. Defects in mitochondrial metabolism and cancer. Cancer metabolism. 2014;2(1):10.

57. Altman BJ, Stine ZE, Dang CV. From Krebs to clinic: glutamine metabolism to cancer therapy. Nat Rev Cancer. 2016;16(10):619.

58. Handa K, Yugawa T, Narisawa-Saito M, Ohno S-i, Fujita M, Kiyono T. E6AP-dependent degradation of DLG4/PSD95 by high-risk human papillomavirus type 18 E6 protein. J Virol. 2007;81(3):1379–89.

59. Xiao F, Wang C, Yin H, Yu J, Chen S, Fang J, Guo F. Leucine deprivation inhibits proliferation and induces apoptosis of human breast cancer cells via fatty acid synthase. Oncotarget. 2016;7(39):63679–89.

60. Liu KA, Lashinger LM, Rasmussen AJ, Hursting SD. Leucine supplementation differentially enhances pancreatic cancer growth in lean and overweight mice. Cancer metabolism. 2014;2(1):6.

61. Weiming X, Liu LZ, Loizidou M, Ahmed M, Charles IG. The role of nitric oxide in cancer. Cell Res. 2002;12(5):311–20.

62. Sakamoto KM, Frank DA. CREB in the pathophysiology of cancer: implications for targeting transcription factors for cancer therapy. Clin Cancer Res. 2009;15(8):2583–7.

63. Hao Y, Samuels Y, Li Q, Krokowski D, Guan B-J, Wang C, Jin Z, Dong B, Cao B, Feng X. Oncogenic PIK3CA mutations reprogram glutamine metabolism in colorectal cancer. Nat Commun. 2016;7:11971.

64. Dang CV. Rethinking the Warburg effect with Myc micromanaging glutamine metabolism. Cancer Res. 2010;70(3):859–62.

65. Smith B, Schafer XL, Ambeskovic A, Spencer CM, Land H, Munger J. Addiction to coupling of the Warburg effect with glutamine catabolism in cancer cells. Cell Rep. 2016;17(3):821–36.

66. Van Heeke G, Schuster SM. The N-terminal cysteine of human asparagine synthetase is essential for glutamine-dependent activity. J Biol Chem. 1989;264(33):19475–7.

67. Orlowski RZ, Dees EC. The role of the ubiquitination-proteasome pathway in breast cancer: applying drugs that affect the ubiquitin-proteasome pathway to the therapy of breast cancer. Breast Cancer Res. 2002;5(1):1.

68. Blakeman V, Williams JL, Meng Q-J, Streuli CH. Circadian clocks and breast cancer. Breast Cancer Res. 2016;18(1):89.

69. Tai Y-L, Chen L-C, Shen T-L. Emerging roles of focal adhesion kinase in cancer. Biomed Res Int. 2015;2015:690690.

70. Jiang P, Enomoto A, Takahashi M. Cell biology of the movement of breast cancer cells: intracellular signalling and the actin cytoskeleton. Cancer Lett. 2009;284(2):122–30.

71. Yamaguchi H, Condeelis J. Regulation of the actin cytoskeleton in cancer cell migration and invasion. Biochim et Biophys Acta (BBA)-Mol. Cell Res. 2007;1773(5):642–52.

72. Osellame LD, Blacker TS, Duchen MR. Cellular and molecular mechanisms of mitochondrial function. Best Pract Res Clin Endocrinol Metab. 2012;26(6):711–23.

73. Kim S, You S, Hwang D. Aminoacyl-tRNA synthetases and tumorigenesis: more than housekeeping. Nat Rev Cancer. 2011;11(10):708–18.

74. Aird KM, Zhang R. Nucleotide metabolism, oncogene-induced senescence and cancer. Cancer Lett. 2015;356(2):204–10.

75. Nagamani SC, Erez A. A metabolic link between the urea cycle and cancer cell proliferation. Mol. Cell. Oncol. 2016;3(2):e1127314.

76. Marini JC, Didelija IC, Castillo L, Lee B. Glutamine: precursor or nitrogen donor for citrulline synthesis? Am J. Physiol-Endocrinol. Metab. 2010; 299(1):E69–79.

77. Tea I, Martineau E, Antheaume I, Lalande J, Mauve C, Gilard F, Barillé-Nion S, Blackburn AC, Tcherkez G. 13C and 15N natural isotope abundance reflects breast cancer cell metabolism. Sci Rep. 2016;6:34251.

78. Roullier C, Bertrand S, Blanchet E, Peigné M, Robiou du Pont T, Guitton Y, Pouchus YF, Grovel O. Time dependency of chemodiversity and biosynthetic pathways: an LC-MS metabolomic study of marine-sourced penicillium. Mar. Drugs. 2016;14(5):103.

79. Repsilber D, Kern S, Telaar A, Walzl G, Black GF, Selbig J, Parida SK, Kaufmann SH, Jacobsen M. Biomarker discovery in heterogeneous tissue samples-taking the in-silico deconfounding approach. BMC bioinform. 2010;11(1):27.