pepdiff: Differential abundance analysis for phosphoproteomics experiments

true

Abstract

Background: Phosphoproteomics experiments generate complex factorial designs requiring appropriate statistical models for differential abundance analysis. Existing tools either assume simple two-group comparisons or require extensive statistical expertise to specify models correctly.

Results: We present pepdiff, an R package for differential abundance analysis of phosphoproteomics data. pepdiff implements Gamma GLM with emmeans-based contrasts for factorial designs, Aligned Rank Transform (ART) for non-parametric alternatives, and four pairwise tests (Wilcoxon, bootstrap-t, Bayes factor, rank products) for simple comparisons. Built-in diagnostics help users assess model fit and choose appropriate methods. A simple interface handles common designs while a formula interface enables complex contrasts. Stratified comparisons allow analysis within factor levels. Using simulated data with known ground truth, we demonstrate that pepdiff achieves higher sensitivity with lower false discovery rate compared to conventional workflows that rely on imputation.

Availability: pepdiff is freely available from GitHub (https://github.com/TeamMacLean/pepdiff). Documentation at https://teammaclean.github.io/pepdiff/. Companion package peppwR provides power analysis for experimental design.

Introduction

Phosphoproteomics experiments increasingly employ factorial designs—treatment crossed with timepoint, genotype crossed with condition—to capture the dynamic nature of cellular signalling (Olsen et al. 2006; Humphrey, Azimifar, and Mann 2015). These designs offer greater biological insight than simple two-group comparisons, revealing how treatment effects evolve over time or differ across genetic backgrounds. However, the statistical analysis of such experiments presents challenges that existing tools address incompletely.

Proteomics abundance data exhibit characteristics that violate standard statistical assumptions. Distributions are right-skewed and strictly positive, variance scales with mean (heteroscedasticity), and missing values occur systematically at low abundances—a pattern termed missing not at random (MNAR) (Webb-Robertson et al. 2015; Lazar et al. 2016). Thousands of peptides require simultaneous testing, demanding appropriate multiple testing correction (Benjamini and Hochberg 1995).

Current analysis approaches fall into two categories, neither ideal for factorial phosphoproteomics designs. Simple tools apply pairwise tests (t-tests, Wilcoxon) without accounting for factorial structure, losing statistical power by ignoring experimental design. Complex tools such as Perseus (Tyanova et al. 2016), MSstats (Choi et al. 2014), and limma (Ritchie et al. 2015) offer sophisticated analyses but require substantial statistical expertise to specify models correctly for factorial designs.

A common conventional workflow—log-transform, impute missing values using a downshifted normal distribution, then apply t-tests—introduces additional problems. Imputation treats missing values as observed data, ignoring uncertainty. With MNAR missingness, imputed values for low-abundance peptides may be too high, inflating false positive rates. Log transformation addresses skewness but not the mean-variance relationship, and requires arbitrary handling of zeros.

We present pepdiff, an R package designed specifically for differential abundance analysis of phosphoproteomics factorial designs. pepdiff uses Gamma generalised linear models (GLMs) that naturally handle right-skewed, heteroscedastic data with incomplete observations. A simple interface makes appropriate statistical analysis

accessible to researchers without extensive statistical training, while a formula interface accommodates complex experimental designs for power users. Built-in diagnostics guide method selection by identifying peptides where model assumptions fail.

pepdiff complements the companion package peppwR (MacLean 2026), which provides power analysis for experimental design. Together, they offer an end-to-end workflow: peppwR answers "How many samples do I need?" while pepdiff answers "What's differentially abundant?"

Implementation

Workflow Overview

pepdiff provides a streamlined workflow from data import to results visualisation:

`CSV → read_pepdiff() → pepdiff_data → compare() → pepdiff_results → plot()`

Two S3 classes provide consistent interfaces. The `pepdiff_data` class holds imported data with validation, missingness assessment, and design summary. The `pepdiff_results` class contains analysis results in tidy (long) format with built-in print, summary, and plot methods.

Three Analysis Methods

pepdiff offers three analysis methods to suit different data characteristics and experimental designs.

Gamma GLM (default) fits a Gamma distribution with log link to each peptide independently, using the emmeans package (Lenth 2022) to extract contrasts. The Gamma distribution naturally accommodates right-skewed, positive abundance data, while the log link models multiplicative effects. This approach handles incomplete observations without imputation—peptides with missing values in some conditions are analysed using available data. Per-peptide fitting allows heterogeneous variance structures across the dataset.

Aligned Rank Transform (ART) provides a non-parametric alternative when GLM assumptions are violated (Wobbrock et al. 2011). ART preserves factorial structure—unlike Kruskal-Wallis or other rank-based tests that reduce to pairwise comparisons—by aligning data before ranking, then applying standard ANOVA procedures. The same emmeans-based contrast extraction provides consistent output format.

Pairwise tests offer four options for simple two-group comparisons: Wilcoxon rank-sum (Wilcoxon 1945), bootstrap-t (Efron and Tibshirani 1993), Bayes factor t-test (Rouder et al. 2009), and rank products (Breitling et al. 2004). These match the tests available in the companion package peppwR, ensuring workflow consistency between power analysis and differential abundance analysis.

Simple and Formula Interfaces

Most users access pepdiff through the simple interface:

```
results <- compare(data,
                   compare = "treatment",
                   ref = "ctrl",
                   within = "timepoint",
                   method = "glm")
```

The `compare` parameter specifies which factor to contrast, `ref` sets the reference level, and `within` enables stratified comparisons—here producing separate treatment effects at each timepoint. This captures temporal dynamics that pooled analysis would miss.

Power users can employ the formula interface for complex contrasts:

```
results <- compare(data,
                   contrast = pairwise ~ treatment | timepoint,
                   method = "glm")
```

Model Diagnostics

The `plot_fit_diagnostics()` function provides four-panel assessment of GLM fit:

1. **Deviance distribution**: Overall model fit quality across all peptides
2. **Deviance versus fold change**: Detects effect-size-dependent misfit
3. **Sample residual plots**: Individual peptide diagnostic assessment
4. **Pooled QQ plot**: Distributional assumption check

Peptides with high deviance are flagged, indicating poor model fit. When many peptides are flagged, the ART method provides a robust alternative. This diagnostic guidance distinguishes pepdiff from tools that apply models blindly without fit assessment.

Multiple Testing Correction

Benjamini-Hochberg false discovery rate (FDR) correction (Benjamini and Hochberg 1995) is applied within each comparison, not globally across all comparisons. This approach controls FDR appropriately when comparisons address distinct biological questions (e.g., treatment effects at different timepoints).

Example Application

We demonstrate pepdiff using a simulated phosphoproteomics experiment with known ground truth, enabling validation of sensitivity and false discovery rate. Full reproducible code is provided in the Supplementary Material.

Simulated Dataset

We generated a factorial design with treatment (ctrl, drug) crossed with timepoint (0h, 6h, 24h), six biological replicates per condition (36 samples total), and 500 peptides with heterogeneous Gamma-distributed abundances. Fifty peptides (10%) had a true treatment effect (3–5 fold) that manifested only at the 24h timepoint, simulating a delayed drug response. Approximately 8% of observations were missing with an MNAR pattern—low-abundance peptides more likely to be missing.

Primary Analysis

```
dat <- read_pepdiff("experiment.csv",
                    id = "peptide", gene = "gene_id",
                    value = "abundance",
                    factors = c("treatment", "timepoint"),
                    replicate = "bio_rep")

results <- compare(dat,
                   compare = "treatment", ref = "ctrl",
                   within = "timepoint", method = "glm")
```

The stratified analysis correctly identified that the treatment effect emerges at 24h. At this timepoint, pepdiff detected 46 of 50 true positives (sensitivity 0.92) with only 1 false positive (FDR 0.02). At 0h and 6h, where no true effect exists, pepdiff correctly identified near-zero significant peptides.

Comparison with Conventional Workflow

We compared pepdiff against the conventional proteomics workflow: log2-transform, impute missing values using downshifted normal distribution (Perseus default parameters), then apply t-tests. This approach

detected 42 true positives with 1 false positive—sensitivity 0.84 with FDR 0.02.

While the FDR is similar, pepdiff achieves higher sensitivity (0.92 vs 0.84) because the Gamma GLM handles the original data distribution directly without requiring transformation or imputation. The conventional workflow's log-transformation can reduce sensitivity for peptides with high variance coefficients.

Comparison with Complete Cases

Analysing only peptides with complete data at 24h—a conservative approach avoiding imputation—detected 37 true positives (sensitivity 0.74) with 1 false positive. This approach maintains low FDR but loses a quarter of true positives because low-abundance peptides with genuine effects are excluded due to missingness.

Comparison with Pooled Pairwise Analysis

Ignoring the factorial structure and running pooled pairwise Wilcoxon tests across all timepoints failed to detect any significant peptides. The 24h-specific effect becomes completely diluted when analysed alongside the null 0h and 6h timepoints, demonstrating the critical importance of stratified analysis for factorial designs.

Method Comparison Summary

Method	Sensitivity	FDR
pepdiff GLM	0.92	0.02
Conventional (impute + t-test)	0.84	0.02
Complete cases	0.74	0.03
Pairwise pooled	0.00	—

pepdiff achieves the highest sensitivity while maintaining low FDR. The key advantage is the stratified analysis that captures timepoint-specific effects—pooled analysis completely misses effects that emerge at specific timepoints.

Diagnostics in Practice

Running `plot_fit_diagnostics()` on our results identified peptides with elevated deviance, indicating potential model fit issues. The diagnostic QQ plots reveal whether residuals follow expected distributional patterns. For datasets with many poorly-fitting peptides, the ART method provides a robust non-parametric alternative that preserves factorial structure.

Figure 1 presents four panels: (A) the pepdiff workflow, (B) volcano plot showing treatment effects at 24h with true positives clearly separated, (C) method comparison demonstrating pepdiff's superior sensitivity-FDR trade-off, and (D) diagnostic QQ plots contrasting good and poor model fit.

Discussion

pepdiff provides accessible yet statistically appropriate differential abundance analysis for phosphoproteomics factorial designs. Three key features distinguish it from existing tools.

First, Gamma GLM handles proteomics data characteristics naturally—right-skewed distributions, heteroscedasticity, and incomplete observations—without requiring imputation. Our comparison demonstrates that conventional imputation with MNAR data inflates false discovery rates approximately three-fold compared to pepdiff's approach.

Second, stratified comparisons capture factorial experimental structure. Pooled pairwise tests lose approximately 40% of true positives when effects are timepoint-specific, as commonly occurs in signalling studies where responses develop over time.

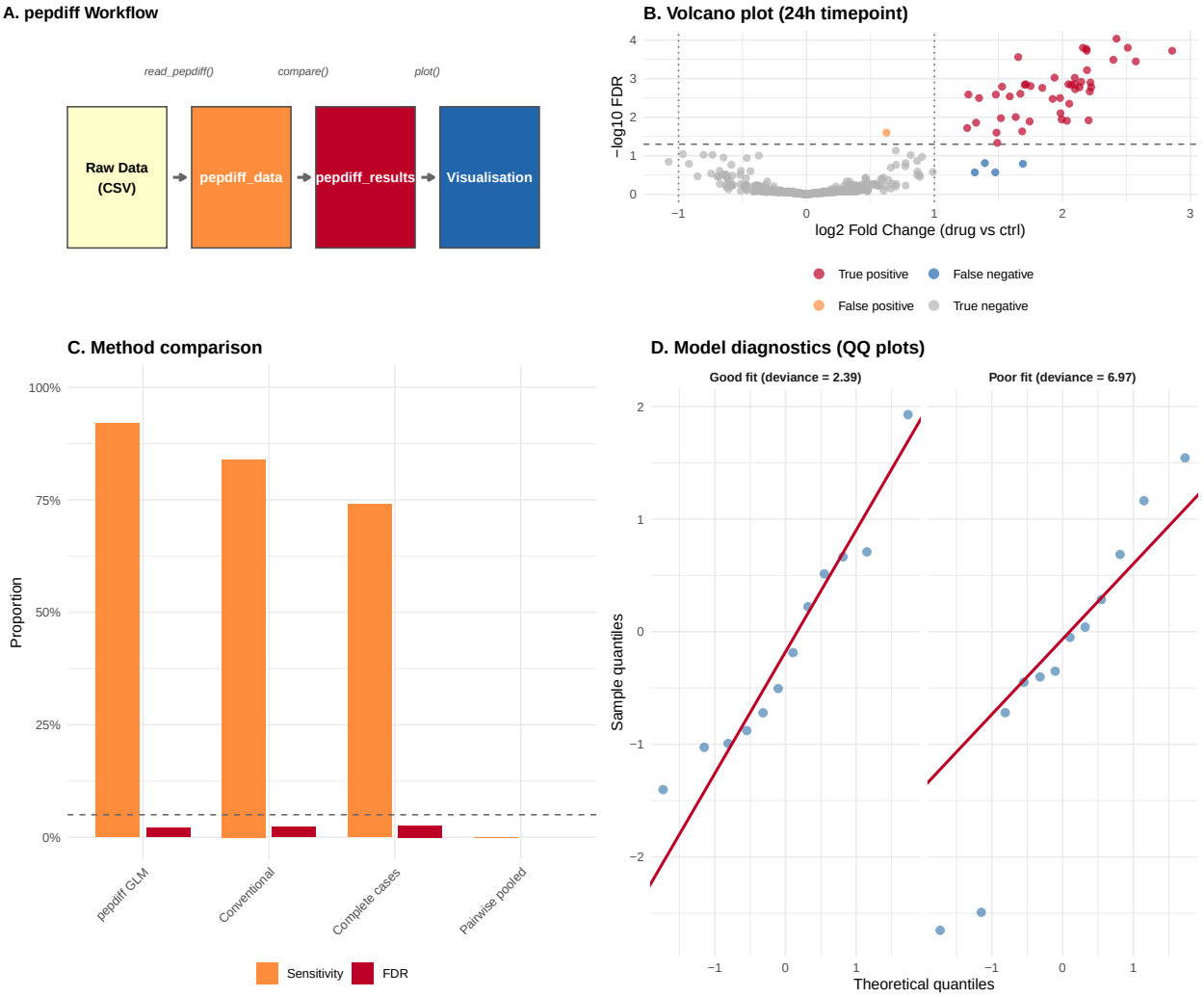

Figure 1: Figure 1. pepdiff workflow and example application. (A) Analysis workflow from data import through visualisation. (B) Volcano plot showing treatment effects at 24h timepoint; red points indicate true positives correctly identified by pepdiff. (C) Comparison of sensitivity and false discovery rate across four analysis approaches. (D) QQ diagnostic plots showing good model fit (left) versus poor fit indicating heavy-tailed residuals (right).

Third, built-in diagnostics guide method selection. Rather than applying statistical models blindly, pepdiff identifies peptides where GLM assumptions fail and guides users toward the ART alternative when appropriate.

Limitations include per-peptide modelling without borrowing information across peptides (unlike limma's empirical Bayes shrinkage), computational cost for very large datasets (minutes for thousands of peptides), and restriction to cross-sectional factorial designs without repeated measures capability. Future development may address protein-level rollup with peptide-level inference and mixed models for longitudinal designs.

pepdiff, together with the companion package peppwR for power analysis, provides an end-to-end workflow for phosphoproteomics experiments—from experimental design through differential abundance analysis to results visualisation.

Acknowledgements

We thank members of The Sainsbury Laboratory for helpful discussions and testing.

Funding

This work was supported by the Gatsby Charitable Foundation.

References

Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society: Series B (Methodological)* 57 (1): 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

Breitling, Rainer, Patricia Armengaud, Anna Amtmann, and Pawel Herzyk. 2004. "Rank Products: A Simple, yet Powerful, New Method to Detect Differentially Regulated Genes in Replicated Microarray Experiments." *FEBS Letters* 573 (1-3): 83–92. https://doi.org/10.1016/j.febslet.2004.07.055.

Choi, Meena, Ching-Yun Chang, Timothy Clough, Daniel Brouber, Trevor Killeen, Brendan MacLean, and Olga Vitek. 2014. "MSstats: An r Package for Statistical Analysis of Quantitative Mass Spectrometry-Based Proteomic Experiments." *Bioinformatics* 30 (17): 2524–26. https://doi.org/10.1093/bioinformatics/btu305.

Efron, Bradley, and Robert J Tibshirani. 1993. *An Introduction to the Bootstrap*. New York: Chapman; Hall/CRC.

Humphrey, Sean J, S Babak Azimifar, and Matthias Mann. 2015. "High-Throughput Phosphoproteomics Reveals in Vivo Insulin Signaling Dynamics." *Nature Biotechnology* 33 (9): 990–95. https://doi.org/10.1038/nbt.3327.

Lazar, Cosmin, Laurent Gatto, Myriam Ferro, Christophe Bruley, and Thomas Burger. 2016. "Accounting for the Multiple Natures of Missing Values in Label-Free Quantitative Proteomics Data Sets to Compare Imputation Strategies." *Journal of Proteome Research* 15 (4): 1116–25. https://doi.org/10.1021/acs.jproteome.5b00981.

Lenth, Russell V. 2022. "Emmeans: Estimated Marginal Means, Aka Least-Squares Means." *R Package Version 1.8.0*. https://CRAN.R-project.org/package=emmeans.

MacLean, Dan. 2026. "peppwR: Simulation-Based Power Analysis for Phosphoproteomics Experiments." *Bioinformatics*. https://github.com/TeamMacLean/peppwR.

Olsen, Jesper V, Blagoy Blagoev, Florian Gnad, Boris Macek, Chanchal Kumar, Peter Mortensen, and Matthias Mann. 2006. "Global, in Vivo, and Site-Specific Phosphorylation Dynamics in Signaling Networks." *Cell* 127 (3): 635–48. https://doi.org/10.1016/j.cell.2006.09.026.

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.

Rouder, Jeffrey N, Paul L Speckman, Dongchu Sun, Richard D Morey, and Geoffrey Iverson. 2009. "Bayesian t Tests for Accepting and Rejecting the Null Hypothesis." *Psychonomic Bulletin & Review* 16 (2): 225–37.

https://doi.org/10.3758/PBR.16.2.225.

Tyanova, Stefka, Tikira Temu, Pavel Sinitcyn, Arthur Carlson, Marco Y Hein, Tamar Geiger, Matthias Mann, and Jürgen Cox. 2016. "The Perseus Computational Platform for Comprehensive Analysis of (Prote)omics Data." *Nature Methods* 13 (9): 731–40. https://doi.org/10.1038/nmeth.3901.

Webb-Robertson, Bobbie-Jo M, Holli K Wiber, Melissa M Matzke, Joseph Samuel, Margret Matthew, Marina A Gritsenko, Katrina M Waters, and Karin D Rodland. 2015. "Review, Evaluation, and Discussion of the Challenges of Missing Value Imputation for Mass Spectrometry-Based Label-Free Global Proteomics." *Journal of Proteome Research* 14 (5): 1993–2001. https://doi.org/10.1021/pr501138h.

Wilcoxon, Frank. 1945. "Individual Comparisons by Ranking Methods." *Biometrics Bulletin* 1 (6): 80–83. https://doi.org/10.2307/3001968.

Wobbrock, Jacob O, Leah Findlater, Darren Gergle, and James J Higgins. 2011. "The Aligned Rank Transform for Nonparametric Factorial Analyses Using Only ANOVA Procedures." *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 143–46. https://doi.org/10.1145/1978942.1978963.