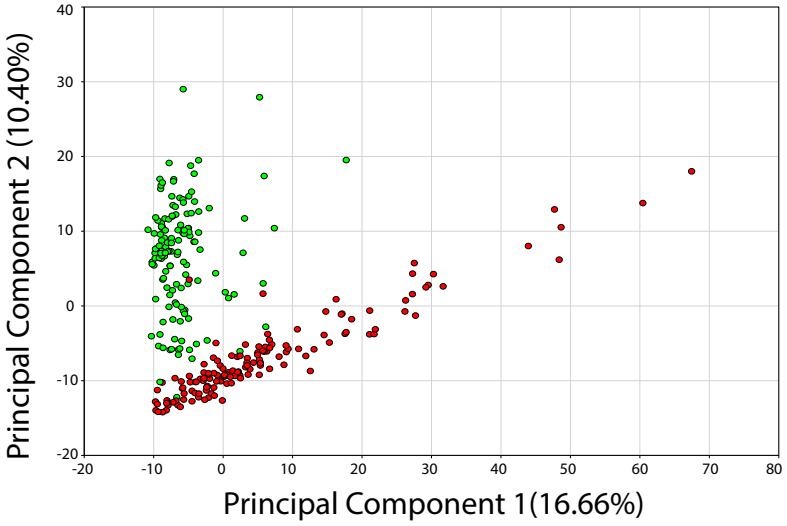
Supplemental Data

A



B

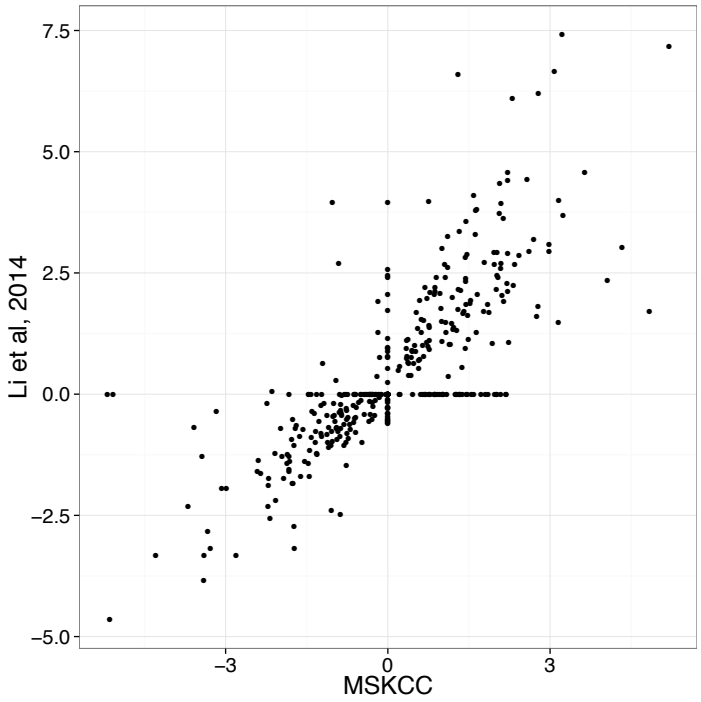


Figure S1, related to Figure 1. (A) Principal component analysis of metabolomics data. Color of dots indicates tumor (red) and normal (green) samples within the MSKCC TKCRP ccRCC metabolomics cohort. **(B)** Comparison of changes in metabolite abundance between tumor and normal tissue from the current study and from Li et al, 2014. Of the 575 metabolites measured in Li et al 2014, 425 were able to be matched to metabolites measured in the current study. Each dot represents one such metabolite. Each axis indicates the log2 ratio of the abundance of a metabolite between tumor and normal tissue (X-axis, Li et al, 2014, Y-axis, MSKCC). Metabolites that exhibited a statistically insignificant change in abundance (q value > 0.05) had fold change set to zero. A statistically significant correlation was observed between the two studies (Spearman correlation 0.76, p value < 2e-16).

Table S1, related to Figure 1. Clinical characteristics of the cohort.

Median Age (quartiles)	63 (55,70)
Gender	Female – 38 (27.5%) Male – 100 (72.5%)
Race	White -122 (88.4%) Black – 7 (5.1%) Asian – 8 (5.8%) Other – 1 (0.7%)
Median Tumor size - cm (quartiles)	4.5(3.5,7.5)
Primary tumor (T Stage)	
pT1	40 (28.9%)
pT2	13 (9.4%)
pT3	81 (58.7%)
pT4	4 (2.9%)
Regional lymph nodes (N Stage)	
pNx	65 (47.1%)
pN0	68 (49.3%)
pN1	5 (3.6%)
Distant metastases at presentation (M Stage)	
M0	118 (85.5%)
M1	20 (14.5%)
AJCC Stage	
1	38 (27.5%)
2	10 (7.2%)
3	70 (50.7%)
4	20 (14.5%)
Fuhrman Nuclear Grade	
2	52 (37.7%)
3	67 (48.6%)
4	19 (13.8%)
Median Followup for survivors (months)	59.8
Overall 5-year survival	81.2%
Metastasis at presentation	14%
Recurrent Disease	14%
Number of deaths	26
Number of death from RCC	18

Table S2 related to Figure 1. Description of metabolites quantified in this study. Provided as an MS Excel file.

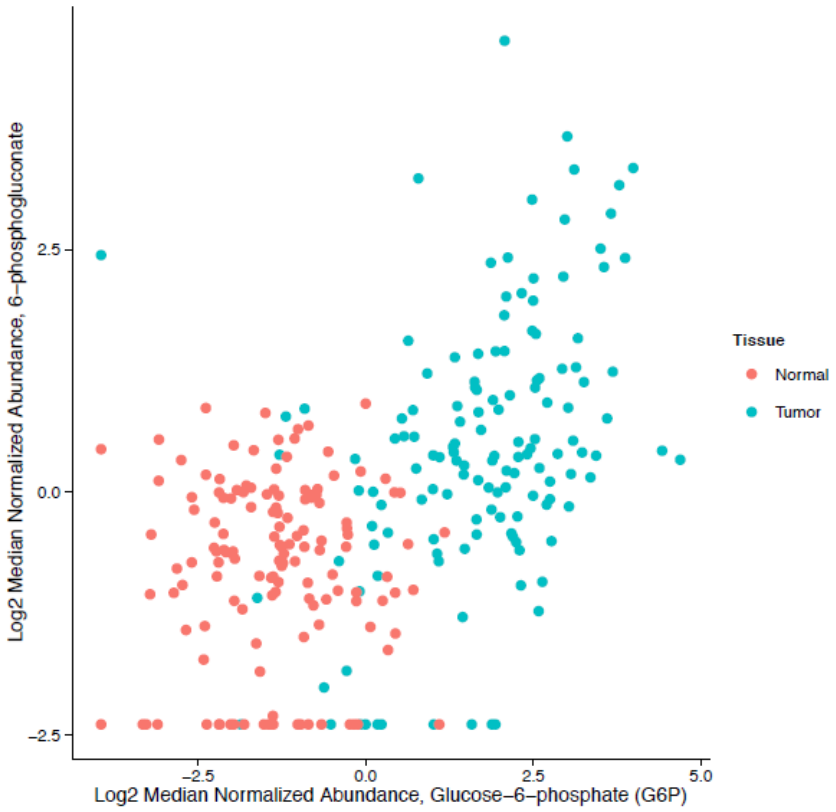


Figure S2, related to Figure 2. The abundances of glucose-6-phosphate and 6-phosphogluconate in tumor (blue, Spearman rho p value 4e-7) and normal samples (red, Spearman rho p value 0.91)

Table S3 related to Figure 2. Results of differential abundance tests for metabolites in this study, comparing tumors to normal tissues. Provided as an MS Excel file.

Table S4 related to Figure 3. Results of differential abundance tests for metabolites in this study, comparing late-stage (III, IV) to early-stage (I, II) tumors. Provided as an MS Excel file.

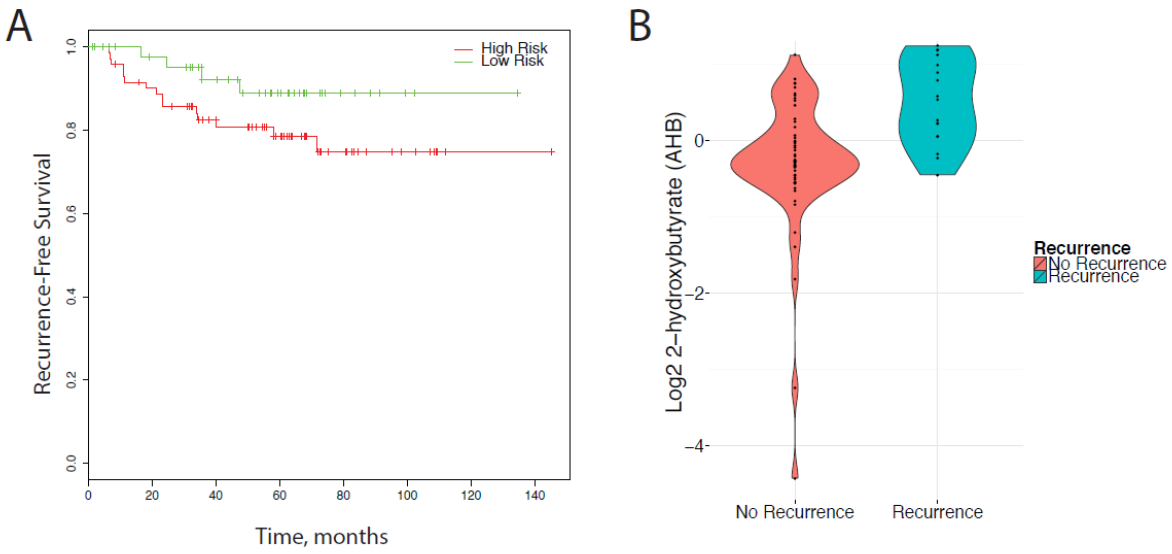
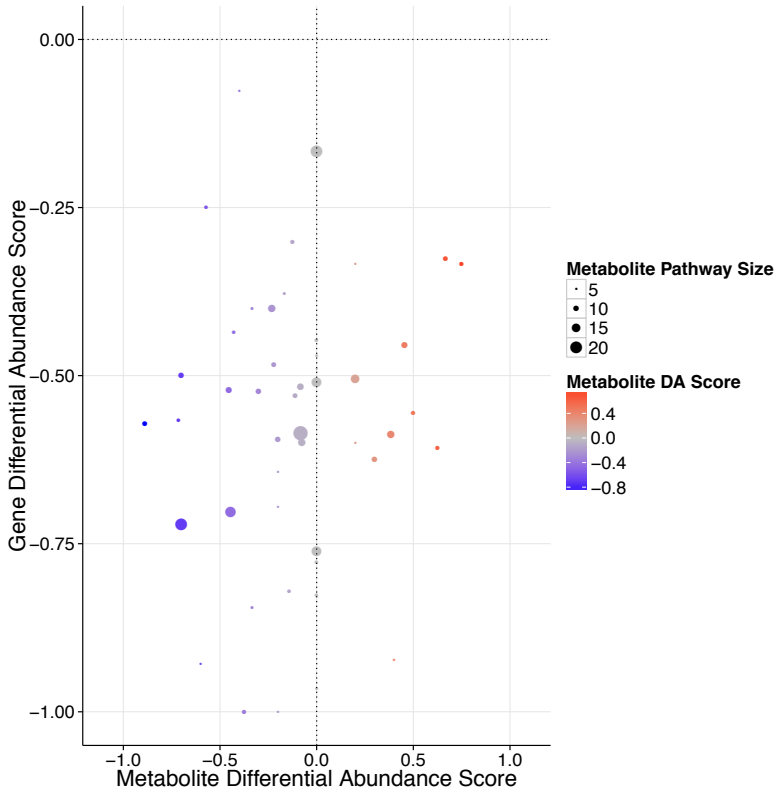


Figure S3, related to Figure 4. (A) Recurrence-free survival of high-risk (mCluster 2,3,4) versus low-risk (mCluster 1) groups for Stage I-III patients. (B) Violin plot of the abundance of 2-hydroxybutyrate (AHB) in Stage III patients who developed recurrent disease (blue) and Stage III patients who did not develop recurrent disease (red).

Table S5 related to Figure 4. Results of differential abundance tests, comparing individual mClusters against each other. Provided as an MS Excel file.

Table S6 related to Figure 4. Results of differential abundance tests, comparing tumors that metastasized to those that did not at preparation of report. Provided as an MS Excel file.

A


0.00
−0.25
−0.50
−0.75
−1.00
Gene Differential Abundance Score
−1.0
−0.5
0.0
0.5
1.0
Metabolite Differential Abundance Score
Metabolite Pathway Size
5
10
15
20
Metabolite DA Score
0.4
0.0
−0.4
−0.8

B

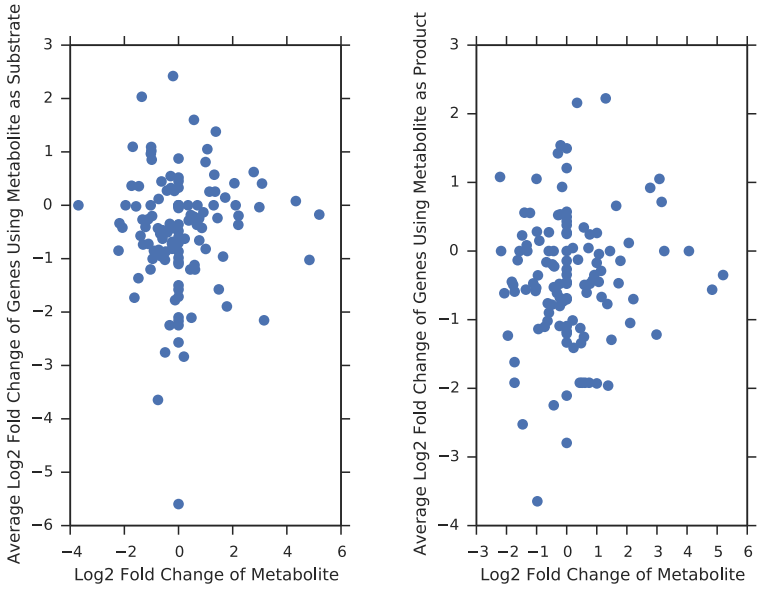
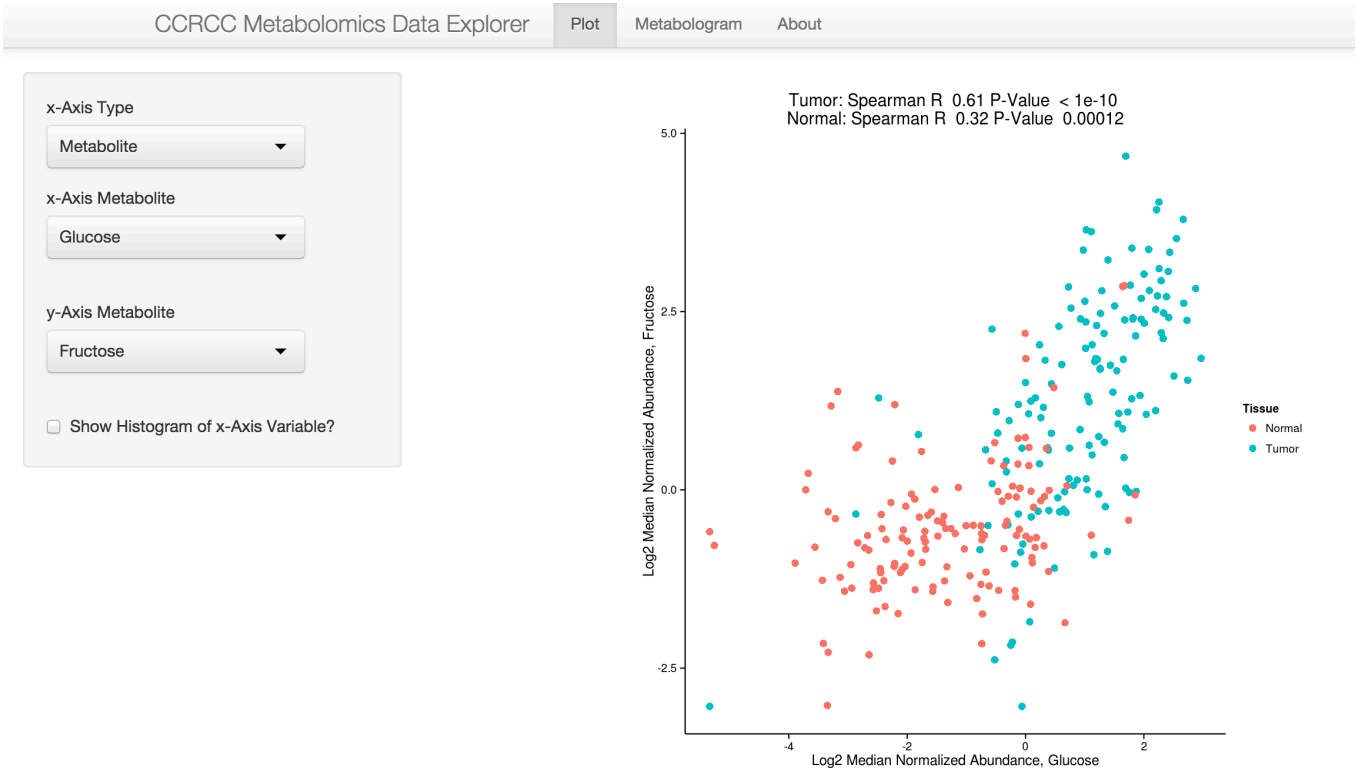

Average Log2 Fold Change of Genes Using Metabolite as Substrate
Log2 Fold Change of Metabolite
Average Log2 Fold Change of Genes Using Metabolite as Product
Log2 Fold Change of Metabolite

Figure S4, related to Figure 5 (A) Comparison of differential abundance scores for metabolomic and transcriptomic data, using only the 10 mCluster 2 tumor samples from the MSKCC cohort. **(B)** Comparison of log-fold change of metabolomics and transcriptomics data at a detailed level. Each dot corresponds to a single metabolite, with the x-axis value indicating its log2 fold change, tumor relative to normal tissue. The y-axis indicates the mean log2 fold change of genes using that metabolite. Plot on the left indicates genes that use metabolite as a substrate, while plot on the right indicates genes that produce metabolite as a product. Analysis is restricted to reactions annotated as irreversible in the Recon2 human metabolic reconstruction.

A


CCRCC Metabolomics Data Explorer
Plot
Metabologram
About
x-Axis Type
Metabolite
x-Axis Metabolite
Glucose
y-Axis Metabolite
Fructose
Show Histogram of x-Axis Variable?
Tumor: Spearman R 0.61 P-Value < 1e-10
Normal: Spearman R 0.32 P-Value 0.00012
Log2 Median Normalized Abundance, Fructose
Log2 Median Normalized Abundance, Glucose
Tissue
Normal
Tumor

B

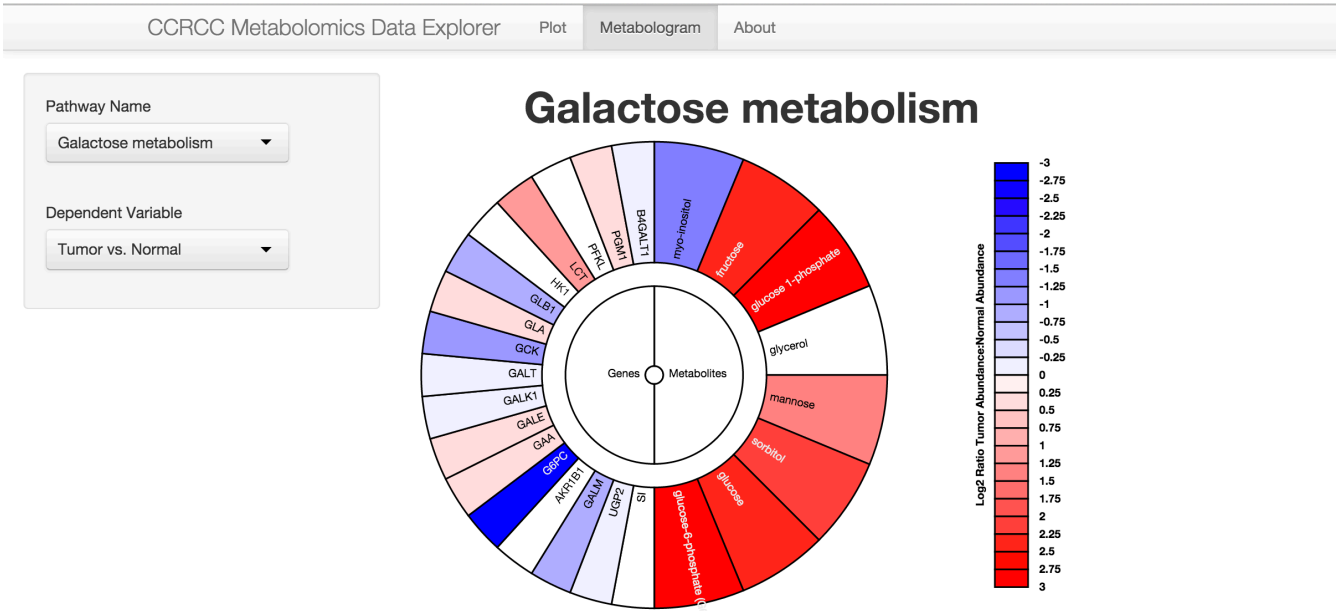

CCRCC Metabolomics Data Explorer
Plot
Metabologram
About
Pathway Name
Galactose metabolism
Dependent Variable
Tumor vs. Normal
Galactose metabolism
Genes
Metabolites
myo-inositol
fructose
glucose 1-phosphate
glycerol
mannose
sorbitol
glucose
glucose-6-phosphate
Log2 Ratio Tumor Abundance:Normal Abundance

Figure S5, related to Figure 6. Exploration of the "Metabologram" Data Portal. **(A)** Scatter-plot view. One can compare log2 median normalized abundance values of two metabolites or compare metabolite abundance to clinical variables (e.g. gender, patient age, etc.) between normal and tumor samples. Shown is a comparison between glucose and fructose abundance. **(B)** Pathway view. One can visualize metabolite variation between tumor and normal samples or tumor stage for given metabolic pathways. Shown are changes in the galactose metabolism pathway changes in both transcripts and metabolites when comparing tumors to adjacent normal kidney tissues.

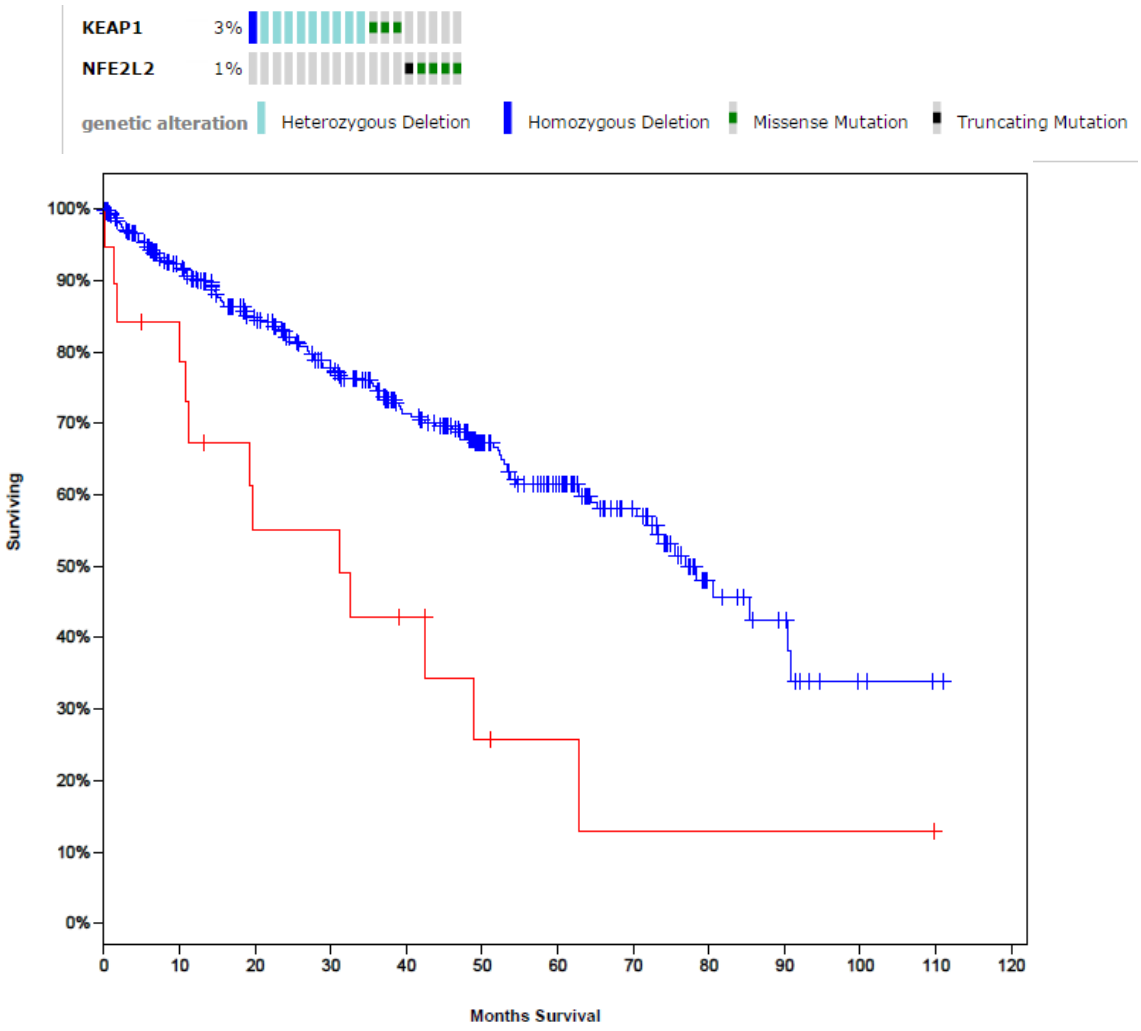


Figure S6, related to Figure 7. *KEAP1* loss and *NFE2L2* mutations in the TCGA KIRC cohort. Kaplan-Meier plot indicates the survival of patients with (red) or without (blue) alterations in *KEAP1* and *NFE2L2* (generated using cBioportal, log rank p value 0.003).

Table S7, related to Figure 7. Comparison of metabolic genes-based clusters (rClusters) and the published KIRC TCGA mRNA clusters.

	rCluster A	rCluster B	rCluster C	rCluster D
MSKCC: High Glutathione	9	1	0	0
TCGA Cluster M1	0	85	24	23
TCGA Cluster M2	22	39	13	10
TCGA Cluster M3	41	2	39	8
TCGA Cluster M4	17	21	8	36

Table S8 related to Figure 7. Results of gene set analysis for rCluster A. Provided as an MS Excel file.

Supplemental Experimental Procedures

Comparison of Metabolomic and Transcriptomic Changes at the Individual Reaction Level

Recon2 (a curated, genome-scale model of human metabolism) was used to identify all pairs of genes/metabolites such that the gene product uses the metabolite as a substrate or produces it as a product in a reaction annotated in Recon2 as irreversible. Having isolated these pairs, we calculated for each metabolite (1) the average fold change (tumor vs. normal) of all genes using the metabolite as a substrate, and (2) the average fold change (tumor vs. normal) of all genes using the metabolite as a product. We compared these values to the fold change of the metabolite itself.

RNA-Seq Alignment and Metabolic Gene Expression Clustering

In order to compare transcriptomic data from our cohort to that from the TCGA (Figure 7), we re-aligned sequencing reads from the TCGA using the same pipeline as applied to the MSKCC RNA-Seq data. First, for TCGA RNA-Seq data, raw output BAMs were converted back to FASTQ using PICARD Sam2Fastq. Then, for both MSKCC and TCGA samples, reads were mapped to the human genome using rnaStar. The genome used was HG19 with junctions from ENSEMBL (GRCh37.69_ENSEMBL) and a read overhang of 49. Gene level counts were computed using htseq-count and the same gene models as used in the mapping step. Principal components analysis was used to analyze the resulting dataset for potential batch effects separating the TCGA and MSKCC samples, and no batch effects were evident.

For metabolic gene expression clustering, counts were normalized and rescaled into log2 counts using the limma R package (Anders et al., 2013). A list of metabolic genes were extracted from the Recon2 Human Metabolic Network Reconstruction. Gene expression profiles for the list of metabolic genes were extracted from the transcript profiles described above, restricting ourselves to genes which were expressed at a level of at least 16 reads per sample. The resulting dataset consisted of 1,506 unique metabolic genes across 488 samples.

Log_2 normalized counts were clustered using consensus clustering via the ConsensusClusterPlus R package using K-means clustering with Euclidean distances. Rank estimation was performed

for k = 2...6 clusters, randomly subsampling 80% of samples and 80% of genes for a total of 200 iterations. The choice of k=4 clusters was identified as the most robust clustering based on analysis of the cumulative distribution function (CDF) of the consensus (co-clustering) matrix. Cluster assignments were robust for k = 2,3,4, and decreases in the area under the curve of the CDF showed no appreciable increase after k=4. Therefore, k = 4 was chosen as the final number of clusters. Consensus clustering assignments were compared with TCGA RNA clusters. We found highly significant association between the two clustering assignments (Table S6, Chi-squared p value < 2e-16).

To identify pathways significantly over- or under-expressed in rCluster A, we used the limma voom package to identify genes which were differentially expressed in rCluster A, relative to all other clusters. The limma function geneSetTest was used to identify GO pathways enriched for over- or under-expressed genes. Analysis was restricted to gene sets with less than 500 members.

Metabolomics Supplemental Methods

Data Quality: Instrument and Process Variability

Instrument variability was determined by calculating the median relative standard deviation (RSD) for the internal standards that were added to each sample prior to injection in the mass spectrometers. Overall process variability was determine by calculating the median RSD for all endogenous metabolites (i.e., non-instrument standards) present in 100% of the Client Matrix samples, which are technical replicates of pooled client samples. Values for instrument and process variability meet Metabolon's acceptance as shown in the table below.

QC Sample	Measurement	Median RSD
Internal Standards	Instrument Variability	6 %
Endogenous Biochemicals	Total Process Variability	13 %

Sample Preparation

At the time of analysis, samples were thawed and extracts prepared according to Metabolon's standard protocol, which is designed to remove protein, dislodge small molecules bound to protein or physically trapped in the precipitated protein matrix, and recover a wide range of chemically diverse metabolites. A separate aliquot of each experimental plasma sample was taken then pooled for the creation of "Client Matrix" (CMTRX) samples. These CMTRX samples were injected throughout the platform run and served as technical replicates allowing variability in the quantitation of all consistently detected biochemicals to be determined and overall process variability and platform performance to be monitored. Extracts of all experimental and CMTRX samples were split for analysis on the GC/MS and LC/MS/MS platforms.

Data Collection and Normalization

The CMTRX technical replicate samples were treated independently throughout the process as if they were client study samples. All process samples (CMTRX, GROBs – a mixture of organic components used to assess GC column performance, process blanks, etc.) were spaced evenly among the injections for each day and all client samples were randomly distributed throughout each day's run. Data were collected over multiple platform run days and thus, 'block normalized' by calculating the median values for each run-day block for each individual compound. This minimizes any inter-day instrument gain or drift, but does not interfere with intra-day sample variability. Missing values (if any) were assumed to be below the level of detection for that biochemical with the instrumentation used and were imputed with the observed minimum for that particular biochemical.

Sample Accessioning

Each sample received was accessioned into the Metabolon LIMS system and was assigned by the LIMS a unique identifier, which was associated with the original source identifier only. This identifier was used to track all sample handling, tasks, results etc. The samples (and all derived aliquots) were bar-coded and tracked by the LIMS system. All portions of any sample were automatically assigned their own unique identifiers by the LIMS when a new task was created;

the relationship of these samples was also tracked. All samples were maintained at -80°C until processed.

Sample Preparation

The sample preparation process was carried out using the automated MicroLab STAR® system from Hamilton Company. Recovery standards were added prior to the first step in the extraction process for QC purposes. Sample preparation was conducted using a proprietary series of organic and aqueous extractions to remove the protein fraction while allowing maximum recovery of small molecules. The resulting extract was divided into two fractions; one for analysis by LC and one for analysis by GC. Samples were placed briefly on a TurboVap® (Zymark) to remove the organic solvent. Each sample was then frozen and dried under vacuum. Samples were then prepared for the appropriate instrument, either LC/MS or GC/MS.

QA/QC

For QA/QC purposes, a number of additional samples are included with each day's analysis. Furthermore, a selection of QC compounds is added to every sample, including those under test. These compounds are carefully chosen so as not to interfere with the measurement of the endogenous compounds. The two tables below describe the QC samples and compounds. These QC samples are primarily used to evaluate the process control for each study as well as aiding in the data curation.

Type	Description	Purpose
MTRX	Large pool of human plasma maintained by Metabolon that has been characterized extensively.	Assure that all aspects of Metabolon process are operating within specifications.
CMTRX	Pool created by taking a small aliquot from every customer sample.	Assess the effect of a non-plasma matrix on the Metabolon process and distinguish biological variability from process variability.
PRCS	Aliquot of ultra-pure water	Process Blank used to assess the contribution to compound signals from the process.
SOLV	Aliquot of solvents used in extraction.	Solvent blank used to segregate contamination sources in the extraction.

Type	Description	Purpose
DS	Derivatization Standard	Assess variability of derivatization for GC/MS samples.
IS	Internal Standard	Assess variability and performance of instrument.
RS	Recovery Standard	Assess variability and verify performance of extraction and instrumentation.

Liquid chromatography/Mass Spectrometry (LC/MS, LC/MS^2)

The LC/MS portion of the platform was based on a Waters ACQUITY UPLC and a Thermo-Finnigan LTQ mass spectrometer, which consisted of an electrospray ionization (ESI) source and linear ion-trap (LIT) mass analyzer. The sample extract was split into two aliquots, dried, then reconstituted in acidic or basic LC-compatible solvents, each of which contained 11 or more injection standards at fixed concentrations. One aliquot was analyzed using acidic positive ion optimized conditions and the other using basic negative ion optimized conditions in two independent injections using separate dedicated columns. Extracts reconstituted in acidic conditions were gradient eluted using water and methanol both containing 0.1% Formic acid, while the basic extracts, which also used water/methanol, contained 6.5mM Ammonium Bicarbonate. The MS analysis alternated between MS and data-dependent MS^2 scans using dynamic exclusion.

Gas chromatography/Mass Spectrometry (GC/MS)

The samples destined for GC/MS analysis were re-dried under vacuum desiccation for a minimum of 24 hours prior to being derivatized under dried nitrogen using bistrimethyl-silyl-triflouroacetamide (BSTFA). The GC column was 5% phenyl and the temperature ramp is from 40° to 300° C in a 16 minute period. Samples were analyzed on a Thermo-Finnigan Trace DSQ fast-scanning single-quadrupole mass spectrometer using electron impact ionization. The instrument was tuned and calibrated for mass resolution and mass accuracy on a daily basis. The information output from the raw data files was automatically extracted as discussed below.

Accurate Mass Determination and MS/MS fragmentation (LC/MS), (LC/MS/MS)

The LC/MS portion of the platform was based on a Waters ACQUITY UPLC and a Thermo-Finnigan LTQ-FT mass spectrometer, which had a linear ion-trap (LIT) front end and a Fourier transform ion cyclotron resonance (FT-ICR) mass spectrometer backend. For ions with counts greater than 2 million, an accurate mass measurement could be performed. Accurate mass measurements could be made on the parent ion as well as fragments. The typical mass error was less than 5 ppm. Ions with less than two million counts require a greater amount of effort to characterize. Fragmentation spectra (MS/MS) were typically generated in data dependent manner, but if necessary, targeted MS/MS could be employed, such as in the case of lower level signals.

Bioinformatics

The informatics system consisted of four major components, the Laboratory Information Management System (LIMS), the data extraction and peak-identification software, data processing tools for QC and compound identification, and a collection of information interpretation and visualization tools for use by data analysts. The hardware and software foundations for these informatics components were the LAN backbone, and a database server running Oracle 10.2.0.1 Enterprise Edition.

LIMS

The purpose of the Metabolon LIMS system was to enable fully auditable laboratory automation through a secure, easy to use, and highly specialized system. The scope of the Metabolon LIMS system encompasses sample accessioning, sample preparation and instrumental analysis and reporting and advanced data analysis. All of the subsequent software systems are grounded in the LIMS data structures. It has been modified to leverage and interface with the in-house information extraction and data visualization systems, as well as third party instrumentation and data analysis software.

Data Extraction and Quality Assurance

The data extraction of the raw mass spec data files yielded information that could loaded into a relational database and manipulated without resorting to BLOB manipulation. Once in the database the information was examined and appropriate QC limits were imposed. Peaks were identified using Metabolon's proprietary peak integration software, and component parts were stored in a separate and specifically designed complex data structure.

Compound identification

Compounds were identified by comparison to library entries of purified standards or recurrent unknown entities. Identification of known chemical entities was based on comparison to metabolomic library entries of purified standards. As of this writing, more than 2000 commercially available purified standard compounds had been acquired registered into LIMS for distribution to both the LC and GC platforms for determination of their analytical characteristics. The combination of chromatographic properties and mass spectra gave an indication of a match to the specific compound or an isobaric entity. Additional entities could be identified by virtue of their recurrent nature (both chromatographic and mass spectral). These compounds have the potential to be identified by future acquisition of a matching purified standard or by classical structural analysis.

Curation

A variety of curation procedures were carried out to ensure that a high quality data set was made available for statistical analysis and data interpretation. The QC and curation processes were designed to ensure accurate and consistent identification of true chemical entities, and to remove those representing system artifacts, mis-assignments, and background noise.

Metabolon data analysts use proprietary visualization and interpretation software to confirm the consistency of peak identification among the various samples. Library matches for each compound were checked for each sample and corrected if necessary.

Normalization

For studies spanning multiple days, a data normalization step was performed to correct variation resulting from instrument inter-day tuning differences. Essentially, each compound was corrected in run-day blocks by registering the medians to equal one (1.00) and normalizing each data point proportionately (termed the "block correction"). For studies that did not require more than one day of analysis, no normalization is necessary, other than for purposes of data visualization.

CCRCC Data Explorer Normalization

All metabolomics data in plotted in the data explorer is log2-transformed for visualization purposes. Before further analysis, users should re-transform the data to natural units.