Detailed data acquisition, processing, and reporting

Plasma metabolome of Holstein cows with and without metritis during the transition into lactation

Glossary

ALEX	automated liner exchange, produced by Gerstel corporation.
CIS	cold injection system, produced by Gerstel corporation
GC	gas chromatography
TOF	time of flight mass spectrometer
MS	mass spectrometry. After hard ionization by electron ionization, one electron gets abstracted from the intact molecules which hence become positively charged. The standardized -70 eV ionization voltage is so high that molecules fragment into multiple product ions, which may also form rearrangements among each other. Fragments are then analyzed by time of flight mass spectrometry which is made by the vendor Leco corporation not to obtain accurate mass information at high resolution but instead to obtain mass spectra at very high sensitivity and speed.
QC	quality control
IS	also istd, internal standards
FAME	fatty acid methyl esters
v/v	volumetric ratio
InChI	International Chemical Identifier key. Denotes the exact stereochemical and atomic description of chemicals and used as universal identifier in chemical databases.
KEGG	Kyoto Encyclopedia of Genes and Genomes
PubChem	a public database of chemicals and chemical information
rt	retention time (seconds)
RI	also ret.index, retention index, a conversion of absolute retention times to relative retention times based on a set of pre-defined internal standards. Classically, Kovats retention indices are used based on hydrocarbons. We use Fiehn retention indices based on FAME istd because FAME mass spectra are much easier to correctly annotate in automatic assays.
mz	also m/z, or mass-to-charge ratio. In metabolomics, ions are almost exclusively detected as singly charged species.
s/n	signal to noise ratios
IUPAC	International Union of Pure and Applied Chemistry NIST National Institute of Standards and Technology

Data acquisition

Data are acquired using the following chromatographic parameters, with more details to be found in Fiehn O. et al. Plant J. 53 (2008) 691–704.
Column: Restek corporation Rtx-5Sil MS (30 m length x 0.25 mm internal diameter with 0.25 µm film made of 95% dimethyl/5%diphenylpolysiloxane)

Mobile phase: Helium
Column temperature: 50-330°C Flow-
rate: 1 mL min^{-1}
Injection volume: 0.5 µL
Injection: 25 splitless time into a multi-baffled glass liner
Injection temperature: 50°C ramped to 250°C by 12°C s^{-1}
Oven temperature program: 50°C for 1 min, then ramped at 20°C min^{-1} to 330°C, held constant for 5 min.

The analytical GC column is protected by a 10 m long empty guard column which is cut by 20 cm intervals whenever the reference mixture QC samples indicate problems caused by column contaminations. We have validated that at this sequence of column cuts, no detrimental effects are detected with respect to peak shapes, absolute or relative metabolite retention times or reproducibility of quantifications. This chromatography method yields excellent retention and separation of primary metabolite classes (amino acids, hydroxyl acids, carbohydrates, sugar acids, sterols, aromatics, nucleosides, amines and miscellaneous compounds) with narrow peak widths of 2–3 s and very good within-series retention time reproducibility of better than 0.2 s absolute deviation of retention times. We use automatic liner exchanges after each set of 10 injections which we could show to reduce sample carryover for highly lipophilic compounds such as free fatty acids.

Mass spectrometry parameters are used as follows: a Leco Pegasus IV mass spectrometer is used with unit mass resolution at 17 spectra s^{-1} from 80-500 Da at -70 eV ionization energy and 1800 V detector voltage with a 230°C transfer line and a 250°C ion source.

Data processing

Raw data files are preprocessed directly after data acquisition and stored as ChromaTOF-specific *.peg files, as generic *.txt result files and additionally as generic ANDI MS *.cdf files. ChromaTOF vs. 2.32 is used for data preprocessing without smoothing, 3 s peak width, baseline subtraction just above the noise level, and automatic mass spectral deconvolution and peak detection at signal/noise levels of 5:1 throughout the chromatogram. Apex masses are reported for use in the BinBase algorithm. Result *.txt files are exported to a data server with absolute spectra intensities and further processed by a filtering algorithm implemented in the metabolomics BinBase database.

The BinBase algorithm (rtx5) used the settings: validity of chromatogram (<10 peaks with intensity >10^7 counts s^{-1}), unbiased retention index marker detection (MS similarity>800, validity of intensity range for high m/z marker ions), retention index calculation by 5th order polynomial regression. Spectra are cut to 5% base peak abundance and matched to database entries from most to least abundant spectra using the following matching filters: retention index window ±2,000 units (equivalent to about ±2 s retention time), validation of unique ions and apex masses (unique ion must be included in apexing masses and present at >3% of base peak abundance), mass spectrum similarity must fit criteria dependent on peak purity and signal/noise ratios and a final isomer filter. Failed spectra are automatically entered as new database entries if s/n >25, purity <1.0 and presence in the biological study design class was >80%. All thresholds reflect settings for ChromaTOF v. 2.32. Quantification is reported as peak height using the unique ion as default unless a different quantification ion is manually set in the BinBase administration software BinView. A quantification report table is produced for all database entries that are positively detected in more than 10% of the samples of a study design class (as defined in the miniX database) for unidentified metabolites. A subsequent post-processing module is employed to automatically replace missing values from the *.cdf files. Replaced values are labeled as ‘low confidence’ by color coding, and for each metabolite, the number of high-confidence peak detections is recorded as well as the ratio of the average height of replaced values to high-confidence peak detections. These ratios and numbers are used for manual curation of automatic report data sets to data sets released for submission.

Data Reporting

Data are reported in the spreadsheet named “**Metabolome_Plasma_Data.csv**”. First column has the metabolites names, and the rest of the columns have the peak heights**.** The actual data are given as **peak heights** for the quantification ion (mz value) at the specific retention index. We give peak heights instead of peak areas because peak heights are more precise for low abundant metabolites than peak areas, due to the larger influence of baseline determinations on areas compared to peak heights. Also, overlapping (co-eluting) ions or peaks are harder to deconvolute in terms of precise determinations of peak areas than peak heights. Raw results data need to be normalized to reduce the impact of between-series drifts of instrument sensitivity, caused by machine maintenance, aging and tuning parameters. Such normalization data sets are called ‘norm data’ worksheets.

The metadata can be found in the spreadsheet named “**Metabolome_Plasma_metadata.csv**”. There is a total of 104 Holstein cows with 3 timepoints each. The first sample was taken 14 days before calving (Prepartum), the second sample was taken at calving (Calving), and the third sample was taken at the diagnosis of metritis (Diagnosis). Control cows were paired by days in milk with cows that developed metritis. There is a total of 52 cows in the MET (metritis) group and 52 cows in the CON (control) group. In column “D” the parity (Multiparous or Primiparous) can be found.

The metabolites information can be found in the “**Metabolites_Information.csv**” spreadsheet. The ‘**BinBase identifier column**’ denotes the unique identifier for the GCTOFMS platform. It is given for both identified and unidentified metabolites in the same manner. The ‘**BinBase name**’ denotes the name of the metabolite, if the peak has been identified. A chemical name is not a unique identifier. We use names recognized by biologists instead of IUPAC nomenclature. If a compound is identified, it has a name, and external database identifiers such as InChI key, PubChem ID and KEGG ID. If a compound is unknown, the name is the same as given in the ‘identifier column’. The ‘**retention index**’ column details the target retention index in the BinBase database system. The ‘**quant mz**’ column details the m/z value that was used to quantify the peak height of a BinBase entry. The ‘**mass spec**’ column details the complete mass spectrum of the metabolite given as mz: intensity values, separated by spaces. The ‘**InChI key**’ identifier gives the unique chemical identifier defined by the IUPAC and NIST consortia. The ‘**KEGG**’ identifier gives the unique identifier associated with an identified metabolite in the community database KEGG LIGAND DB. The ‘**PubChem**’ column denotes the unique identifier of a metabolite in the PubChem database.