# Metabolomics Pipeline Overview

## Table of contents
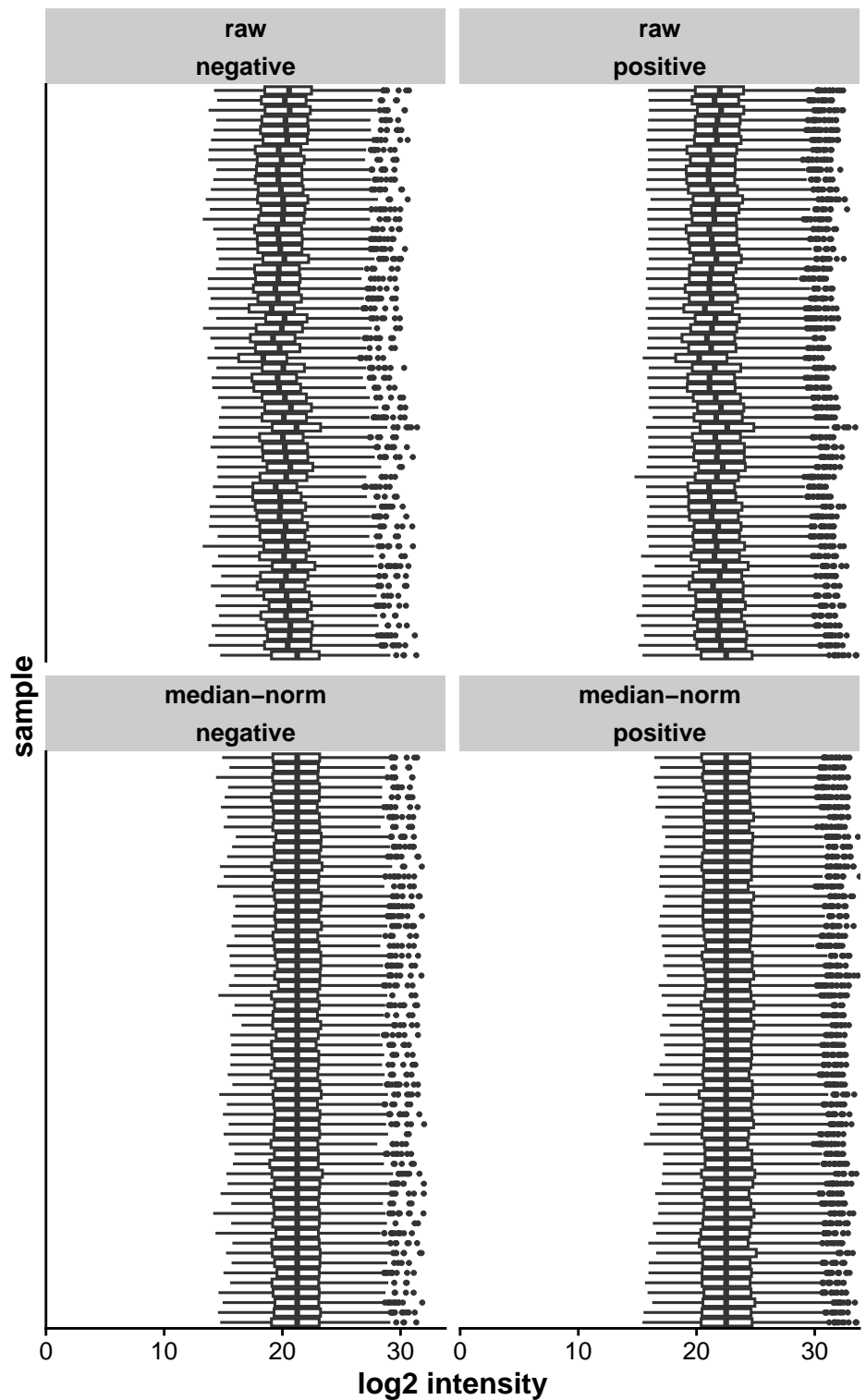
## 1 Introduction

A basic analysis pipeline in R for untargeted LC-MS metabolomics, including quality control, normalization, PCA overview of sample behavior, and differential feature analysis with linear models. The data are panels of untargeted LC–MS features from positive and negative ionization mode runs, such as metabolites measured from tissue, stool samples, or microbial cultures. In this hypothetical scenario, the data have been structured to resemble samples belonging to control subjects, one of two treatments (tr1 or tr2), or a dual-treatment group (tr1+tr2). The pipeline demonstrates a conventional approach to characterizing metabolomic differences among these groups.
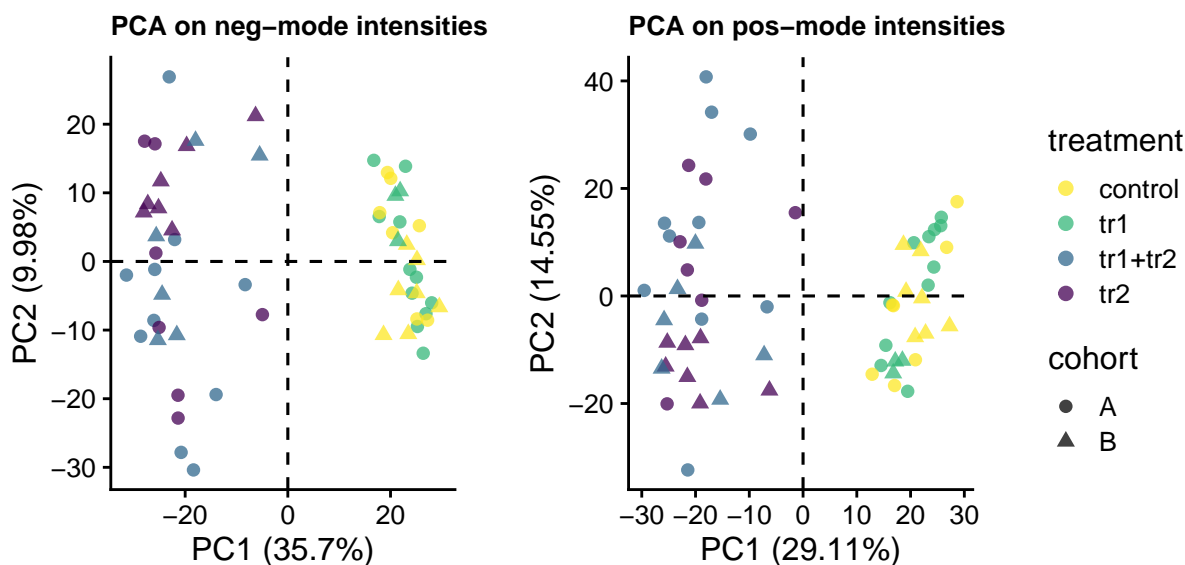
# 2 Distribution

The LC-MS intensities of metabolomics data typically skew to the right. Data normalization is required to improve comparability across samples and help stabilize variance before actual analysis can begin. The figure below demonstrates how median normalization reduces between-sample differences in overall intensity while preserving the underlying variability that's of interest in the data. The data are also transformed into a log2 scale to reduce domination by extreme values and bring the data closer to normality.

# 3 Sample Analysis — PCA

To gauge the overall structure of the metabolomic profiles (samples), a Principal Components Analysis is used. The median-normalized, log2-transformed intensities are next centered and scaled, then decomposed into a set of orthogonal principal components that capture the dominant sources of variance in the dataset.

In the figure below, principal component (PC) 1 illustrates a divide amongst the samples based on the treatment group, with controls and tr1 contrasted from tr2 and the dual-treatment (tr1+tr2). Additionally, PC 1 captures 35.7% of the variance in the data for the negative-mode intensities and 29.11% for the positive-mode intensities, meaning the treatment variable has substantial influence on the overall data structure (as might be expected). PC 2 does not appear to highlight any meaningful secondary structure in the data.

# 4 Feature Analysis — limma

(describe and depict the feature-level analyses)