LimROTS: A Hybrid Method Integrating Empirical Bayes and Reproducibility-Optimized Statistics for Robust Analysis of Proteomics Data

Ali Mostafa Anwar[a*], Akewak Jeba[b], Leo Lahti[b] & Eleanor Coffey[a]

[a]Turku Bioscience Centre, University of Turku and Åbo Akademi University, 20520 Turku, Finland
[b]Department of Computing, University of Turku, Turku, Finland

***Correspondence: aymost@utu.fi**

Abstract

Differential expression analysis plays a vital role in proteomics research, as it enables the precise identification of proteins that are specifically associated with different phenotypes. This process is critical for uncovering biological differences between conditions, such as disease versus healthy states. In proteomics, several statistical methods have been used, ranging from simple t-tests to more sophisticated methods like limma. In this study, we introduce LimROTS, which extends Empirical Bayes from limma into an optimization framework using reproducibility-optimized statistics, enabling differential expression analysis that accounts for technical and biological variation. To validate LimROTS performance, 22 proteomics gold standard spike-in datasets with different protein mixtures, MS instruments, and techniques were used for benchmarking. This hybrid approach improves the accuracy and reproducibility in proteomics studies, making LimROTS a powerful tool for high-dimensional omics data analysis.

Introduction

Advances in mass spectrometry technologies and data processing methods have expanded the depth of proteomic detail attainable from biospecimens. These large datasets encode complex biological information that can help to advance precision medicine. The ability to detect phenotype-specific proteins not only enhances our understanding of underlying biological mechanisms, it also holds promise for the discovery of novel biomarkers for early disease diagnosis, prognosis, and monitoring (Fuchu He et al. 2025; Yue-Ting Deng et al. 2025). It also aids the identification of potential drug targets for the development of new therapeutic strategies (Niu et al. 2022; Meissner et al. 2022).

One of the statistical tools that recently gained attention for its high performance compared to other tools in proteomics data (Peng et al. 2024) is ROTS (Reproducibility-Optimized Test Statistic) (Suomi et al. 2017). ROTS is a statistical method specifically designed for differential expression analysis, including applications in proteomics (originally developed for microarray RNA data). It identifies differentially expressed proteins by optimizing the test statistic based on reproducibility across replicate datasets. Instead of relying on traditional fixed test statistics, ROTS searches for the optimal combination of statistical parameters that maximize the reproducibility of significant results. This data-driven approach improves the detection of true biological signals while controlling for false discoveries (Elo et al. 2008), making ROTS a robust tool for analyzing complex proteomics datasets where measurement variability is common.

Another well-known tool that has been developed in the first place for microarray RNA data as well as RNA-sequencing is limma (Ritchie et al. 2015). It utilizes linear models to analyze data, accounting for complex experimental designs. limma applies empirical Bayes methods to improve statistical power and control for false discoveries, making it particularly effective for high-dimensional data like proteomics (Ooijen et al. 2017; Peng et al. 2024). By combining robust statistical techniques with flexible model designs, limma has become a cornerstone tool for identifying differentially expressed proteins in a variety of biological contexts (Ooijen et al. 2017).

The addition of covariates is common in differential expression analysis (DEA) to account for known biological or technical variations, such as age or gender adjustments (Yingxiao Yan et al. 2024; Rebecca Ting Jiin Loo et al. 2024). ROTS lacks the capability to adjust for these covariate effects, which makes it less flexible for more complex omics experimental designs. In this study, we developed LimROTS, a hybrid method integrating Empirical Bayes from the limma framework and Reproducibility-Optimized Statistics from ROTS for robust analysis of proteomics data. LimROTS takes advantage of being flexible, as it allows for more complicated experimental designs to be modeled with the linear model. Furthermore, as ROTS has shown notably higher performance, especially in proteomics data, integrating both methods improves the results by leveraging the strengths of both approaches. To test LimROTS, 22 gold standard spike-in datasets (Peng et al. 2024) were utilized and compared to limma, ROTS, ANOVA, t-test, and SAM.

Materials and Methods

Integrating Empirical Bayes and Reproducibility-Optimized Statistics

LimROTS integrates the statistical principles of Empirical Bayes as implemented before in limma (Ritchie et al. 2015) and extends that as an optimization problem, solved by the reproducibility optimization statistics (Suomi et al. 2017). Therefore, in principle, LimROTS inherits some parts of the workflow and the mathematical rationale from both methods.

LimROTS introducing a new moderated ranking statistics $\tilde{d}_{\alpha_{(i)}}$ represented as:

$$\tilde{d}_{\alpha_{(i)}} = \frac{\hat{\beta}_i}{\alpha_1 + \left(\alpha_2 (u_i . \tilde{s}_i)\right)} \qquad (1)$$

where, β_i is the effect (coefficient) due to the experiment conditions, α_1 and α_2 are reproducibility optimized parameters, u_i is the un-scaled standard deviation, and $\tilde{s}_i$ is the posterior residual standard deviation.

To calculate $\tilde{d}_{\alpha_{(i)}}$, first, a feature-wise linear model is fitted which allow to handle complex experimental design, this can be represented as:

$$E[x_i = X\beta_i] \qquad (2)$$

where, x_i is the observed expression value of a protein, X is the design matrix of an experiment, and β_i is the effect (coefficient) trying to be estimated by the linear model, in other words; how much expression changes due to the experiment conditions.

Using parametric empirical Bayes technique to borrow information between proteins in a dynamic way, the $\widetilde{s_i}$ which is the posterior residual standard deviation is used to to adjust the variance to account for uncertainty in the proteins measurements and combines it with prior information estimated from the data, leading to empirical bayes variance shrinkage, and can be calculated as;

$$\widetilde{s_i}^2 = \frac{f_o s_o^2 + f_i s_i^2}{f_o + f_i} \qquad (3)$$

where, f_o is the prior degree of freedom (a global parameter) calculated from the whole dataset, s_o^2 is the prior variance (a global parameter) calculated from the whole dataset, f_i is the residual degree of freedom calculated for each protein (a protein specific parameter), and s_i^2 is the observed residual variance for protein i (a protein specific parameter).

Equation three allows, more reliable estimate of variance for each protein, as well as more statistical confidence in the estimates, because the final variance estimate for each protein ($\widetilde{s_i}$) is a balance between two things: the protein specific variance, and the global variance across all the proteins in the experiment. In other words, each protein specific variance is shrinkage to a common variance estimated from all the proteins in the dataset.

The optimization parameters α_1 and α_2 are estimated by maximizing the reproducibility by z-type statistics $Z_k(\widetilde{d_\alpha})$. To calculate the $Z_k(\widetilde{d_\alpha})$, the average reproducibility score $R_k(\tilde{d}_\alpha)$ has to be calculated from a bootstrap dataset as:

$$R_k(\tilde{d}_\alpha) = \frac{1}{B}\sum_{B}^{b=1} R_k^{(b)}(\tilde{d}_\alpha) \qquad (4)$$

where, for each pair of bootstrap datasets $D_1^{(b)}$ and $D_2^{(b)}$ the reproducibility $R_k^{(b)}(\tilde{d}_\alpha)$ can be computed as:

$$R_k^{(b)}(\tilde{d}_\alpha) = \frac{\#\{p | r(\alpha, D_1^b) \leq k, r(\alpha, D_2^b) \leq k\}}{k} \qquad (5)$$

where, $\{p|\ldots\}$ represents the set of proteins that satisfy the condition of being less than or equal to k (k equal to cutoff of considering a protein as a top rank protein) in both $D_1^{(b)}$ and $D_2^{(b)}$, $r(\alpha, D_1^b)$ is the rank of protein p in the dataset $D_1^{(b)}$ with respect to the statistic $\tilde{d}_\alpha$ and the same for $r(\alpha, D_2^b)$, then scaled by the k.

Finally, the $Z_k(\widetilde{d_\alpha})$ can be estimated by:

$$Z_k(\widetilde{d_\alpha}) = \frac{R_k\widetilde{d_\alpha} - R_k^o\widetilde{d_\alpha}}{S_k(\widetilde{d_\alpha})} \qquad (6)$$

The stander divination $S_k(\widetilde{d_\alpha})$ can be computed by square root the variance calculated in by;

$$S_k^2 = \sum_{b=1}^{B} \left[R_k^{(b)}(\tilde{d}_\alpha) - R_k(\tilde{d}_\alpha)/B - 1 \right] \qquad (7)$$

where, B is the number of bootstrapping datasets. The $R_k^o \widetilde{d_\alpha}$ is the null reproducibility estimate of $\tilde{d}_\alpha$ in random (per mutated) data. Finally, to avoid the need of pre-specified k, LimROTS maximize the $Z_k(\widetilde{d_\alpha})$ over a lattice of (α, k)-pairs.

Datasets used for workflow benchmarking

Table 1 shows 22 gold standard spike-in datasets used in this study to test LimROTS performance against other tools. Datasets were adapted from a published benchmark study (Peng et al. 2024). In all datasets, appropriate software was applied for the quantification step. For DIA, Spectronaut (Roland Bruderer et al. 2017) and DIANN (Vadim Demichev et al. 2020) were used, while for DDA, MaxQuant (Lili Niu et al. 2020) and FragPipe (Andy T Kong et al. 2017) were used. For the preprocessing of each dataset, we followed the recommendations from the same study. Datasets may have more than one contrast; therefore, all the contrasts for a dataset have been considered in the comparison. In Case Study 1, all the datasets have been analyzed in their original form (after preprocessing). In Case Study 2, we merged both HEof_w60 and HEof_n600_DIA, as they are the same samples analyzed twice with different quantification settings (wide and narrow DIA settings). This was done to add some complexity to the experimental design to mimic some of the bias that may exist in a real biological experiment rather than gold standard spike-in datasets. In Case Study 3 and 4, both used the same mixture of samples as in Case Study 2. However, in Case Study 3, a random effect size (from 1 to 10) was added to the narrow setting samples only. Moreover, to increase the complexity, we assigned toy batches to each sample to mimic the effect of digestion batches in proteomics experiments.

Table 1*: Datasets used for workflow benchmarking*

Dataset	ID	Technique	Mixture	Instrument
HYEtims735_DIA	PXD028735	DIA	Human+yeast+*E. coli*	TimsToFpro
MYtims709_DIA	PXD034709	DIA	Mouse + yeast	TimsTOF Pro
HEof_n600_DIA	PXD026600	DIA	UPS1 + E. coli	Orbitrap Fusion ETD
HEof_w600_DIA	PXD026600	DIA	UPS1 + E. coli	Orbitrap Fusion ETD
HYtims134_DIA	PXD036134	DIA	human + yeast	TimsToF pro
HEqe777_DIA	PXD019777	DIA	Human + E. coli	Q Exactive HF

HEqe408_DIA	PXD018408	DIA	Human + E. coli	Q Exactive
HEof_n600_DIA+HEof_w600_DIA	PXD036134	DIA	UPS1 + E. coli	Orbitrap Fusion ETD
HEof_n600_DIA+HEof_w600_DIA	PXD036134	DIA	UPS1 + E. coli	Orbitrap Fusion ETD
HEof_n600_DIA+HEof_w600_DIA	PXD036134	DIA	UPS1 + E. coli	Orbitrap Fusion ETD
HYE560073 5_LFQ	PXD028735	DDA	Human+yeast + *E. coli*	SCIEX Triple TOF5600
HYE660073 5_LFQ	PXD028735	DDA	Human+yeast+ *E. coli*	SCIEX Triple TOF5600
HYEqe735_LFQ	PXD028735	DDA	Human+yeast+*E. coli*	Orbitrap QE-HFX
HYEtims735_LFQ	PXD028735	DDA	Human+yeast+*E. coli*	TimsToF pro
HYtims134_LFQ	PXD036134	DDA	Human+yeast	TimsToF pro
HEtims425_LFQ	PXD021425	DDA	Human+*E. coli*	TimsToF pro
YUltq006_LFQ	PDC000006	DDA	Yeast+UPS1	LTQ-Orbitrap
YUltq099_LFQ	PXD002099	DDA	Yeast+UPS1	LTQ Orbitrap Velos
YUltq819_LFQ	PXD001819	DDA	Yeast+UPS1	LTQ Orbitrap Velos
HEqe408_LFQ	PXD018408	DDA	Human+*E. coli*	Q Exactive
HYqfl683_LFQ	PXD007683	DDA	Human+yeast	Orbitrap Fusion Lumos

HYEtims777_LFQ	PXD014777	DDA	Human+yeast+*E. coli*	TimsToF pro

Performance evaluation metrics

Throughout this study, significant proteins were identified with both 0.01 and/or 0.05 adjusted p-values. Additionally, five performance evaluation metrics were used: nMCC, F1 score, G-mean, balanced accuracy, and pAUC with a 5% cutoff. nMCC (normalized Matthews correlation coefficient) metric is used to assess the quality of each method. It considers true positives, true negatives, false positives, and false negatives. A value closer to 1 indicates better classification performance, and 0 suggests poor performance. F1 score is the harmonic mean of precision and recall, balancing the accuracy of positive predictions and the ability to identify true positives. G-mean is the geometric mean of sensitivity and specificity, indicating how well the model balances correctly identifying both positives and negatives. Balanced accuracy is the average of sensitivity and specificity, ensuring the model is not biased by class imbalances. Finally, the pAUC is the area under the ROC curve within a specified false positive rate range (5% cutoff), measuring how well the model distinguishes between classes.

LimROTS as bioconductor package

LimROTS workflow has been implemented as an R package to facilitate its use. It can be downloaded from Bioconductor (Mostafa Anwar et al. 2024) (https://bioconductor.org/packages/LimROTS/), and a documentation on how to use it is also available here https://aliyoussef96.github.io/LimROTS/. Also, a GitHub repository are available https://github.com/AliYoussef96/LimROTS

Results and discussion

Case study 1

In Case study 1, all possible contrasts have been analyzed for each dataset, with an average of 3 contrasts per dataset, corresponding to approximately 68 comparisons overall. When using DIANN as quantification software, with a significance threshold of 5% FDR and $\log_2$ fold change, LimROTS and ROTS outperformed other methods in terms of nMCC (median values: 0.88 and 0.87, respectively), F1 score (median values: 0.79 and 0.75, respectively), and pAUC 5% (median values: 0.94 and 0.92, respectively) (Figure 1a). However, LimROTS consistently displayed slightly higher values than ROTS across all these evaluation metrics. A similar trend was observed when Spectronaut was used as the quantification software. The results for all evaluation metrics across datasets and tools used in this study are available in Supplementary File 1, Table S1 (A–H). Furthermore, with the data dependent analysis (DDA) dataset using MaxQuant as the quantification software, LimROTS and ROTS remained as the top-performing methods in terms of F1 score, nMCC, and pAUC 5% (Supplementary File 1, Table S1 G–H). Nevertheless, when FragPipe was used as the quantification software, LimROTS and ROTS continued to be the top-performing methods overall; however, performance across all tools was notably lower in terms of F1 score and nMCC (Supplementary File 1, Table S1 E–F).

The findings from Case Study 1 suggest that LimROTS and ROTS offer superior overall performance compared to alternative methods (e.g., limma, t-test, ANOVA, and SAM) when

DIANN or Spectronaut is used as the quantification software under a significance threshold of 5% FDR and $\log_2$ fold change. The high nMCC values, which reflect the overall classification performance, indicate that both LimROTS and ROTS accurately detect most true positive proteins while effectively distinguishing true negatives. The F1 score, a metric that balances precision and recall, was also higher for LimROTS and ROTS, further supporting their superior capability in detecting significant proteins while minimizing false positives and false negatives. Additionally, both methods showed strong performance in pAUC 5%, highlighting their ability to rank the most significant proteins in the top 5%. Furthermore, for DDA-based analysis, particularly when using MaxQuant for quantification, LimROTS and ROTS continued to outperform other methods. These results establish LimROTS and ROTS as promising candidates for DEA in proteomics DIA workflows. In particular, LimROTS demonstrated marginally better performance in this low-complexity experimental design.

Case Study 2

As described in the Materials and Methods section, this case study is more complex than the previous one, as it focuses on the HEof dataset, combining both wide DIA and narrow DIA settings using the same samples. Using DIANN as the quantification software, LimROTS showed a notably higher performance in F1 score (median 0.61, with ROTS second at 0.48), nMCC (median 0.82, with ROTS second at 0.75), and pAUC 5% (median 0.86, with ROTS second at 0.81) (Figure 1b). A similar trend was observed with Spectronaut. The results for all evaluation metrics across all datasets and tools used in this case study are available in Supplementary File 1, Table S2 A–D.

The findings from Case Study 2, where the experimental design complexity increased by combining the same set of samples analyzed using both narrow and wide DIA settings, indicate that LimROTS showed significantly higher overall performance compared to all other tools. The highest F1 score, 0.61, was achieved by LimROTS, indicating its superior ability to identify true positives while minimizing both false positives and false negatives. In contrast, the second-highest score of 0.48 was recorded by ROTS, which further suggests that LimROTS is more effective, even in the presence of strong covariates. Furthermore, LimROTS achieved the highest scores for both nMCC and pAUC 5%, indicating better discriminative power and stronger overall performance in distinguishing between true differential and non-differential proteins. These results can be attributed to the additional advantages LimROTS offers over ROTS, as it flexibly integrates both narrow and wide DIA settings into the linear model as covariates, along with its advantage over limma by optimizing the rank statistics of the proteins.

Case Study 3

In this case study, we excluded SAM, ROTS, and the t-test, as these methods are not specifically designed to account for covariates in an experiment, apart from the target groups, which is crucial for obtaining accurate results, as demonstrated in Case Study 2. Additionally, ANOVA was excluded due to its overall poor performance. (All results for all tools are available in Supplementary File 1, Table S3 A–D.) Therefore, LimROTS was compared against limma in this analysis. Using Spectronaut as the quantification software, LimROTS demonstrated significantly higher performance across all evaluation metrics (F1 score, nMCC, pAUC 5%, G-mean) compared to limma. Specifically, LimROTS achieved values of 0.8 (F1 score), 0.9 (nMCC), 0.96 (pAUC 5%), 0.96 (G-mean), and 0.96 (balanced accuracy), while limma achieved values of 0.38,

0.73, 0.92, 0.92, and 0.92, respectively, as shown in Figure 1c. Similarly, using DIANN as the quantification software, LimROTS outperformed limma in nMCC (0.84 compared to 0.67) and F1 score (0.64 compared to 0.27) (Supplementary File 1, Table S3 A–D).

The findings from Case Study 3 demonstrate that LimROTS significantly outperforms limma (and all other tools, as shown in the supplementary file 1) in balancing precision and recall when using either Spectronaut or DIANN. LimROTS effectively identifies true positives while minimizing false positives and false negatives, as indicated by its high F1 score. Additionally, the G-mean, which represents the balance between sensitivity (true positive rate) and specificity (true negative rate), was much higher for LimROTS compared to limma. Furthermore, the higher pAUC 5% score of LimROTS highlights its superior ability to rank the top 5% of the most differentially expressed proteins with greater confidence in its top-ranking results. Moreover, the nMCC score of LimROTS exceeded the performance of limma. Overall, LimROTS consistently outperformed limma when using either Spectronaut or DIANN, which indicates that LimROTS is a more reliable and accurate tool for in proteomics DIA and DDA data.

Case Study 4

This case study introduces synthetic bias into the data and incorporates unbalanced synthetic digestion batches, making the experimental design more complex. As in Case Study 3, LimROTS and limma were compared, with all other tools excluded (comprehensive results for all tools are provided in Supplementary File 1, Table S4 A and B). As shown in Figure 2, LimROTS significantly outperformed limma across all evaluated metrics. Specifically, LimROTS and limma achieved the following scores: F1 score (0.78 compared to 0.15), nMCC (0.90 compared to 0.61), pAUC 5% (0.96 compared to 0.90), G-mean (0.99 compared to 0.81), and balanced accuracy (0.99 compared to 0.83), respectively.

Case Study 4 features a more complex experimental design than the other case studies, requiring a flexible tool to effectively model covariates, even in unbalanced settings. The results from this semi-synthetic dataset show that LimROTS performs better at accurately identifying differentially expressed proteins while maintaining a strong balance between sensitivity and specificity.

Conclusion

LimROTS consistently demonstrated superior performance compared to other differential expression analysis statistical tools, such as limma, ROTS, and SAM, for all case studies. For Case Study 1, LimROTS performed better in nMCC, F1 score, and pAUC 5% over other tools when quantitation was conducted either by DIANN or Spectronaut. Case Study 2, featuring the more complicated experimental design, further demonstrated the capability of LimROTS in processing complex data as it outcompeted ROTS in every evaluation metric. In Case Study 3, LimROTS performed better than limma across all the metrics used for evaluation, demonstrating its capacity to balance recall and precision. Further, in Case Study 4, LimROTS did better in accommodating synthetic biases and unbalanced batches, performing better than limma across all the metrics used. Generally, LimROTS was more accurate, more reliable, and more flexible for differential expression analysis in proteomics research.

References

Elo, Laura L., Sanna Filen, Riitta Lahesmaa, and Tero Aittokallio. 2008. “Reproducibility-Optimized Test Statistic for Ranking Genes in Microarray Studies.” *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 5 (3): 423–31. https://doi.org/10.1109/tcbb.2007.1078.

Meissner, Felix, Jennifer Geddes-McAlister, Matthias Mann, and Marcus Bantscheff. 2022. “The Emerging Role of Mass Spectrometry-Based Proteomics in Drug Discovery.” *Nature Reviews Drug Discovery* 21 (9): 637–54. https://doi.org/10.1038/s41573-022-00409-3.

Mostafa Anwar, Ali, Leo Lahti, Akewak Jeba, and Eleanor Coffey. 2024. *LimROTS: A Hybrid Method Integrating Empirical Bayes and Reproducibility-Optimized Statistics for Robust Analysis of Proteomics and Metabolomics Data*. https://doi.org/10.18129/B9.bioc.LimROTS.

Niu, Lili, Maja Thiele, Philipp E. Geyer, Ditlev Nytoft Rasmussen, Henry Emanuel Webel, Alberto Santos, Rajat Gupta, et al. 2022. “Noninvasive Proteomic Biomarkers for Alcohol-Related Liver Disease.” *Nature Medicine* 28 (6): 1277–87. https://doi.org/10.1038/s41591-022-01850-y.

Ooijen, Michiel P van, Victor L Jong, Marinus J C Eijkemans, Albert J R Heck, Arno C Andeweg, Nadine A Binai, and Henk-Jan van den Ham. 2017. “Identification of Differentially Expressed Peptides in High-Throughput Proteomics Data.” *Briefings in Bioinformatics* 19 (5): 971–81. https://doi.org/10.1093/bib/bbx031.

Peng, Hui, He Wang, Weijia Kong, Jinyan Li, and Wilson Wen Bin Goh. 2024. “Optimizing Differential Expression Analysis for Proteomics Data via High-Performing Rules and Ensemble Inference.” *Nature Communications* 15 (1). https://doi.org/10.1038/s41467-024-47899-w.

Ritchie, Matthew E., Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. 2015. “Limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies.” *Nucleic Acids Research* 43 (7): e47–47. https://doi.org/10.1093/nar/gkv007.

Suomi, Tomi, Fatemeh Seyednasrollah, Maria K. Jaakkola, Thomas Faux, and Laura L. Elo. 2017. “ROTS: An r Package for Reproducibility-Optimized Statistical Testing.” Edited by Timothée Poisot. *PLOS Computational Biology* 13 (5): e1005562. https://doi.org/10.1371/journal.pcbi.1005562.

Demichev, V., Messner, C. B., Vernardis, S. I., Lilley, K. S., & Ralser, M. (2019). DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. In Nature Methods (Vol. 17, Issue 1, pp. 41–44). Springer Science and Business Media LLC. https://doi.org/10.1038/s41592-019-0638-x

Bruderer, R., Bernhardt, O. M., Gandhi, T., Xuan, Y., Sondermann, J., Schmidt, M., Gomez-Varela, D., & Reiter, L. (2017). Optimization of Experimental Parameters in Data-Independent Mass Spectrometry Significantly Increases Depth and Reproducibility of Results. In Molecular & Cellular Proteomics (Vol. 16, Issue 12, pp. 2296–2309). Elsevier BV. https://doi.org/10.1074/mcp.ra117.000314

Niu, L., Thiele, M., Geyer, P. E., Rasmussen, D. N., Webel, H. E., Santos, A., Gupta, R., Meier, F., Strauss, M., Kjaergaard, M., Lindvig, K., Jacobsen, S., Rasmussen, S., Hansen, T., Krag, A., & Mann, M. (2022). Noninvasive proteomic biomarkers for alcohol-related liver disease. In Nature Medicine (Vol. 28, Issue 6, pp. 1277–1287). Springer Science and Business Media LLC. https://doi.org/10.1038/s41591-022-01850-y

Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D., & Nesvizhskii, A. I. (2017). MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics. In Nature Methods (Vol. 14, Issue 5, pp. 513–520). Springer Science and Business Media LLC. https://doi.org/10.1038/nmeth.4256

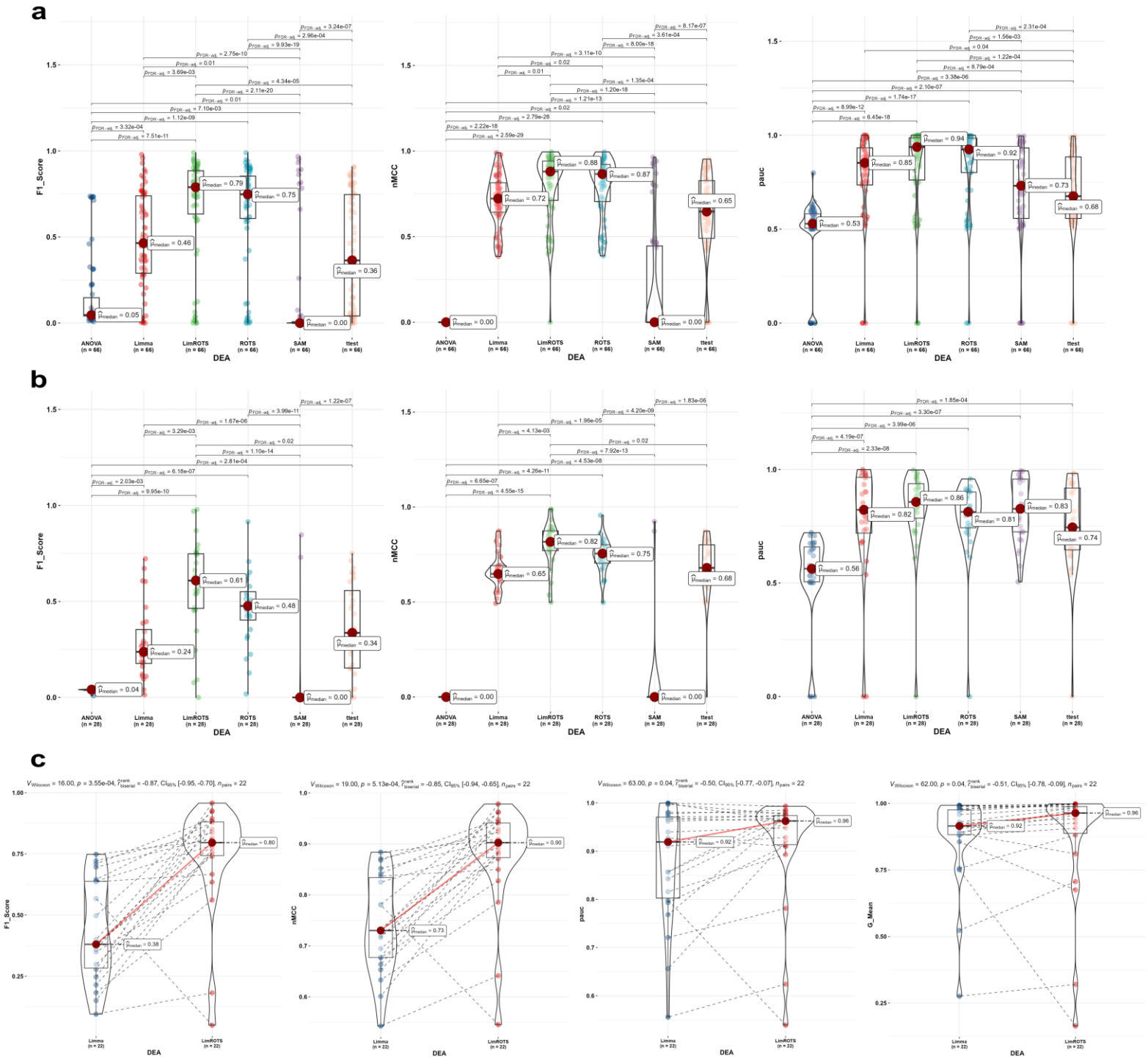
He, F., Aebersold, R., Baker, M. S., Bian, X., Bo, X., Chan, D. W., Chang, C., Chen, L., Chen, X., Chen, Y. J., Cheng, H., Collins, B. C., Corrales, F., Cox, J., E, W., Van Eyk, J. E., Fan, J., Faridi, P., Figeys, D., Gao, G. F., … π-HuB Consortium (2025). Author Correction: π-HuB: the proteomic navigator of the human body. Nature, 637(8046), E22. https://doi.org/10.1038/s41586-024-08555-x

Deng, Y.-T., You, J., He, Y., Zhang, Y., Li, H.-Y., Wu, X.-R., Cheng, J.-Y., Guo, Y., Long, Z.-W., Chen, Y.-L., Li, Z.-Y., Yang, L., Zhang, Y.-R., Chen, S.-D., Ge, Y.-J., Huang, Y.-Y., Shi, L.-M., Dong, Q., Mao, Y., … Yu, J.-T. (2025). Atlas of the plasma proteome in health and disease in 53,026 adults. In Cell (Vol. 188, Issue 1, pp. 253-271.e7). Elsevier BV. https://doi.org/10.1016/j.cell.2024.10.045Figures and supplementary files

Yan, Y., Schillemans, T., Skantze, V., & Brunius, C. (2024). Adjusting for covariates and assessing modeling fitness in machine learning using MUVR2. In M. Rattray (Ed.), Bioinformatics Advances (Vol. 4, Issue 1). Oxford University Press (OUP). https://doi.org/10.1093/bioadv/vbae051

Loo, R. T. J., Soudy, M., Nasta, F., Macchi, M., & Glaab, E. (2024). Bioinformatics approaches for studying molecular sex differences in complex diseases. In Briefings in Bioinformatics (Vol. 25, Issue 6). Oxford University Press (OUP). https://doi.org/10.1093/bib/bbae499

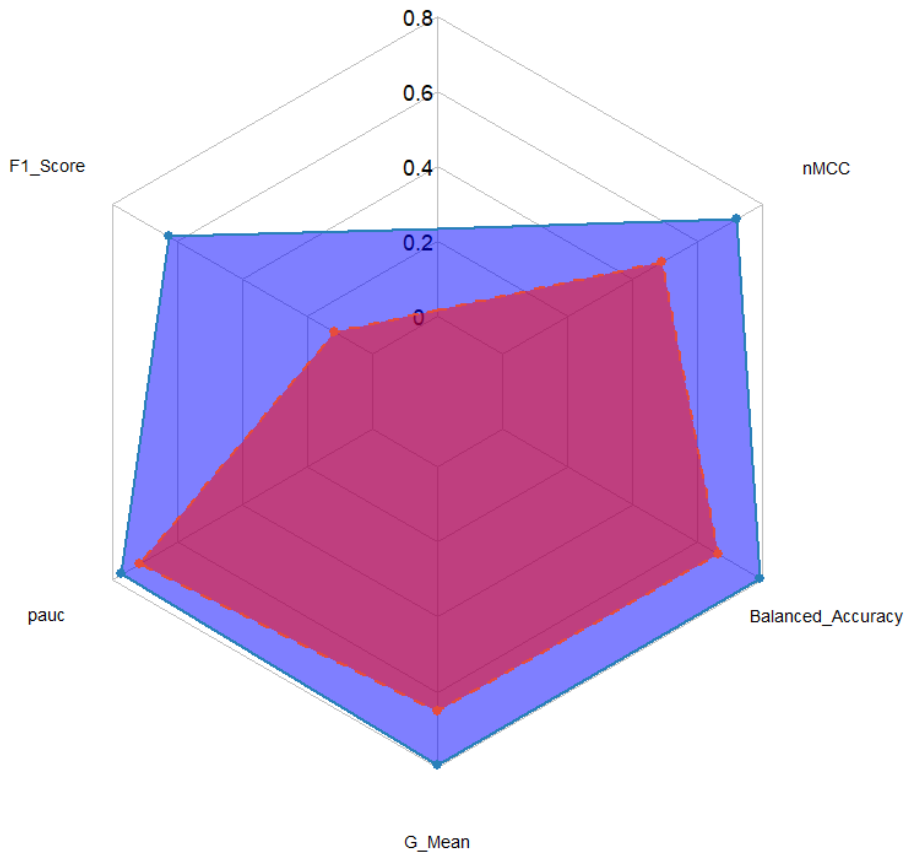
Figure 1: Benchmarking Performance Metrics; F1 score, nMCC, and pAUC5%.



a: Presents the results of the benchmarking performances (F1 score, nMCC, and pAUC5%) boxplots for each method from case study 1, with adjusted p-values for the methods showing statistically significant differences. b: Presents the results of the benchmarking performances (F1 score, nMCC, and pAUC5%) boxplots for each method from case study 2, with adjusted p-values for the methods showing statistically significant differences. c: Presents the results of the

benchmarking performances (F1 score, nMCC, and pAUC5%) boxplots between LimROTS and limma, from case study 3, with adjusted p-values for the methods showing statistically significant differences.

Figure 2: The spider plot illustrates the performance differences between LimROTS (blue) and limma (red) in case study 4, using five metrics: F1 score, nMCC, pAUC5%, G-mean, and balanced accuracy.



Supplementary file 1.

The supplementary file contains four tables, each presenting the performance metrics evaluated across all case studies for all methods.

Table S1: Results from case study 1.

Table S2: Results from case study 2.

Table S3: Results from case study 3.

Table S4: Results from case study 4.