

Improve shotgun proteomics identifications with b10prot R package.

Create tidyverse-style workflows.

Proteomics Identification Workflow in R

Gorka Prieto¹, 

 @GorkaEhu

 gorka.prieto@ehu.eus

Nerea Osinalde² Pedro Navarro³ Jesús Vázquez⁴

¹ Department of Communications Engineering, University of the Basque Country (UPV/EHU)

² Department of Biochemistry and Molecular Biology, University of the Basque Country (UPV/EHU)

³ Chromatography & Mass Division Software, Thermo Fisher Scientific (Bremen) GmbH

⁴ Cardiovascular Proteomics Laboratory, Centro Nacional de Investigaciones Cardiovasculares Carlos III (CNIC) and CIBER de Enfermedades Cardiovasculares

Motivation

- **Shotgun proteomics:** Preferred method for large-scale protein identification.
- **Challenge:** Need for robust statistical methods to automate identification and reduce false positives.
- **Our contribution:**
 - Adapted our existing protein inference and scoring tools (PAnalyzer, LPGF, refined FDR).
 - Integrated seamlessly with shotgun proteomics workflows using R.
- **Goal:** Make these tools more accessible to a wider scientific audience via R.

Approach

PSMs

- A `data.frame` with the following columns:
 - `psmScore`, `rank`, `isDecoy`, `peptideRef`, `proteinRef`, `geneRef` (optional)
- Select `rank=1` PSMs

Peptides

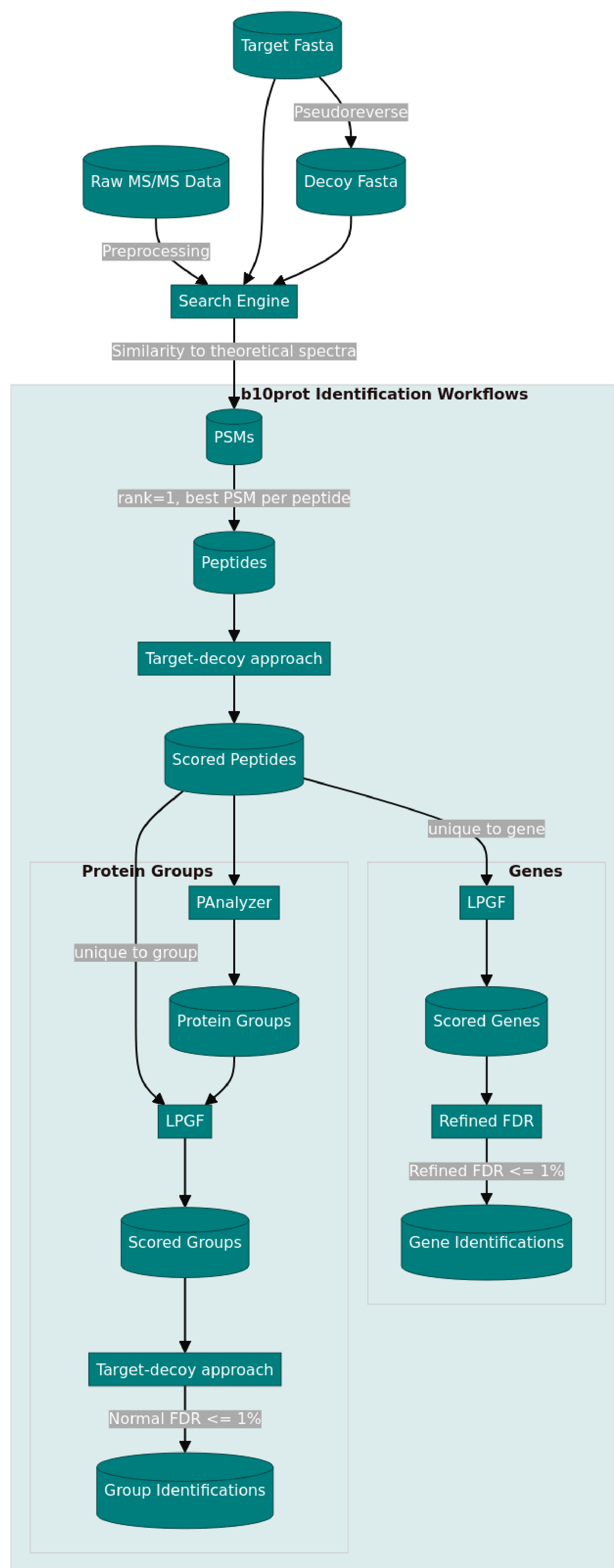
- Best PSM per peptide as `pepScore`
- Target-decoy approach (Elias and Gygi 2007) to assign statistical significance (Käll et al. 2008):
 - p-value, FDR, q-value
- Identified peptides: q-value $\leq 1\%$
- For upper level (eg. protein) scores we will use:
 - Peptide-level `LP` = \log_{10} of p-value

Proteins/Genes

- Consider only peptides **unique** to one protein/gene
- Calculate **LPGF** score (Prieto and Vázquez 2020) from:
 - `LP` peptide scores
 - `n` number of matched peptides
 - `m` number of identified peptides
- Protein/Gene-level **refined FDR** (Prieto and Vázquez 2020) using LPGF score

Protein Groups

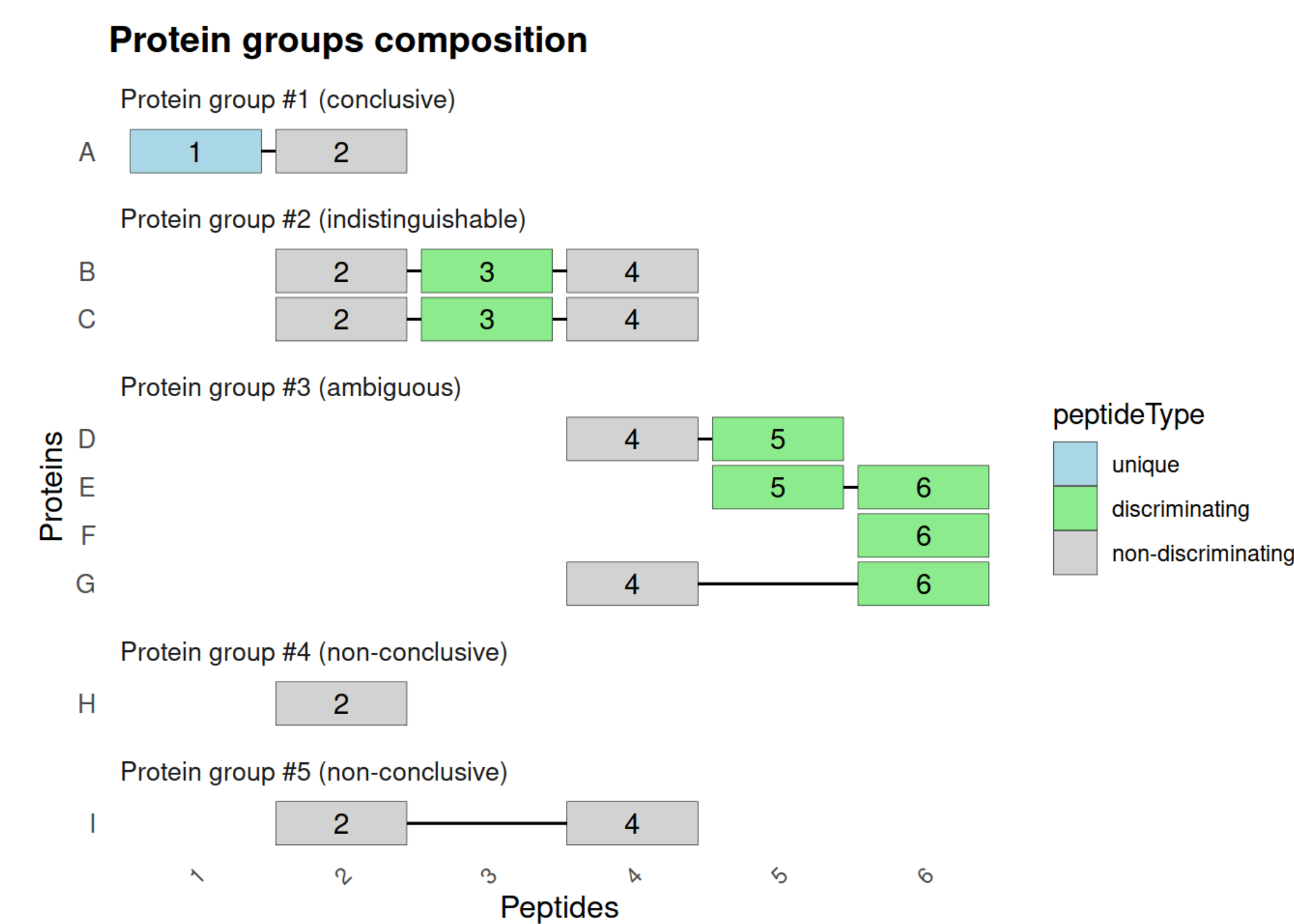
- Construct protein groups using **PAnalyzer** (Prieto et al. 2012)
- Consider only peptides **unique** to one group
- Calculate group-level **LPGF** score
- **Group-level FDR** using LPGF score



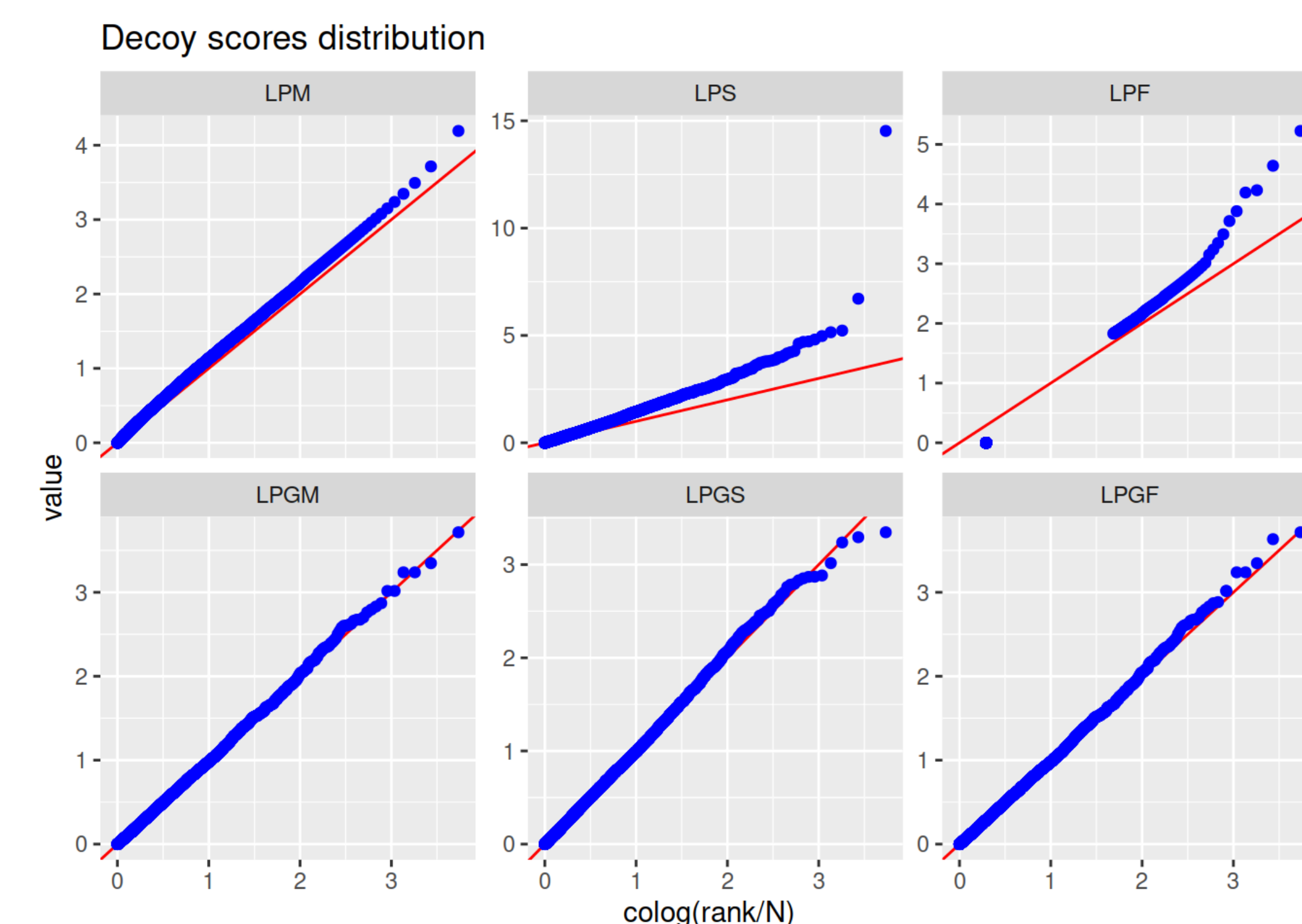
Results

- New **b10prot** R package: <https://akrogp.github.io/b10prot/>

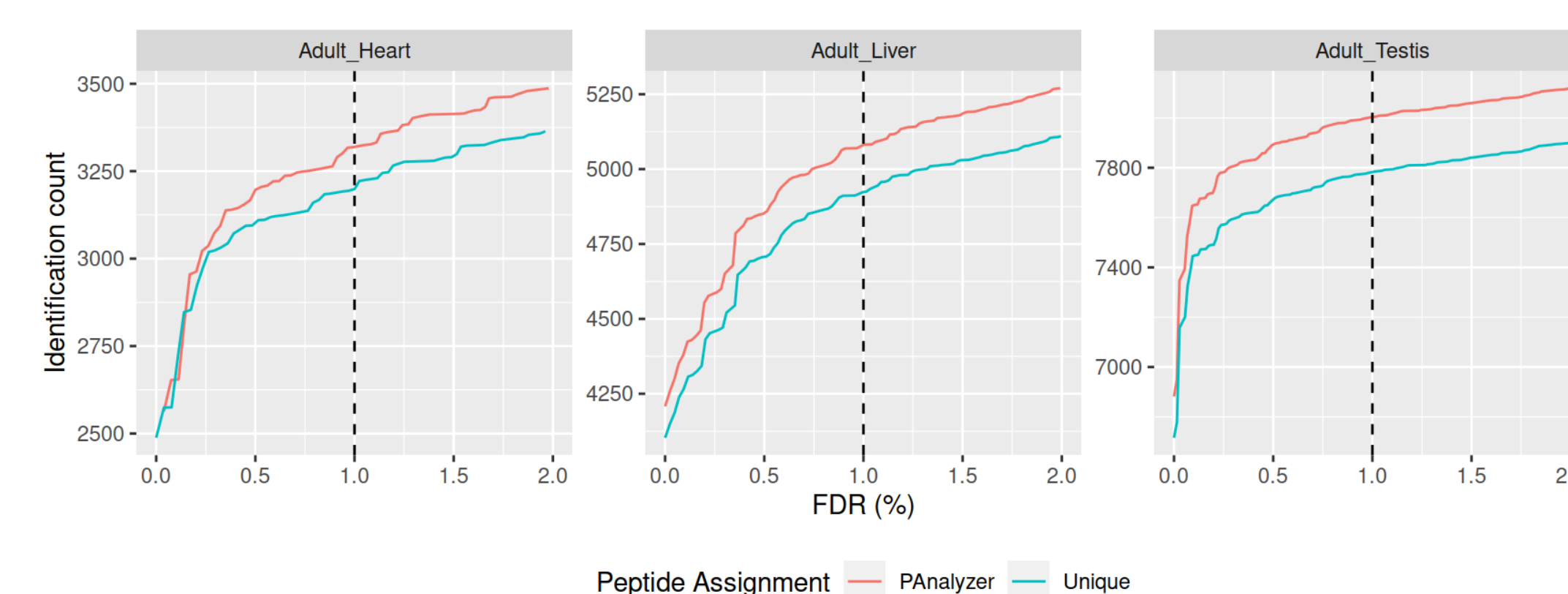
- **PAnalyzer** can now be used from R:



- **LPGF** and **FDRr** implemented in R:



- Approx. 3% improvement using protein groups:



- Approx. 5% improvement compared to other scores:
 - Navarro et al. (2024)

References

Elias, Joshua E., and Steven P. Gygi. 2007. "Target-Decoy Search Strategy for Increased Confidence in Large-Scale Protein Identifications by Mass Spectrometry." *Nature Methods* 4 (3): 207–14.

Käll, Lukas, John D. Storey, Michael J. MacCoss, and William Stafford Noble. 2008. "Assigning Significance to Peptides Identified by Tandem Mass Spectrometry Using Decoy Databases." *Journal of Proteome Research* 7 (1): 29–34.

Navarro, Pedro, Waqas Nasir, Kai Fritzemeier, Gorka Prieto, Víctor M. Guerrero-Sánchez, Jesús Vázquez, and Christoph Henrich. 2024. "Performance of LPGF Protein Validation in diverse sample types and mass spectrometry workflows." *72nd ASMS Conference on Mass Spectrometry and Allied Topics*. Anaheim, California.

Prieto, Gorka, Kerman Aloria, Nerea Osinalde, Asier Fullaondo, Jesus M. Arizmendi, and Rune Matthiesen. 2012. "PAnalyzer: A software tool for protein inference in shotgun proteomics." *BMC Bioinformatics* 13 (1): 288.

Prieto, Gorka, and Jesús Vázquez. 2020. "Protein Probability Model for High-Throughput Protein Identification by Mass Spectrometry-Based Proteomics." *Journal of Proteome Research* 19 (3): 1285–97.

`==]b10prot[==`



eman ta zabal zazu

