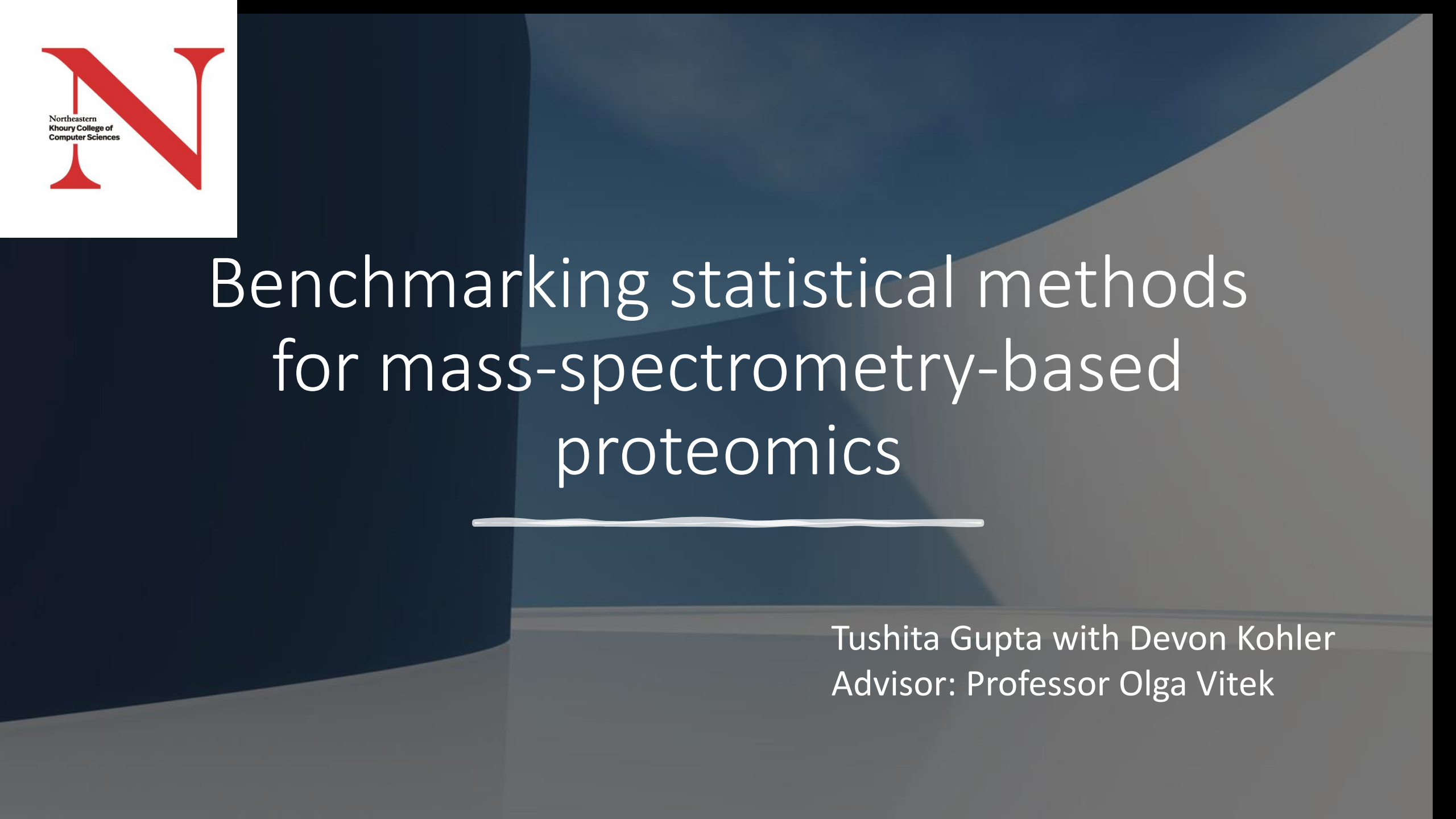



# Benchmarking statistical methods for mass-spectrometry-based proteomics

Tushita Gupta with Devon Kohler
Advisor: Professor Olga Vitek

# Introduction

- Our research compares statistical methods for analyzing mass spectrometry proteomics data to help biologists choose the best method for their specific scenario.
- This is important due to the various processing tools and biological variations that can affect the results, making it challenging to choose an appropriate method.
- By providing insights into the comparative performance of different methods, we aim to assist biologists in making informed decisions in their data analysis.

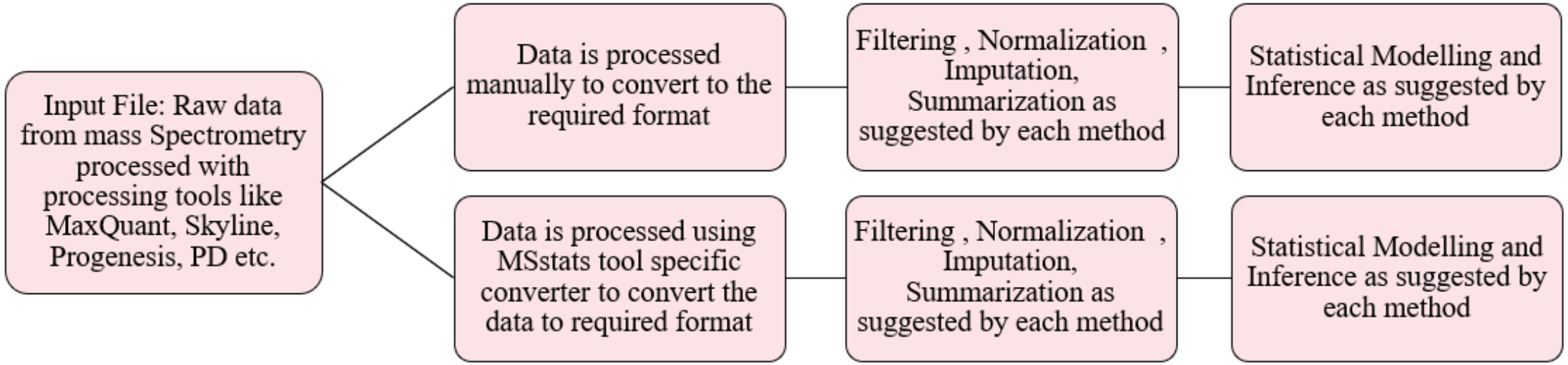
# Data Overview

- The datasets used in our research are summarized in Table 1.
- Two controlled mixtures and one biological mixture were employed representing the DDA data acquisition strategy.
- To compare the impact of preprocessing tools on the data, different preprocessing tools were used to generate multiple datasets from a single raw dataset.

| Dataset | Comparison | Number of Conditions | Number of Biological Replicates | Number of Technical Replicates | Data Processing Tool |
|---|---|---|---|---|---|
| Dataset1-DDA ControlledMix | Group | 5 | 1 | 3 | Skyline MaxQuant Progenesis P.D. |
| Dataset2-DDA:Choi2017 | Group | 4 | 1 | 3 | Skyline MaxQuant Progenesis |
| Dataset3-DDA:Meierhofer2016 | Paired | 2 | 6 | 2 | MaxQuant |

# Methodology Workflow

- The research compared the performance of eight different statistical methods, namely MSstats, MSqRob2, DEqMS, pmartR, DEP, proDA, prolfqua, and limma.
- The statistical methods were evaluated using controlled mixtures and biological experiments to assess their performance in different scenarios.

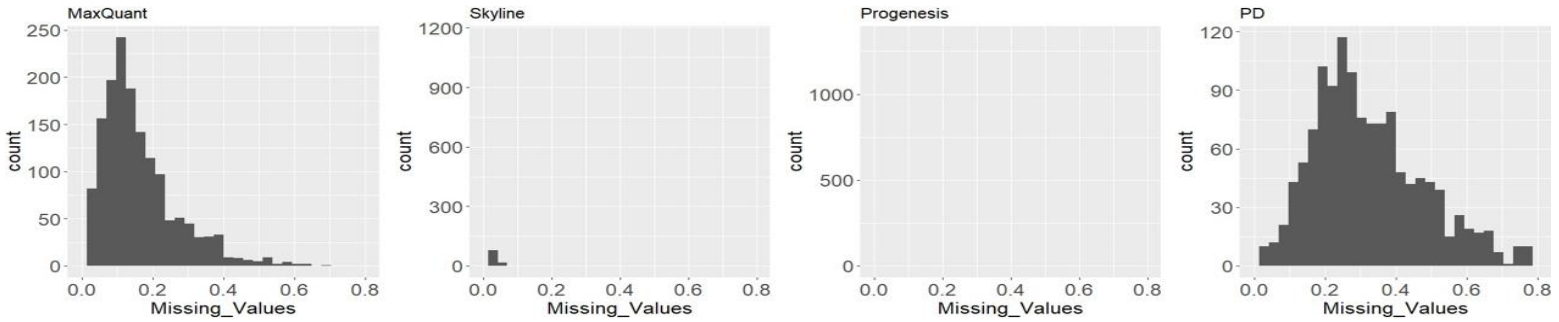


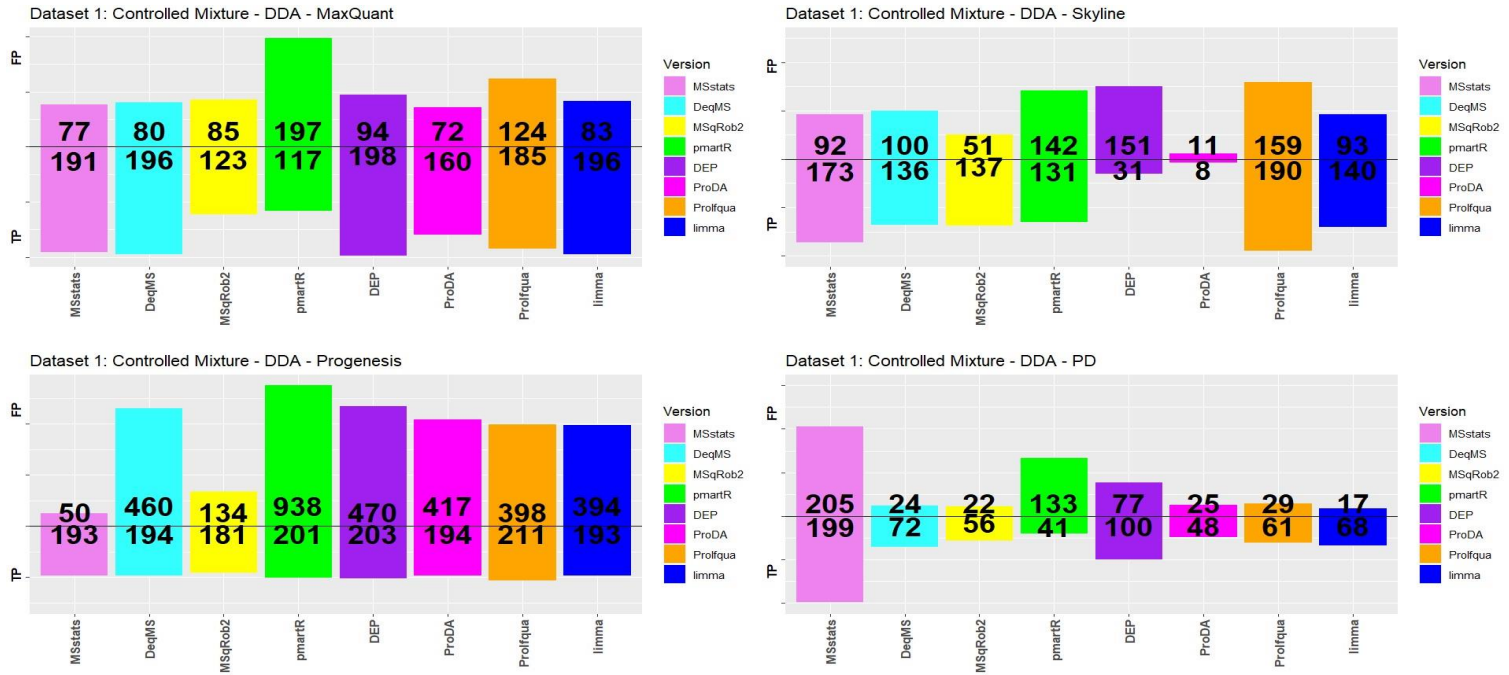
# Practical details of the existing workflows used in the evaluation

| | MSstats | PmartR | DEP | ProDA | DeqMS | MSqRob2 | Prolfqua | Limma |
|---|---|---|---|---|---|---|---|---|
| Input File | Two options used: 1. Native preprocessing 2. MSstats Converter preprocessing | | | | | | | |
| Filtering | Based on shared peptides, decoy proteins | Based on proteins, statistical fitting | Based on missing values | None | None | Based on number of peptides | Based on Intensity, number of peptides | None |
| Normalization | equalizeMedians | Median | Variance stabilizing transformation | Median | None | Median | Robust z-score | None |
| Imputation | AFT model | None | MinProb | None | None | None | None | None |
| Summarization | TMP | Roll up | Already Summarized Experiment object | Matrix with column per sample and row per protein | None | Aggregate peptides for each protein | None | None |
| Statistical model | Linear Mixed effect Model | Anova Model | Protein-wise linear models and empirical Bayes statistics using limma | Linear probabilistic dropout model | Limma workflow and a function to moderate variance based on feature counts | Linear mixed-effects model with Empirical Bayes moderation | Linear Model | Linear model fit with Bayes moderation |

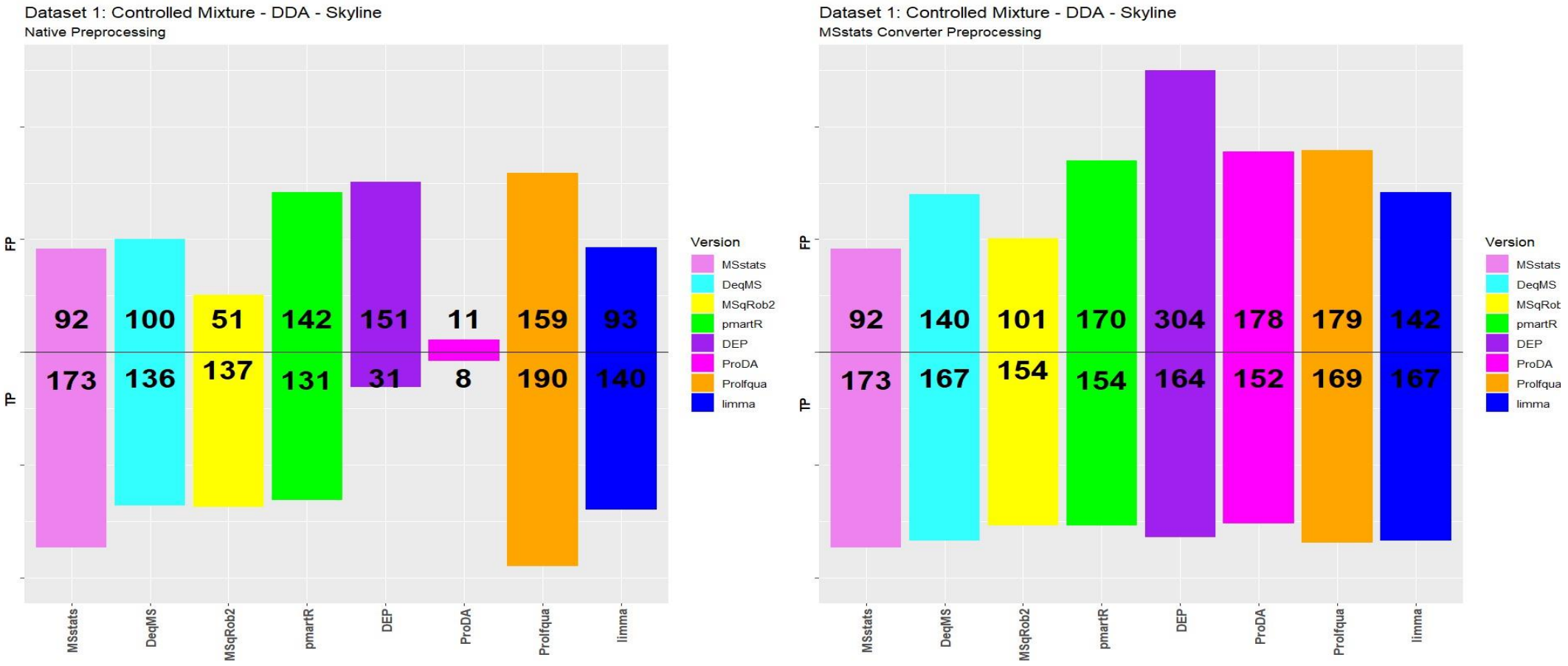
# Results

# 1. **Processing tool** affects the performance of statistical methods

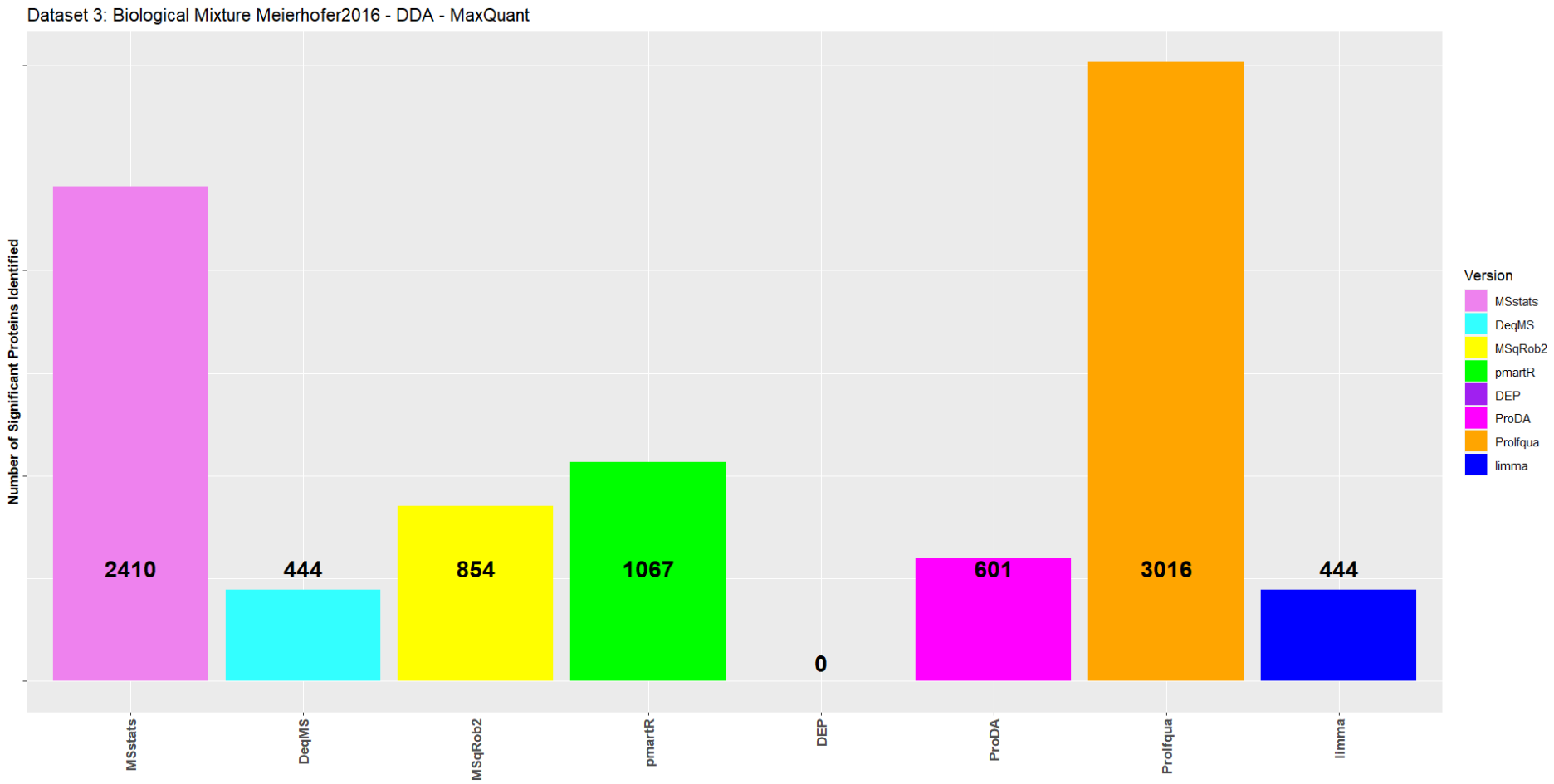


# 2. **Upstream processing** affects the performance of statistical methods



# 3. **Biological variation** affects the performance of statistical methods



Thank you