UNIVERSITY OF COPENHAGEN
FACULTY OF SCIENCE

Master Thesis

Tu Hu

Investigation of Whole Grain Wheat Intake Biomarkers

Supervisor: Gözde Gürdenz

Jun, 2019

Contents

Acronyms

GC-MS Gas Chromatography-Mass Spectrometry. 4

LC-MS Liquid Chromatography-Mass Spectrometry. 4

1 Preface

2 Introduction

2.1 Statistics method

2.1.1 Comparisions between Univariable and Multivariable statistics

t-test has multiple testing problem. because when we do a t-test, normally we use a cutoff value of 0.05, it also means we take the risk of 5% probability that it's NOT significantly different, but classified as different. this is called multiple testing problem.

FALSE DISCOVERY problem in metabolomics.

how to overcome this problem? adjusted t-test, or reduce the cutoff to a reasnable value.

multivariable data analysis and univariable data analysis show different aspects of data. It is very common to observe analysis results are significant univariablely but not multivaribalely, also, it is common to see that another way. This means uni-/multi- variable data analysis both have their limitations. that's why it is recommanded that do both uni and multi variable data analysis for the same dataset.

However, how to integrate these analysis? are they chemically correlated? maybe one feature significant in univariable analysis is associated with another one in multivaribale data analysis? Maybe, one way is to first merge all these results together. in addition, because based on current technology limitation, it's impossible to identify OR intereprete all Metabolomics results, actually also time and resources. it actually exists priorities in identifying. better chance to identify, if they're correlated. meanwhile, if intensities are high.

2.2 Identification

2.2.1 Level of Identification and communication confidence

Reporting level of identification together with identification results can enhance communication confidence. Identification is recognized by far the most difficult part of Metabolomics research, especially concerning novel compound or biomarker discovery.

In a single research project, not all structures or chemical information could be confirmed. Therefore, besides reporting chemical information (such as mass and structures), it is equally important to report the confidence of identification.

Five levels of confidence were proposed and applied extensively in xxx areas of LC-MS based compound identification[**?**].

2.3 Validation of the Biomarker

3 Materials and methods

3.1 Software

Several software packages were used for different purposes.

MATLAB R2018a (9.4.0.813654) coupled with PLS toolbox was used for data processing, modeling.

MZmine 2.31, an open source data processing software for LC-MS and GC-MS.

MassLynx was used to check mass spectra.

DataBridge, an LC-MS data file conversion program built-in MassLynx developed by Waters.

XCMS Online was used for uni-variable data analysis.

3.2 Data-preprocessing

Data-preprocessing consists x steps.

First, data format was converted by DataBridge from '.raw' to '.cdf'. '.raw' was the format directly generated by Waters analytical platform. In order to be readable by MZmine, data was converted[1].

Then, the data was preprocessed by MZmine (2.31) following the steps: peak detection, deisotoping, alignment and gap filling.

Positive mode and negative mode were separately processed because of different noise level and in-source reaction. Blank samples were also excluded in pre-processing.

[1] N.B. Although in MZmine manual, '.raw' file is described as a compatible format, in practice some weird errors were generated when '.raw' format was input into MZmine.

In the end, the detected features, including information of mass to charge ratio (m/z), retention time (rt) and intensities were output as '.csv' files for further investigation.

3.3 Statistics

3.3.1 Paired t test

Paired-t test and unpaired-t test were conducted on XCMS Online (xcmsonline.scripps.edu).

3.4 PCA

PCA was used for quality control and outlier detection.

3.5 PLSDA modeling

PLSDA modeling was used to select variables that have significant differences.

3.6 Literature search

using qian's article as a reference

4 Results

4.1 Unpaired t-test of Negative Mode Urine Samples

hi — hi |