NAR Genomics and Bioinformatics, 2026, **8**, lqag015
https://doi.org/10.1093/nargab/lqag015
Pipelines and Workflows for Biological Data Analysis

OXFORD

DEprot: a comprehensive R-package for the analysis of label-free quantitation mass spectrometry data

Nils Eickhoff[1,2], Liesbeth Hoekman[3], Onno Bleijerveld[3], Andries M. Bergman[1,4,5], Wilbert Zwart[1,2,6,*], Sebastian Gregoricchio[1,2,*]

[1]Division of Oncogenomics, The Netherlands Cancer Institute, Plesmanlaan 121, Amsterdam 1066 CX, The Netherlands
[2]Oncode Institute, Utrecht 3521 AL, The Netherlands
[3]Mass Spectrometry/Proteomics Facility, Netherlands Cancer Institute, Amsterdam 1066 CX, The Netherlands
[4]Division of Medical Oncology, The Netherlands Cancer Institute, Plesmanlaan 121, Amsterdam 1066 CX, The Netherlands
[5]Department of Medical Oncology, Amsterdam University Medical Center, Amsterdam 1081 HV, The Netherlands
[6]Department of Biomedical Engineering, Eindhoven University of Technology, Eindhoven 5600 MB, The Netherlands

*To whom correspondence should be addressed. Email: s.gregoricchio@nki.nl
Correspondence may also be addressed to Wilbert Zwart. Email: w.zwart@nki.nl

Abstract

In recent years, studies investigating protein–protein interactions on a proteome-wide scale have increased exponentially. These developments were achieved through implementation of new techniques, as well as improvements in mass spectrometry hardware. However, issues specific to this type of technique, such as distinct imputation of random and not-at-random missing proteins, have not been completely addressed. Here, we present *DEprot*, an R-package providing a comprehensive toolset for end-to-end analysis and visualization of proteomics data. Moreover, we implemented a serial imputation strategy for handling non-random missing values, particularly suitable for protein enrichment techniques and knockout conditions.

Introduction

In recent years, in parallel to the improvement of mass spectrometry (MS) instruments, detection methods, and spectral processing algorithms, a number of new methodologies for protein enrichment studies have been developed. These novel methods include protein interactome detection, such as immunoprecipitation followed by mass spectrometry (IP-MS), and its derivatives. One of these novel methods is rapid immunoprecipitation mass spectrometry of endogenous proteins (RIME) [1]; a technique initially developed to investigate nuclear chromatin complexes, but also applicable for interrogation of cytoplasmic complexes [2]. In brief, biological material (e.g., cells or frozen tissues) is cross-linked, after which nuclei are isolated and chromatin is sheared by sonication. Subsequently, nuclear lysates are incubated with beads coupled with an antibody against the protein-of-interest, for which the protein interactome is aimed to be characterized. Beads are washed and, after on-bead tryptic digestion, peptides from the pulled-down proteins are typically processed for detection by label-free quantitation (LFQ) liquid chromatography-mass spectrometry (LC-MS).

Several tools are available for the analyses of proteomics data. The most often used are *MSnbase* [3], *MSstats* [4, 5], and *DEP/DEP2* [6, 7]. However, all these tools must be used in combination with other software to accomplish a comprehensive analysis of the data. In particular, these tools lack features dedicated to protein enrichment techniques, such as RIME, and the possibility to correct for data derived from different batches of MS runs.

To enable a comprehensive proteomics analysis in a "one-stop-shop" fashion, we developed *DEprot* (https://github.com/sebastian-gregoricchio/DEprot) [8, 9], a new R-package that aims to fill current computational gaps by implementing all essential tools required for the analyses of whole-cell proteomics (e.g., batch correction, data imputation, normalization, differential analyses, enrichment analyses, etc.), as well as protein-enrichment-based techniques, in one user-friendly package (Fig. 1).

Materials and methods

Estimation of LFQ intensities

Raw MS data were analyzed by MaxQuant (RIME, standard settings) [10] or DIA-NN (whole cell proteomics; robust LC high accuracy and MBR enabled) [11] for LFQ. MS/MS data were searched against the SwissProt human database, complemented with a list of common contaminants, and concatenated with the reversed version of all sequences. The maximum allowed mass tolerance was 4.5 ppm in the main search and 0.5 Da for fragment ion masses. False discovery rates for peptide and protein identification were set to 1%. Trypsin/P was chosen as cleavage specificity allowing two missed cleavages. Carbamidomethylation (C) was set as a fixed modification, while oxidation (M) and deamidation (N, Q) were

Received: November 5, 2025. **Revised:** January 13, 2026. **Accepted:** January 22, 2026

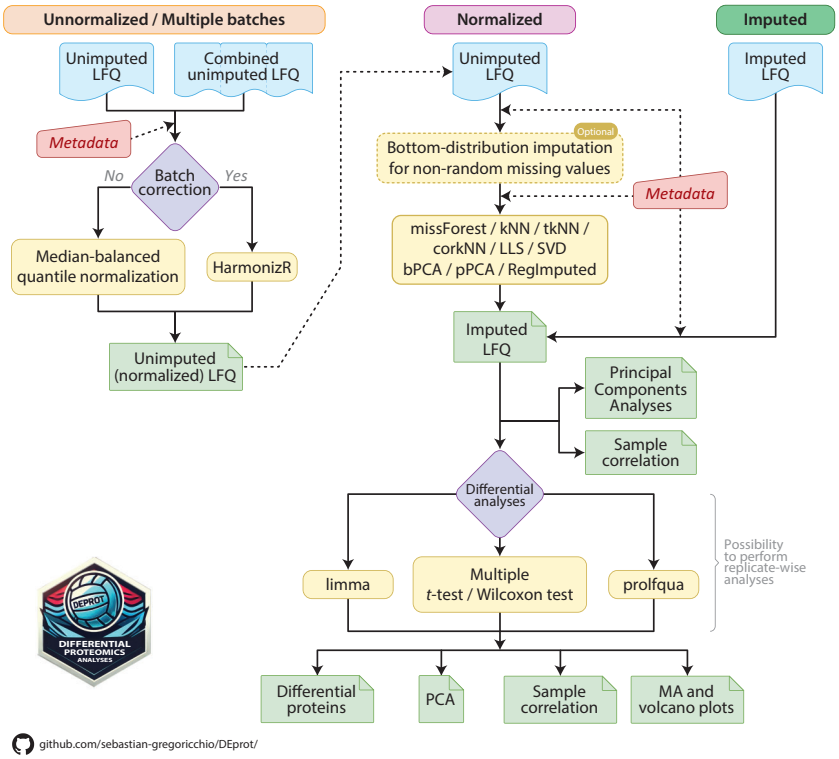

Figure 1. *DEprot* analyses workflow. Three possible types of analyses are facilitated through *DEprot*, depending on the input data provided (unnormalized/multiple batches, normalized, and imputed data), indicated by "wavy-bottom" blue boxes. Software/tools/functions used in each step are indicated by yellow rectangular boxes, while output files generated are indicated by green folded-corner boxes (text files and plots). Computational interactions are indicated by diamonds, while key user-defined parameters collected from the "metadata" table are contained in the red trapezoid box. kNN: k-Nearest Neighbors algorithm; tkNN: truncated kNN; corkNN: correlation kNN; LLS: Local Least Squares imputation; SVD: Singular Value Decomposition; PCA: Principal Component Analysis; bPCA: Bayesian PCA; pPCA: probabilistic PCA; LFQ: label-free quantitation.

used as variable modifications. LFQ intensities were $\log_2$-transformed in Perseus [12], after which proteins were filtered.

Details on the mass spectrometers used, MaxQuant/DIA-NN, Perseus, and database versions and parameters are available in Supplementary Table S1.

Batch correction and normalization

Within *DEprot*, it is possible to apply, to unimputed data, the Modified Balanced Quantile Normalization (MBQN) from the *MBQN* (v2.16.0) R-package developed by Brombacher *et al.* [13]. The modification balances the median (or mean) intensity of features (rows) that are rank invariant (RI) or nearly rank invariant (NRI) across samples (columns) before quantile normalization. This approach prevents an over-correction of the intensity profiles of RI and NRI features by classical quantile normalization and therefore supports the reduction of systematics in downstream analyses.

Batch correction of unimputed data can instead be performed using *HarmonizR* (v1.2.0), developed by Voß *et al.* [14]. This tool can handle datasets with large amounts of missing values—both of random and non-random type—as well as data derived by the combination of both DIA (Data-Independent Acquisition) and DDA (Data-Dependent Acquisition) mass spectrometry analysis approaches.

Imputation methods benchmarking

To select the best-performing imputation method, we made use of the *imputomics* [15] web tool and benchmarked all the ~40 methods provided on our RIME and proteomics data. We proceeded by using the unimputed LFQ tables for both datasets, removing all proteins that had at least one missing value (MV) across all samples, resulting in a table containing only known values. Furthermore, we applied a mask of artificial MVs that reflected the fraction of total MVs in the original dataset as well as the pattern within each sample group. To maintain the latter, we computed the number of proteins that had N replicates presenting MVs within a group and maintained this proportion in the artificial mask. We then submitted these masked tables to *imputomics* and computed the root mean square error (RMSE) for each method as well as Pearson's correlation between imputed and expected values (*residuals* $= I_i - O_i$). To compute the RMSE, we use the following formula:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (I_i - O_i)^2}{n}},$$

where I_i indicates an imputed value for the ith observation; O_i indicates the observed (measured) value for the ith observation; and n is the total number of observations (total missing values).

For each data type, we ranked the methods by RMSE and correlation and selected the top 5 performing methods in both datasets. We then excluded the methods that were redundant (using the same methods in background) or that did not allow for groups of different sample sizes.

Example data imputation

The imputation of missing values of RIME and whole-cell proteomics data involved two steps: (i) imputation of the missing not at random (MNAR) values using the *randomize.missing.values* function (75% of missing values in a group, selection of the data within bottom 3% of the distribution); and (ii) imputation of the missing completely at random (MCAR) values, using the *RegImpute* (RIME) [16] or *SVD* (proteomics) [17] algorithms.

Differential analyses

Three different modalities allow for the identification of differentially expressed proteins. The first modality involves a multiple Student's *t*/Wilcoxon ranked-sum test between the values of two groups—both in unpaired and paired mode. The *P*-value of the statistics can subsequently be corrected for multiple testing with canonical methods (e.g., Bonferroni, Benjamini–Hochberg, etc.).

The second approach employs the R-package *limma* (v3.60.6) [18]. If paired analyses are performed, the correlation between replicates (*limma::duplicateCorrelation*) is included in the fitting model (*limma::lmFit*). However, *limma* assumes that the value distribution of each sample is normal. This can be verified using the built-in *check.normality* function, which will perform an Anderson–Darling normality test, and plot density distributions and Q–Q plots for each sample using *ggpubr* (v0.6.2) R-package (https://rpkgs.datanovia.com/ggpubr).

The last method employs the package *prolfqua* (v1.3.9) [19], which uses linear models (including mixed-effects models) to assess protein abundance differences across experimental conditions.

For the results presented in this work, we used *limma* (v3.60.6), and defined proteins as "differentially expressed/enriched" when $|\log_2(\text{FoldChange})| \geq 1$ and $P_{adj} < 0.05$; "unresponsive" when $0.90 \leq \text{FoldChange} \leq 1.11$ and $P_{adj} \geq 0.05$; while remaining proteins were labeled as "null." The *P*-values were adjusted using the Benjamini–Hochberg (BH) method.

SAINT computation

To perform Significance Analysis of INTeractome (SAINT) [20], protein peptide–spectrum match (PSM) data for bait IPs (AR RIME) and corresponding control IPs (IgG RIME) were uploaded for determination of likely AR interactors using the *SAINTexpress* [21] computational tool through the CRAPome interface [22]. Proteins were considered enriched in AR RIME over IgG RIME when the SAINT score was ≥0.95.

Overrepresentation analyses (ORA)

Enrichment for subunits of specific complexes was assessed by ORA against the CORUM (version 5.0) protein complexes database [23]. For this, analyses were performed using *clusterProfiler* (v4.12.6) [24] and results were subsequently visualized using *aPEAR* (v1.0.0) [25].

GeneSet enrichment analyses

GeneSet enrichment analyses (GSEA) are performed by ranking the proteins depending on their $\log_2$(Fold Change) obtained from differential analyses or by computing the correlation between the two groups (for each protein). These two different ranking methods can be compared using the built-in *compare.ranking* function. GSEA analyses were performed using *clusterProfiler* (v4.12.6) [24]. Visualization of networks is performed using *aPEAR* (v1.0.0) [25].

For GSEA shown in Fig. 7, we used the "Hallmarks" [26] dataset from the Human MSigDB Collections (*msigdbr* (v25.1.1) using *P*- and *q*-value cutoffs of 0.05. GSEA plot was generated using *plot.gsea* function from the *Rseb* R-package (v0.3.4) [27], which is also implemented in *DEprot*.

Statistical analysis

All analyses and software implementation were performed in R version 4.4.3 (R Core Team 2020, https://www.R-project.org).

Results

Case study: AR interactome analyses in prostate cancer

To illustrate the functionalities of *DEprot*, we reanalyzed RIME data derived from Eickhoff *et al.* [28]. This dataset involves experiments of RIME for the Androgen Receptor (AR) in the LNCaP prostate cancer cell line. Notably, these data have been generated in three different batches, and each batch has been analyzed independently using MaxQuant [10] (see the "Materials and methods" section and Supplementary Table S1).

More specifically, the data include LNCaP and LNCaP-abl (a derivative cell line resistant to hormone deprivation, yet expressing and fully dependent on AR signaling) prostate cancer cells. LNCaP cells are either wild-type (WT) or monoclonal knockout (KO) for the E3 ubiquitin-protein ligase TRIM33 (condition referred to as TRIM33-5#MC-C2)—a bona fide AR interaction partner involved in the regulation of a distinct subset of AR-responsive genes—and respective non-targeting control (referred to as NT-3). Cells were cultured in medium supplied with regular fetal bovine serum (FBS) or were hormone-deprived for 3 days prior to stimulation with 5 nM R1881 (synthetic androgen) for 4 h. For each condition, immunoprecipitation with AR or IgG (negative isotype control) antibodies was performed in four biological replicates. Pulled-down proteins were subsequently detected using LFQ-based mass spectrometry (for details see the "Materials and methods" section and Supplementary Table S1).

Batch correction and normalization

Technologies like proteomics face significant experimental variability due to low data completeness and different quantification platforms and setups [29]. On the other hand, the possibility of combining multiple acquisitions would enhance the usability of publicly available data. For this aim, a batch correction must be applied to avoid misinterpretation of the results due to skewed data. This data transformation can be achieved within *DEprot* using a recently developed algorithm implemented in the *HarmonizR* R-package [14].

As aforementioned, the example unimputed AR-RIME data were generated in three different batches. Despite each batch being internally normalized, the differences in LFQ-intensity distribution across batches are substantial (Fig. 2A). These discrepancies are heavily reduced after that *HarmonizR* is applied (Fig. 2B), rendering the data suitable for direct comparison between samples regardless of the original batch.

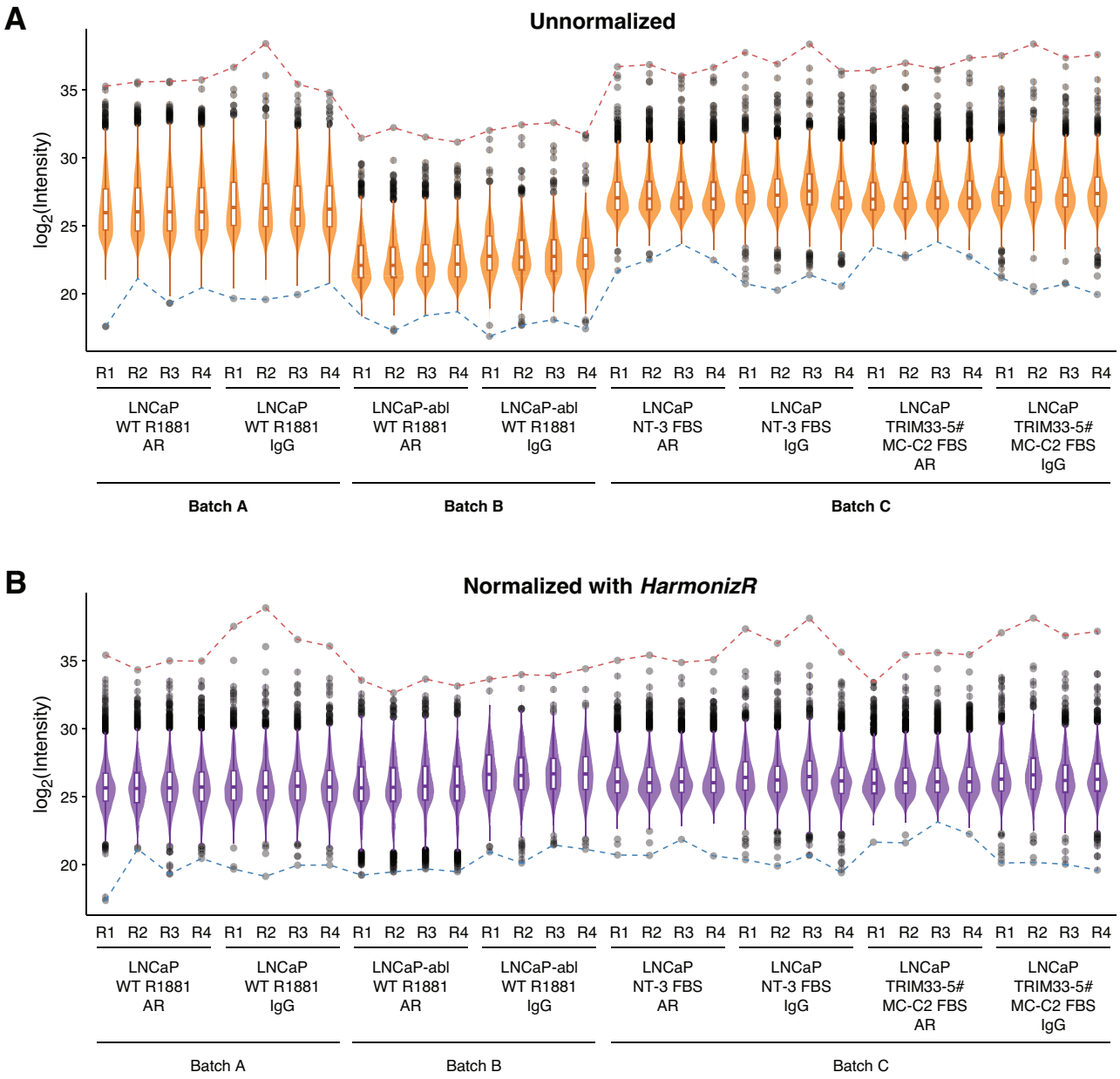

Figure 2. Effects of *HarmonizR* batch correction on LFQ-intensity distribution. Combination of violin and boxplots of the raw (**A**), or batch-corrected (**B**) unimputed LFQ intensity distribution for each sample in the RIME dataset. In both panels, the minimum and maximum values of each sample are connected by a blue or red dashed line, respectively. Boxplots depict the quartiles with the central horizontal line indicating the median. Black dots represent the outliers. R1, R2, R3, and R4 indicate the four biological replicates.

Data imputation

A common challenge in the analyses of proteomic data is the presence of a high number of missing values (MVs) (Supplementary Fig. S1A). The origin of these MVs can be due to "not-random" events, such as KO proteins or not expressed proteins—that *a priori* will not be measured—and are defined as MNAR values. For the specific case of immunoprecipitation techniques (such as RIME), another case of MNAR values can be found in the nature of the IgG/isotype control. Indeed, the latter is used as negative control for immunoprecipitation (IP), designed to capture only nonspecific background interactions. This implies that for the isotype control the variability between replicates will be much higher when compared to the targeted IP samples (Fig. 3A). Furthermore, a missing protein is, with a certain degree of confidence, not pulled down rather than not detected.

On the other hand, due to technical limitations in the MS-based protein detection, proteomics data also present many "random" MVs (reviewed in [30, 31])—defined as MCAR values. These MVs can have numerous causes, including sample preparation and processing or ion fragmentation and detection, which is especially the case for peptides at the limit of detection or from lowly expressed proteins.

Overall, the presence of too many MVs reduces the power of differential analyses for proteomics data. To overcome this hurdle, several methods to impute MVs have been developed

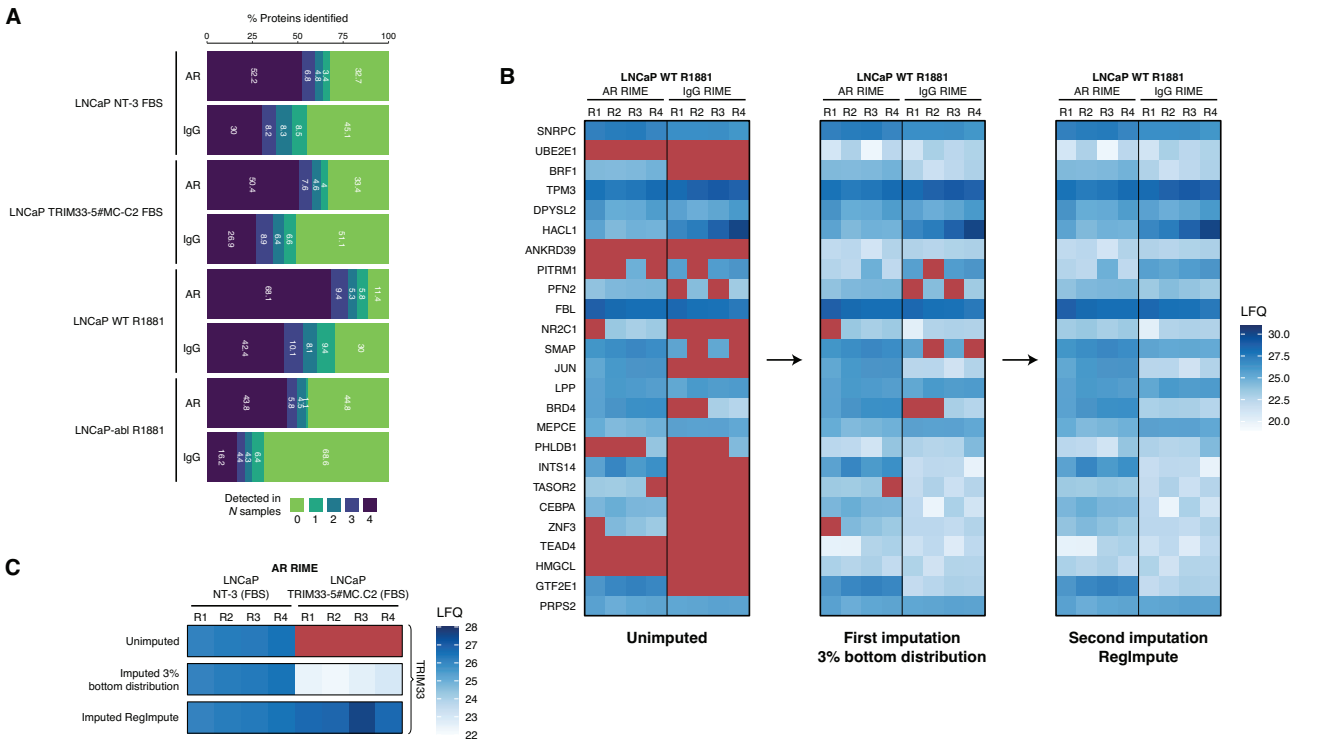

Figure 3. Comparison of imputation methods. (**A**) Stacked bar plot showing the percentage of proteins, out of the total (n = 3,153), identified in each condition. Color scale indicates the number of samples within the group of replicates in which a protein has been identified (0 = none, 4 = all biological replicates). (**B**) Heatmap showing the LFQ intensity of a random selection of 25 proteins on raw unimputed data (left), first-step imputation using the sampling of values from the global distribution tail (middle), and second-step imputation using the *RegImpute* algorithm (right). Red squares indicate missing values. (**C**) Heatmap of LFQ intensities for the protein TRIM33, showing raw unimputed values (top), imputed using the tail distribution (middle), and imputed using *RegImpute* alone (bottom). Red squares indicate missing values.

[32, 33]. However, none of these methods distinguishes the different types of MVs during the imputation process. Here, we propose to use a combination of methodologies in a two-step approach in order to handle first the MNAR values and sequentially the MCAR ones (Fig. 3B).

For the first step (MNAR), within each group (e.g., the replicates of one condition), values are randomly assigned, sampled from a pool represented by the bottom X% (default: 3%) of the whole dataset distribution (normally distributed) (Fig. 3B, left and middle panels). Potential MNAR values can be identified by missingness across all, or the majority of, replicates within a group of samples as compared to another group (here 3 or 4 MVs across 4 replicates in a group). Further imputation (second step) of remaining MCAR values—in this experiment one to two MVs within a group—can be performed using one of the 9 different methods implemented in *DEprot*: (i) *missForest* (default), developed by Stekhoven & Buehlmann [34]. This tool will impute the MVs and assign an estimated value. It also yields an out-of-bag (OOB) imputation error estimate (general or for each sample). This method can be run on multiple cores to save computation time; (ii) *k-Nearest Neighbors* (kNN) algorithm [35]; (iii) truncated kNN (tkNN) [15]; (iv) correlation kNN (corkNN) [15]; (v) *Local Least Squares* (LLS) imputation [36], using the Pearson's correlation coefficient; (vi) Singular Value Decomposition (SVD) [17]; (vii) Bayesian principal component analyses (bPCA) [17]; (viii) probabilistic PCA (pPCA) [17]; and (ix) *RegImpute*, a machine learning-based method from the DreamAI package [16].

These methods were selected on the comparison of 40 different imputation methods using the *imputomics* web tool [15]. From both batch-corrected RIME and whole-cell proteomics data (see the "Whole-cell proteomics data" section for more details), we filtered out all the proteins that presented at least one MV. Then, in the resulting matrix—containing only known values—we introduced *in silico* the same fraction of MVs identified in the original data, reflecting the sample proportions within each sample group (i.e. condition/treatment). Subsequently, we imputed 40 sets of values using *imputomics* [15] and estimated the RMSE. All methods were then ranked by RMSE, and the top 5 for each data type were implemented in *DEprot* (Supplementary Fig. S2). Of note, we excluded methods that were using the same algorithm for their calculations (e.g., *metabimpute bpca* [37] uses *pcaMethods bpca*), as well as those methods that do not allow for different sample sizes among groups (e.g., *metabimpute* [37]). We also implemented a function that can perform similar analyses and estimate an RMSE for all 9 methods implemented in *DEprot* directly on the user-provided data. This comparison identified the *RegImpute* [16] method as the best performing for our RIME data set (Fig. 3B, right panel; Supplementary Fig. S3).

The consequences of using our combined imputation approach, as opposed to individual imputation, are evident for KO proteins (Fig. 3C). Indeed, application of *RegImpute* alone leads to an artificial overestimation of the TRIM33 protein in the monoclonal KO: LFQ intensities upon imputation are comparable across all samples, while TRIM33 is in fact absent in the KO condition [28] (Fig. 3C). This also stands at

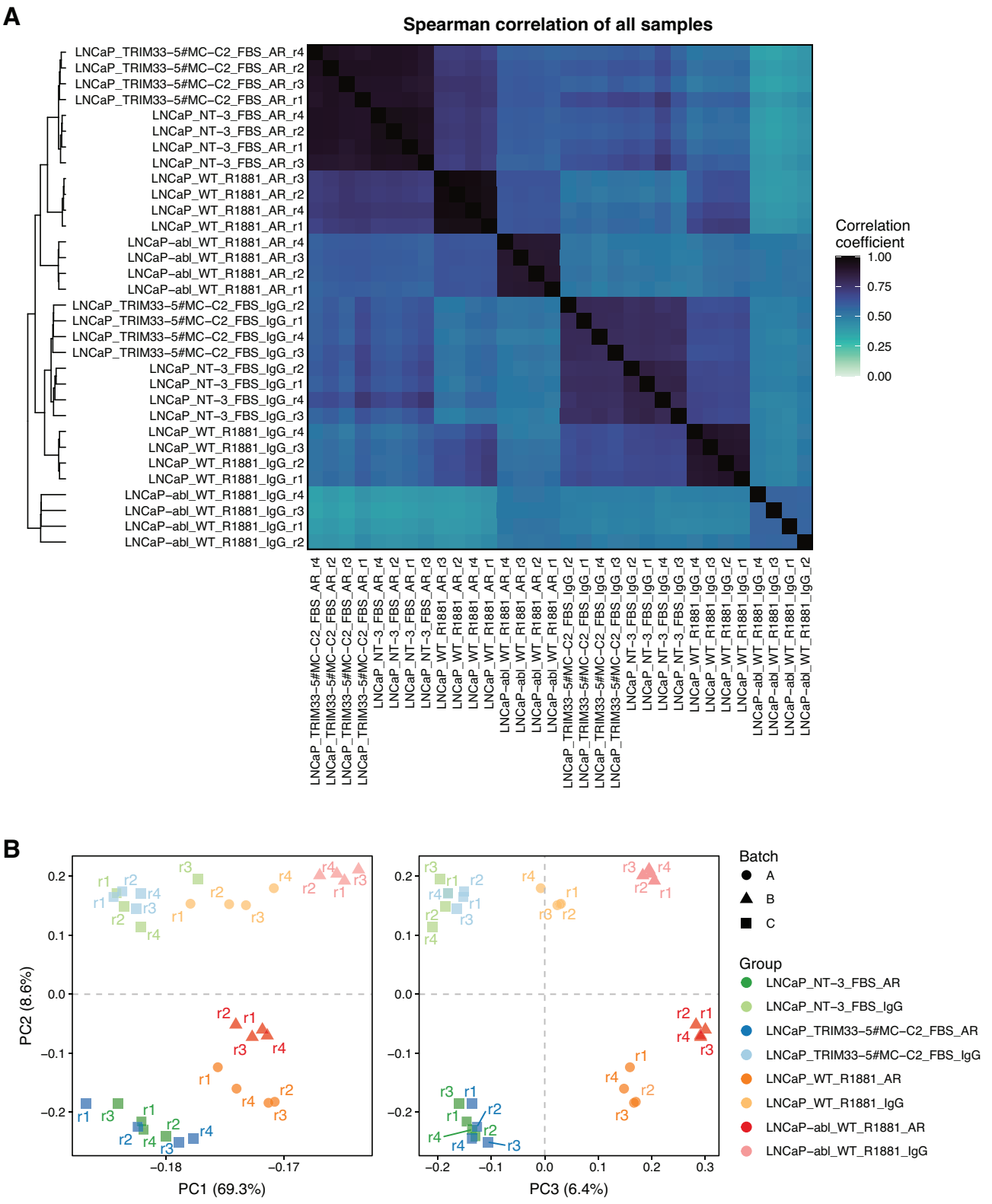

Figure 4. Quality controls of RIMEs. (**A**) Spearman's correlation heatmap of all samples. The dendrogram uses (1 – *correlation*) as distance, and the clustering method is "complete". (**B**) Scatter plots depicting the results of principal components (PC) analyses. Left: PC1-versus-PC2; right: PC2-versus-PC3. Dark colors indicate the AR RIMEs, while light colors indicate the IgG negative control. The original batches are indicated by different dot shapes, while the replicate number is labeled in vicinity of the corresponding dot.

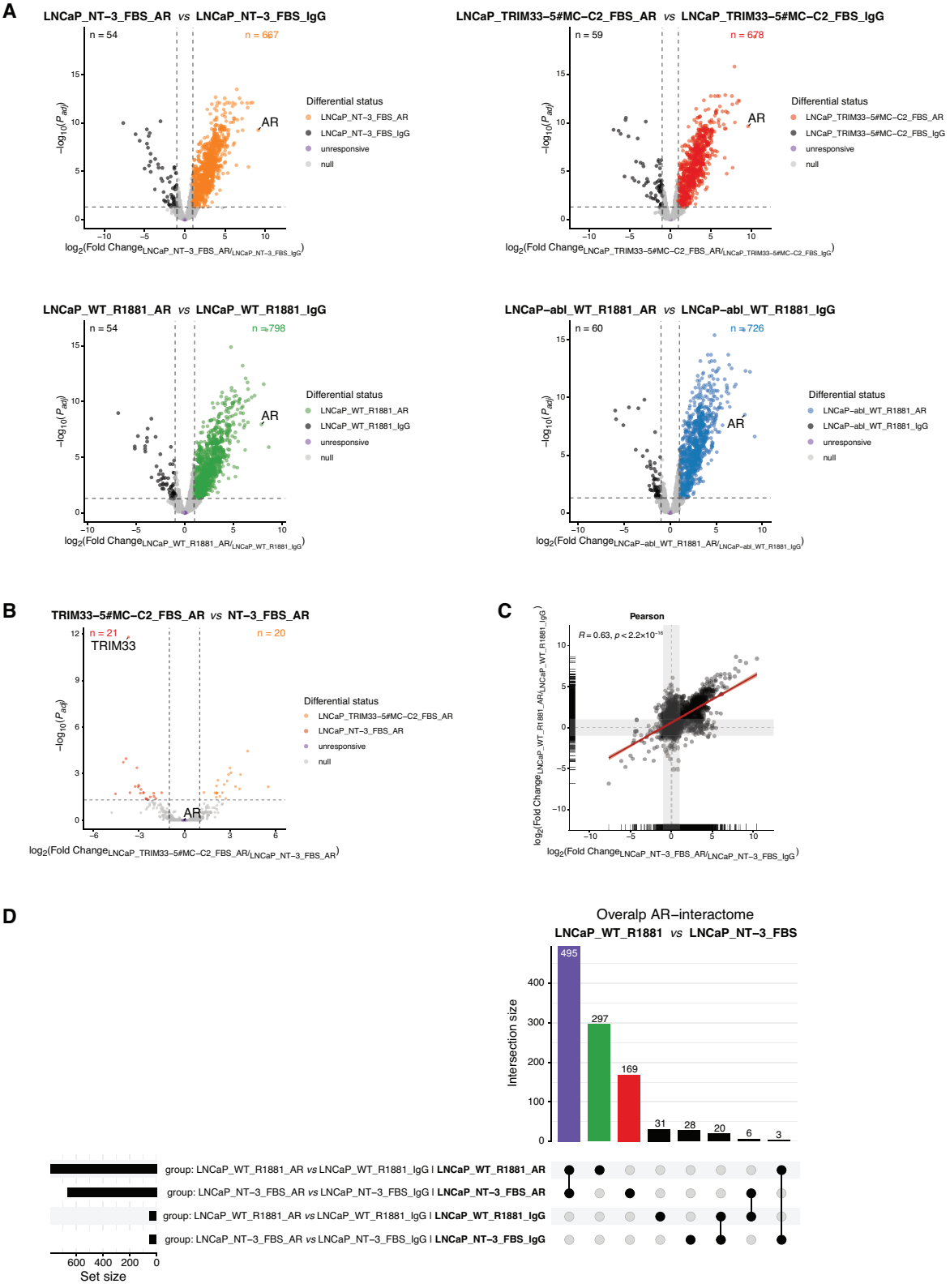

Figure 5. Differential analysis results for RIMEs. (**A**) Volcano plots depicting protein enrichment for each group comparing AR RIME versus IgG RIME. Proteins are defined as differentially enriched when $|\log_2(\text{Fold Change})| \geq 1$ and $P_{adj} < 0.05$. Unresponsive proteins present a (linear) fold change in the [0.90–1.11] range and $P_{adj} \geq 0.05$. Remaining proteins are labeled as "null". (**B**) As in panel (A), volcano plot for the comparison between AR-RIME in LNCaP NT-3 versus TRIM33-KO. (**C**) Scatter plot of the $\log_2$(Fold Change AR/IgG) between the LNCaP WT (R1881 stimulated) and LNCaP NT-3 (cultured in FBS). Pearson's coefficient correlation is indicated. The red line indicates linear regression (glm, y ~ x). Gray zones mask values with $|\log_2(\text{Fold Change})| \leq 1$ ("not differential"). *R* indicates Pearson's correlation coefficient. (**D**) Upset plot of the differential proteins identified in the AR-versus-IgG RIME comparisons in the LNCaP-WT (R1881 stimulated) and LNCaP-NT-3 (cultured in FBS) groups.

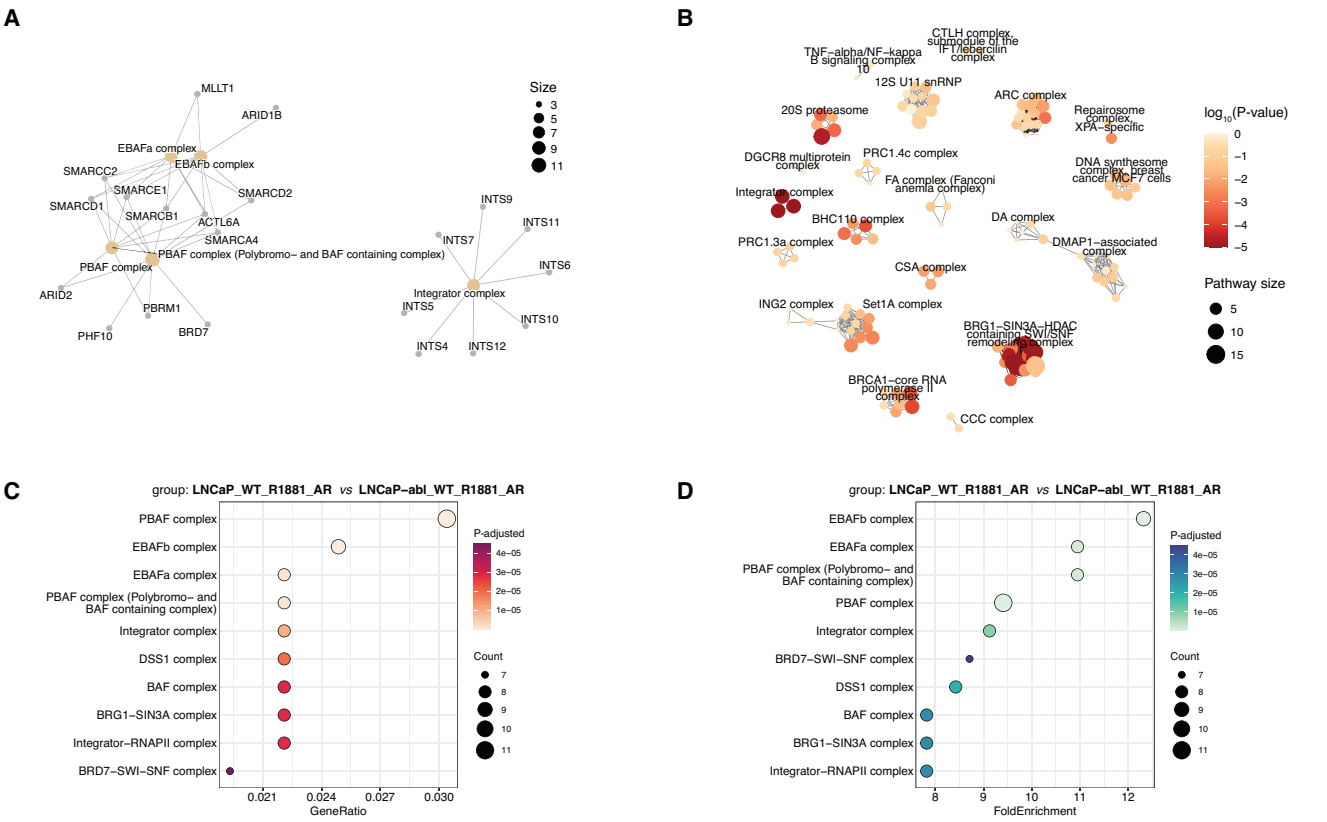

Figure 6. Overrepresentation analysis results for AR RIME studies in LNCaP-versus-LNCaP-abl cells. AR interactors in WT LNCaP cells, identified by comparison with LNCaP-abl, were used for ORA against the CORUM (v5.0) protein complexes database ($P < 0.05$). (**A**) Protein complexes enriched in the LNCaP-WT AR-RIME studies are indicated by a colored dot, whose size reflects the number of proteins composing the corresponding complex. The members of a complex are indicated by gray dots and connected by gray segments to the colored complex dot; the same protein can belong to multiple complexes. (**B**) In this protein complex network, each dot is a protein complex of the CORUM database. Complexes are clustered by enrichment of their members, which are visualized by the color of the dots (indicating the *P*-value of the enrichment). Dot plot depicting the protein ratio (**C**) or fold enrichment (**D**) of the top-enriched complexes.

a more global level; *RegImpute* alone is not able to properly impute non-random missing values and tends to "homogenize" the intensities around the mean of the distribution (Supplementary Fig. S1B).

The performance of data imputation can be ultimately assessed by inter-sample correlation (Fig. 4A and Supplementary Fig. S4A) and principal component analyses (PCA) (Fig. 4B and Supplementary Fig. S4B). For the example of our AR RIME data, we can observe a distinct segregation of AR versus IgG RIME as well as the cell culturing conditions (FBS or R1881 stimulation).

Differential abundance analyses

To determine which proteins are differentially expressed, or enriched, between two groups, *DEprot* provides three possible methods: (i) multiple t-/Wilcoxon tests, (ii) *limma* [18] package (originally developed for microarray expression data), and (iii) *prolfqua* [19] package. While the first two methods require data with only complete observations (imputed data), *prolfqua* can also be applied on unimputed data. Furthermore, as *limma* assumes a normal distribution of the data, we implemented in *DEprot* a function that allows for the application of the Anderson–Darling normality test for each sample and plots the corresponding Q–Q plot (Supplementary Fig. S5) and density distribution (Supplementary Fig. S6).

Due to the elevated number of MVs present in the RIME data, we tested multiple combinations of imputation strategies and differential expression methods (Supplementary Fig. S7): (a) *RegImpute* + *limma*, (b) bottom distribution: *RegImpute* + *limma*, (c) bottom distribution: *RegImpute* + t-test, (d and e) *prolfqua* on unimputed data, using either a linear model (lm) or a linear mixed-effects model (lmer) strategy. To compare the used methods with an orthogonal approach to identify bait interactors, we computed SAINT scores [20] for the AR RIME (over IgG control) in LNCaP-Abl (Supplementary Table S2) and LNCaP TRIM33-KO (Supplementary Table S3). Given a specific bait, SAINT assigns the number of identified peptides for each interactor to a probability distribution, which is then exploited to estimate the likelihood of a true interaction (SAINT score $\geq$0.95). Applying the different strategies described earlier, we defined different sets of AR interactors—as enriched in AR RIME over IgG RIME—(Supplementary Fig. S7A). Comparing the fraction of SAINT-defined interactors that are also identified as AR interactors in each strategy, we observed that the "bottom distribution: *RegImpute* + *limma*" method is the one that can retrieve the largest number of SAINT-defined interactors (93-99% range) (Supplementary Fig. S7B). Importantly, these results show that *prolfqua* not only identifies a smaller proportion of SAINT-defined interactors, but also that many proteins were not evaluable with *prolfqua*—TRIM33, for instance—due to its

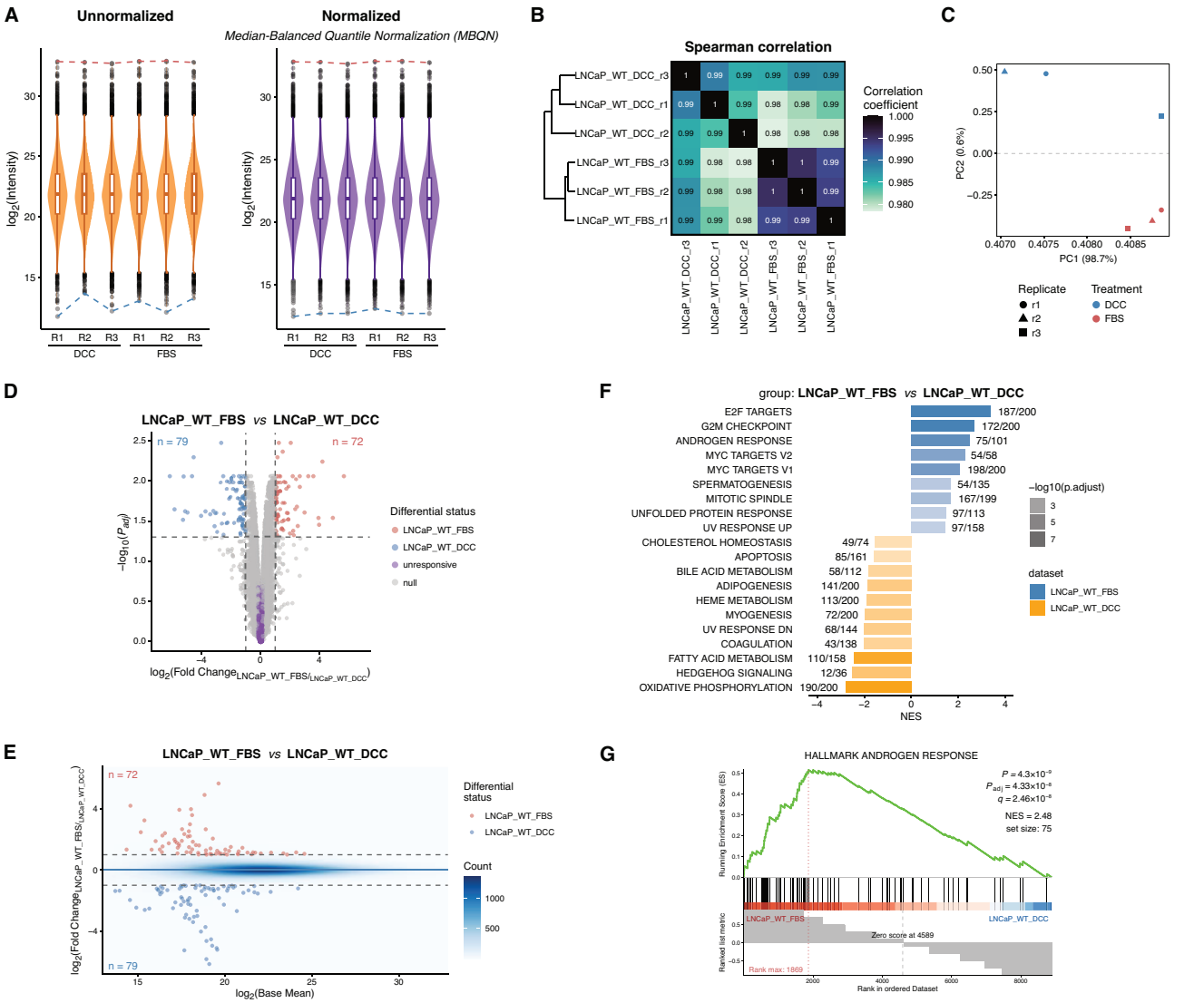

Figure 7. Whole-cell proteomics analysis results. (**A**) Combination of violin and boxplot of unimputed raw (left, orange) or MBQN-normalized (right, purple) unimputed LFQ intensity distribution for each sample in the whole-cell proteomics dataset. Minimum and maximum values of each sample are connected by a blue or red dashed line, respectively. Boxplots depict the quartiles with the central horizontal line indicating the median. Black dots represent the outliers. R1, R2, and R3 indicate the three biological replicates. (**B**) Spearman's correlation heatmap of all samples. The dendrogram uses (1 – *correlation*) as distance, and the clustering method is "complete". (**C**) Scatter plot depicting the results of principal components (PC) analyses. The biological replicates are indicated by different dot shapes. Volcano plot (**D**) and MA-plot (**E**) depicting protein enrichment for the FBS-versus-DCC comparison. Proteins are defined as differentially enriched when $|\log_2(\text{Fold Change})| \geq 1$ and $P_{adj} < 0.05$. Unresponsive genes present a (linear) fold change in the [0.90–1.11] range and $P_{adj} \geq 0.05$. Remaining proteins are labeled as "null". (**F**) Bar plot representing the normalized enrichment score (NES) for the GSEA analyses performed on the cancer "Hallmarks" dataset. The numbers at the end of the bars indicate the number of proteins identified in the total amount of proteins belonging to the dataset (i.e., "Hallmarks"). (**G**) GSEA plot for the Androgen Response Hallmark pathway. Protein ranking is based on protein correlation.

absence across KO samples (MNAR). Furthermore, it is possible to conclude that omission of the bottom-distribution imputation leads to an important underestimation of protein interactors. Altogether, we conclude that the preferable approach to identify RIME interactors is represented by the "bottom distribution: *RegImpute* + *limma*" strategy and that *prolfqua* is too conservative when applied to RIME data.

Hence, using *limma* (bottom distribution + *RegImpute* imputation strategy), we first assessed the quality of RIME datasets comparing each AR IP with the negative isotype control (IgG) (Fig. 5A). As expected, the number of enriched proteins in the negative control is lower than in the AR RIME. Furthermore, AR itself is always found among the top-enriched proteins in each comparison (Fig. 5A). Another internal control is represented by the comparison between AR RIMEs performed in LNCaP NT-3 or TRIM33-KO. In this case, as expected, TRIM33 is enriched only in the NT-3 condition, while depleted in the TRIM33-KO, yet AR enrichment is unchanged between both cell lines (Fig. 5B).

The two groups represented by LNCaP-WT (R1881 stimulated) and LNCaP-NT-3 (cultured in FBS) should show only limited differences in protein enrichments in AR-RIME, since both conditions are genetically quasi-identical, and in both cases AR is activated. To test this hypothesis, we correlated the fold change enrichment of AR-versus-IgG RIME for the two groups (Fig. 5C). The high Pearson's correlation ($R = 0.65$,

Table 1. Feature comparison for different R-based proteomics packages

	DEprot (this work)	*DEP/DEP2* [5, 6]	*MSstats* [4, 5]	*msqrob2* [39]	*prolfqua* [19]
Typical input formats	LFQ tabular inputs (e.g., MaxQuant-style)	MaxQuant, IsobarQuant; tabular inputs	Converters for many tools: PD, MaxQuant, Skyline, Spectronaut, OpenMS	Peptide or protein intensities (QFeatures)	Tidy long format
Batch-effect handling	Yes, via harmonizR	No	Yes, handled in design	Yes, via covariates/random effects	Yes, via covariates
Normalization options	MBQN/HarmonizR	Variance stabilization	Global, equalize medians, quantile	Equalize medians, quantile, vsn	Global, vsn, robust scaling
Missing-data handling	Imputation-based (9 methods) + 2-step approach for MNAR values	Imputation (e.g., MinProb, QRILC, knn)	Model-based with optional imputations; censored/ties handled in modeling	Hurdle/robust mixed models	Multiple strategies; left-censored/MAR workflows
Imputation methods benchmarking	Yes	No	No	No	No
Peptide-to-Protein summarization	Works at protein level (from MaxQuant proteinGroups)	Works at protein level (from MaxQuant proteinGroups)	Tukey's median polish	Yes, peptide-level modeling possible	Aggregation included
Differential analyses	Multiple t-test/Wilcoxon, limma, linear, and linear mixed-effects models (via prolfqua)	limma	linear mixed-effects models	robust linear mixed models	linear and linear mixed-effects models
Enrichments	Yes (ORA/GSEA, networks)	No	No	No	No

$P < 2.2 \times 10^{-16}$), as well as the large proportion of differential proteins detected (Fig. 5D), confirmed the similarity in AR-interactome composition between the two culturing conditions and cell lines.

Altogether, we show that *limma*, applied to the double imputed data, successfully detects expected differences in protein enrichment.

Overrepresentation analyses for AR-interactome composition

RIME allows for the unbiased characterization of proteins interacting with the target protein, in our example AR. To determine how long-term hormone deprivation impacts the AR-interactome composition, we used the differentially enriched proteins identified in the AR-RIME in WT LNCaP compared to the LNCaP-abl to perform ORA of members of specific protein complexes. For these analyses, we used the CORUM (version 5.0) protein complexes database [23]. Using *aPEAR* [25], we defined networks of both proteins (Fig. 6A) and complexes (Fig. 6B). These analyses, in agreement with previous findings [38], underlined a strong enrichment for chromatin remodeling proteins (Fig. 6C and D) that are lost from the AR complex in conditions of long-term hormone deprivation.

Whole-cell proteomics data

In the aforementioned analyses, we illustrated the advantages of *DEprot* in the analyses of protein pull-down techniques, such as RIME. Nevertheless, *DEprot* can also be applied to the analyses of whole-cell proteomics data. Hereafter, we will present an example of whole-cell proteomics analyses using data generated by Eickhoff *et al.* [28]. More precisely, LNCaP prostate cancer cells were cultured either in media supplemented with regular FBS or hormone-stripped serum (using dextran-coated charcoal, DCC) for 3 days.

The data display a relatively low number of missing values (Supplementary Fig. S8A) and high protein detection reproducibility (Supplementary Fig. S8B). However, the LFQ intensity distributions present a bias at the tails (Fig. 7A, left panel). Indeed, proteomics data displays protein intensities that are consistently high/low across samples, therefore falling in the tails of the distribution. Further, MVs originating from proteins below the detection limit may result in a virtual over-representation of high values from more abundant proteins. Normalization with conventional methods (e.g., non-balanced quantile normalization, variance-stabilizing transformation, etc.) might lead to an underestimation of the variance with an increased risk of false positive discoveries. With the ultimate goal to correct for this "tail bias," we implemented in *DEprot* a function that applies a Median-Balanced Quantile Normalization (MBQN) [13], which allows for the unbiased estimation of the sample variance using a tail-robust approach (Fig. 7A, right).

As previously described for RIME, we computed the RMSE for all 9 imputation methods implemented in *DEprot* and determined that, for this dataset, *SVD* is the best-performing algorithm (Supplementary Fig. S9). Upon dual imputation (bottom distribution followed by *SVD*), sample correlation (Fig. 7B) and PCA (Fig. 7C) displayed a segregation of the samples by condition.

In prostate cancer cells, we expect hormone deprivation to impact nuclear receptor response, i.e. androgen response. To test this hypothesis, we performed differential analyses using *limma* and compared the FBS condition to the DCC one, and identified 40 proteins upregulated and 33 downregulated in FBS condition (Fig. 7D and E). We then performed GSEA using the cancer "Hallmarks" dataset (Fig. 7F). Among the top upregulated pathways in FBS condition, we identified—as expected—the androgen response (Fig. 7F) with a normalized enrichment score (NES) of 2.58 ($P = 1.3 \times 10^{-9}$, Fig. 7G).

Altogether our results recapitulated biological effects of hormone deprivation in prostate cancer cells, demonstrating the employability of *DEprot* in the analyses of whole-cell proteomics.

Discussion

DEprot is an R-package for "end-to-end" analyses of LFQ-based proteomics data. This package implements data normalization, batch correction, 9 distinct imputation methods, several differential analysis methods, and downstream enrichment analysis and visualization tools. Within this framework, the operator can explore the different available methods and check their effectiveness within the package. Furthermore, we propose a new dual imputation strategy to improve differential analyses in the case of MNAR values; a known challenge in IP-MS data when comparing to IgG controls or when studying genetic KOs. As all methodologies come with advantages and drawbacks, we provide an overview table to compare the principal features of *DEprot* with other commonly used R-based proteomics packages (Table 1).

All together, we provide a user-friendly analysis tool with simple input requirements that allows for flexibility in downstream analysis.

Limitations

DEprot is reliant on aggregated LFQ values per protein as input, and therefore individual peptide and spectral information cannot be evaluated. Despite the high overlap with SAINT scores, there are also proteins that have not been identified as interactors using SAINT, but are nonetheless identified with our combined imputation and *limma*-based strategy. This could be due to the stringency in SAINT but also the occurrence of false positives is possible. Therefore, careful investigation of the data pre- and post-imputation as well as functional validations of potential new interactors is crucial.

Acknowledgements

We thank members of the Zwart and Bergman labs for valuable feedback, suggestions, and input. We would like to acknowledge the Research High Performance Computing (RHPC) facility of the Netherlands Cancer Institute (NKI) for enabling us to perform all the computations required to analyze the data and develop this tool, the Mass Spectrometry/Proteomics facility for the support in the proteomics data handling, and the Biostatistics department for the support on the statistical application.

Author contributions: Nils Eickhoff (Conceptualization [equal], Data curation [lead], Formal analysis [supporting], Investigation [lead], Methodology [equal], Software [supporting], Validation [lead], Writing – original draft [equal], Writing – review & editing [lead]), Liesbeth Hoekman (Data curation [supporting], Formal analysis [supporting]), Onno Bleijerveld (Data curation [supporting], Formal analysis [supporting], Supervision [supporting]), Wilbert Zwart (Project administration [supporting], Resources [lead], Supervision [supporting], Writing – original draft [supporting], Writing – review & editing [equal]), and Sebastian Gregoricchio (Conceptualization [lead], Data curation [equal], Formal analysis [lead], Investigation [equal], Methodology [equal], Software [lead], Supervision [lead], Validation [equal], Visualization [lead], Writing – original draft [lead], Writing – review & editing [lead])

Supplementary data

Supplementary data is available at NAR Genomics & Bioinformatics online.

Conflict of interest

None declared.

Funding

This work was supported by Oncode Institute. The Zwart lab is part of the Oncode Institute, which is partly funded by the Dutch Cancer Society. Research at the Netherlands Cancer Institute is supported by institutional grants of the Dutch Cancer Society and of the Dutch Ministry of Health, Welfare and Sport. This work received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 813599. L.H. and O.B. are supported by the X-omics Initiative (project 184.034.019), part of the NWO National Roadmap for Large-Scale Research Infrastructures.

Data availability

Mass spectrometry data, derived from Eickhoff *et al.* [28], are available at the ProteomeXchange Consortium [40] through the PRIDE [41] partner repository with the identifier PXD058914. The *DEprot* R-package is publicly available as a GitHub repository at https://github.com/sebastian-gregoricchio/DEprot [9]. Codes used to generate the figures and tables presented in this work are available on Zenodo (doi: 10.5281/zenodo.17472993) [8]. Additional examples and functionalities for *DEprot* can be found at the publicly available tutorial (https://sebastian-gregoricchio.github.io/DEprot/doc/DEprot.overview.vignette.html) and technical web manual (https://sebastian-gregoricchio.github.io/DEprot/manual/index.html).

References

1. Mohammed H, Taylor C, Brown GD *et al.* Rapid immunoprecipitation mass spectrometry of endogenous proteins (RIME) for analysis of chromatin complexes. *Nat Protoc* 2016;11:316–26. https://doi.org/10.1038/nprot.2016.020
2. Silva J, Alkan F, Ramalho S *et al.* Ribosome impairment regulates intestinal stem cell identity via ZAKα activation. *Nat Commun* 2022;13:4492. https://doi.org/10.1038/s41467-022-32220-4
3. Gatto L, Lilley KS. MSnbase-an R/Bioconductor package for isobaric tagged mass spectrometry data visualization, processing and quantitation. *Bioinformatics* 2011;28:288–9. https://doi.org/10.1093/bioinformatics/btr645
4. Choi M, Chang C-Y, Clough T *et al.* MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. *Bioinformatics* 2014;30:2524–6. https://doi.org/10.1093/bioinformatics/btu305
5. Kohler D, Staniak M, Tsai TH *et al.* MSstats version 4.0: statistical analyses of quantitative mass spectrometry-based proteomic experiments with chromatography-based quantification at scale. *J Proteome Res* 2023;22:1466–82. https://doi.org/10.1021/acs.jproteome.2c00834

6. Zhang X, Smits AH, van Tilburg GBA *et al.* Proteome-wide identification of ubiquitin interactions using UbIA-MS. *Nat Protoc* 2018;13:530–50. https://doi.org/10.1038/nprot.2017.147
7. Feng Z, Fang P, Zheng H *et al.* DEP2: an upgraded comprehensive analysis toolkit for quantitative proteomics data. *Bioinformatics* 2023;39:btad526. https://doi.org/10.1093/bioinformatics/btad526
8. Eickhoff N, Hoekman L, Bleijerveld O *et al.* DEprot: a comprehensive R-package for the analyses of label-free quantitation mass-spectrometry data. *Zenodo. https://doi.org/10.5281/zenodo.17472993. 14 February 2025.*
9. Gregoricchio S, Eickhoff N. sebastian-gregoricchio/DEprot: v0.1.0 (0.1.0). *Zenodo. https://doi.org/10.5281/zenodo.18233891. 13 January 2026.*
10. Cox J, Hein MY, Luber CA *et al.* Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol Cell Proteomics* 2014;13:2513–26. https://doi.org/10.1074/mcp.M113.031591
11. Demichev V, Messner CB, Vernardis SI *et al.* DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nat Methods* 2020;17:41–4. https://doi.org/10.1038/s41592-019-0638-x
12. Tyanova S, Temu T, Sinitcyn P *et al.* The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat Methods* 2016;13:731–40. https://doi.org/10.1038/nmeth.3901
13. Brombacher E, Schad A, Kreutz C. Tail-robust quantile normalization. *Proteomics* 2020;20:2000068. https://doi.org/10.1002/pmic.202000068
14. Voß H, Schlumbohm S, Barwikowski P *et al.* HarmonizR enables data harmonization across independent proteomic datasets with appropriate handling of missing values. *Nat Commun* 2022;13:3523.
15. Chilimoniuk J, Grzesiak K, Kała J *et al.* imputomics: web server and R package for missing values imputation in metabolomics data. *Bioinformatics* 2024;40:btae098. https://doi.org/10.1093/bioinformatics/btae098
16 Ma W, Kim S, Chowdhury S *et al.* DreamAI: algorithm for the imputation of proteomics data. bioRxiv, https://doi.org/10.1101/2020.07.21.214205, 19 May 2021, preprint: not peer reviewed
17. Stacklies W, Redestig H, Scholz M *et al.* pcaMethods—a bioconductor package providing PCA methods for incomplete data. *Bioinformatics* 2007;23:1164–7. https://doi.org/10.1093/bioinformatics/btm069
18. Ritchie ME, Phipson B, Wu D *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 2015;43:e47. https://doi.org/10.1093/nar/gkv007
19. Wolski WE, Nanni P, Grossmann J *et al.* prolfqua: a comprehensive R-package for proteomics differential expression analysis. *J Proteome Res* 2023;22:1092–104. https://doi.org/10.1021/acs.jproteome.2c00441
20. Breitkreutz A, Choi H, Sharom JR *et al.* A global protein kinase and phosphatase interaction network in yeast. *Science* 2010;328:1043–6. https://doi.org/10.1126/science.1176495
21. Teo G, Liu G, Zhang J *et al.* SAINTexpress: improvements and additional features in Significance Analysis of INTeractome software. *J Proteomics* 2014;100:37–43. https://doi.org/10.1016/j.jprot.2013.10.023
22. Mellacheruvu D, Wright Z, Couzens AL *et al.* The CRAPome: a contaminant repository for affinity purification-mass spectrometry data. *Nat Methods* 2013;10:730–6. https://doi.org/10.1038/nmeth.2557
23. Steinkamp R, Tsitsiridis G, Brauner B *et al.* CORUM in 2024: protein complexes as drug targets. *Nucleic Acids Res* 2024;53:D651–7. https://doi.org/10.1093/nar/gkae1033
24. Wu T, Hu E, Xu S *et al.* clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *The Innovation* 2021;2:100141. https://doi.org/10.1016/j.xinn.2021.100141
25. Kerseviciute I, Gordevicius J. aPEAR: an R package for autonomous visualization of pathway enrichment networks. *Bioinformatics* 2023;39:btad672. https://doi.org/10.1093/bioinformatics/btad672
26. Liberzon A, Birger C, Thorvaldsdóttir H *et al.* The Molecular Signatures Database hallmark gene set collection. *Cell Syst* 2015;1:417–25. https://doi.org/10.1016/j.cels.2015.12.004
27. Gregoricchio S, Polit L, Esposito M *et al.* HDAC1 and PRC2 mediate combinatorial control in SPI1/PU.1-dependent gene repression in murine erythroleukaemia. *Nucleic Acids Res* 2022;50:7938–58. https://doi.org/10.1093/nar/gkac613
28. Eickhoff N, Janetzko J, Padrão N *et al.* TRIM33 loss reduces androgen receptor transcriptional output and H2BK120 ubiquitination. *Commun Biol* 2025;8:1043. https://doi.org/10.1038/s42003-025-08449-2
29. Aslam B, Basit M, Nisar MA *et al.* Proteomics: technologies and their applications. *J Chromatogr Sci* 2017;55:182–96. https://doi.org/10.1093/chromsci/bmw167
30. Kong W, Hui HWH, Peng H *et al.* Dealing with missing values in proteomics data. *Proteomics* 2022;22:2200092. https://doi.org/10.1002/pmic.202200092
31. Schumann Y, Gocke A, Neumann JE. Computational methods for data integration and imputation of missing values in omics datasets. *Proteomics* 2025;25:e202400100. https://doi.org/10.1002/pmic.202400100
32. Jin L, Bi Y, Hu C *et al.* A comparative study of evaluating missing value imputation methods in label-free proteomics. *Sci Rep* 2021;11:1760. https://doi.org/10.1038/s41598-021-81279-4
33. Shen M, Chang YT, Wu CT *et al.* Comparative assessment and novel strategy on methods for imputing proteomics data. *Sci Rep* 2022;12:1067. https://doi.org/10.1038/s41598-022-04938-0
34. Stekhoven DJ, Bühlmann P. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics* 2011;28:112–8. https://doi.org/10.1093/bioinformatics/btr597
35. Fix E, Hodges JL. Discriminatory analysis. Nonparametric discrimination: consistency properties. *International Statistical Review* 1989;57:238–47.
36. Kim H, Golub GH, Park H. Missing value estimation for DNA microarray gene expression data: local least squares imputation. *Bioinformatics* 2004;21:187–98. https://doi.org/10.1093/bioinformatics/bth499
37. Davis TJ, Firzli TR, Higgins Keppler EA *et al.* Addressing missing data in GC × GC metabolomics: identifying missingness type and evaluating the impact of imputation methods on experimental replication. *Anal Chem* 2022;94:10912–20. https://doi.org/10.1021/acs.analchem.1c04093
38. Stelloo S, Nevedomskaya E, Kim Y *et al.* Endogenous androgen receptor proteomic profiling reveals genomic subcomplex involved in prostate tumorigenesis. *Oncogene* 2018;37:313–22. https://doi.org/10.1038/onc.2017.330
39. Goeminne LJE, Sticker A, Martens L *et al.* MSqRob takes the missing hurdle: uniting intensity- and count-based proteomics. *Anal Chem* 2020;92:6278–87. https://doi.org/10.1021/acs.analchem.9b04375
40. Deutsch EW, Csordas A, Sun Z *et al.* The ProteomeXchange consortium in 2017: supporting the cultural change in proteomics public data deposition. *Nucleic Acids Res* 2017;45:D1100–6. https://doi.org/10.1093/nar/gkw936d1106.
41. Perez-Riverol Y, Bai J, Bandla C *et al.* The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Res* 2022;50:D543–52. https://doi.org/10.1093/nar/gkab1038

Received: November 5, 2025. **Revised:** January 13, 2026. **Accepted:** January 22, 2026