peppwR: Simulation-based power analysis for phosphoproteomics experiments

Dan MacLean, The Sainsbury Laboratory, Norwich Research Park, Norwich NR4 7UH, UK

Abstract

Background: Phosphoproteomics experiments require careful experimental design to ensure adequate statistical power, yet no dedicated tools exist for power analysis that account for the unique statistical properties of peptide-level quantification data.

Results: We present peppwR, an R package for simulation-based power analysis tailored to phosphoproteomics. peppwR fits distributions to pilot data on a per-peptide basis, capturing the heterogeneity in abundance and variance across the peptidome. It supports multiple statistical tests (Wilcoxon, bootstrap-t, Bayes factor), tracks missingness patterns common in mass spectrometry data, and provides FDR-aware power calculations. Users can determine required sample sizes, estimate power for planned experiments, or identify minimum detectable effect sizes.

Availability: peppwR is freely available from GitHub (https://github.com/TeamMacLean/peppwR) with documentation at https://teammaclean.github.io/peppwR/.

Introduction

Statistical power—the probability of detecting a true effect when one exists—is fundamental to experimental design. Under-powered experiments waste resources and risk missing genuine biological signals, while over-powered experiments are inefficient (Cohen 1988; Button et al. 2013). Power analysis is standard practice in clinical trials and genomics, yet remains neglected in proteomics despite the field's increasing quantitative sophistication.

Generic power analysis tools fail to capture the unique statistical properties of mass spectrometry-based proteomics data. First, peptides exhibit substantial heterogeneity: abundance levels span orders of magnitude, and variance differs dramatically across the peptidome (Olsen et al. 2006; Humphrey, Azimifar, and Mann 2015). Second, abundance distributions are typically right-skewed rather than normal, better described by gamma or lognormal distributions. Third, missing values follow non-random patterns—low-abundance peptides are systematically more likely to be undetected, creating missing-not-at-random (MNAR) data (Webb-Robertson et al. 2015; Lazar et al. 2016). Fourth, phosphoproteomics experiments simultaneously test thousands of peptides, requiring consideration of multiple testing correction.

Existing tools address these challenges incompletely. Perseus (Tyanova et al. 2016), the most widely-used proteomics analysis platform, provides comprehensive statistical analysis but lacks power analysis functionality. The clippda package (Nyangoma et al. 2012) was designed for SELDI-TOF peak data and assumes homogeneous variance across features. MultiPower (Tarazona et al. 2020) focuses on multi-omics integration, requiring multiple data types. G*Power (Faul et al. 2007) is a general-purpose tool that assumes normality and lacks proteomics-specific features.

We present peppwR, an R package that addresses these limitations through per-peptide distribution fitting, simulation-based power estimation, and missingness-aware calculations. peppwR answers three key experimental design questions: “What sample size do I need?”, “What power will I achieve?”, and “What is the minimum detectable effect size?”

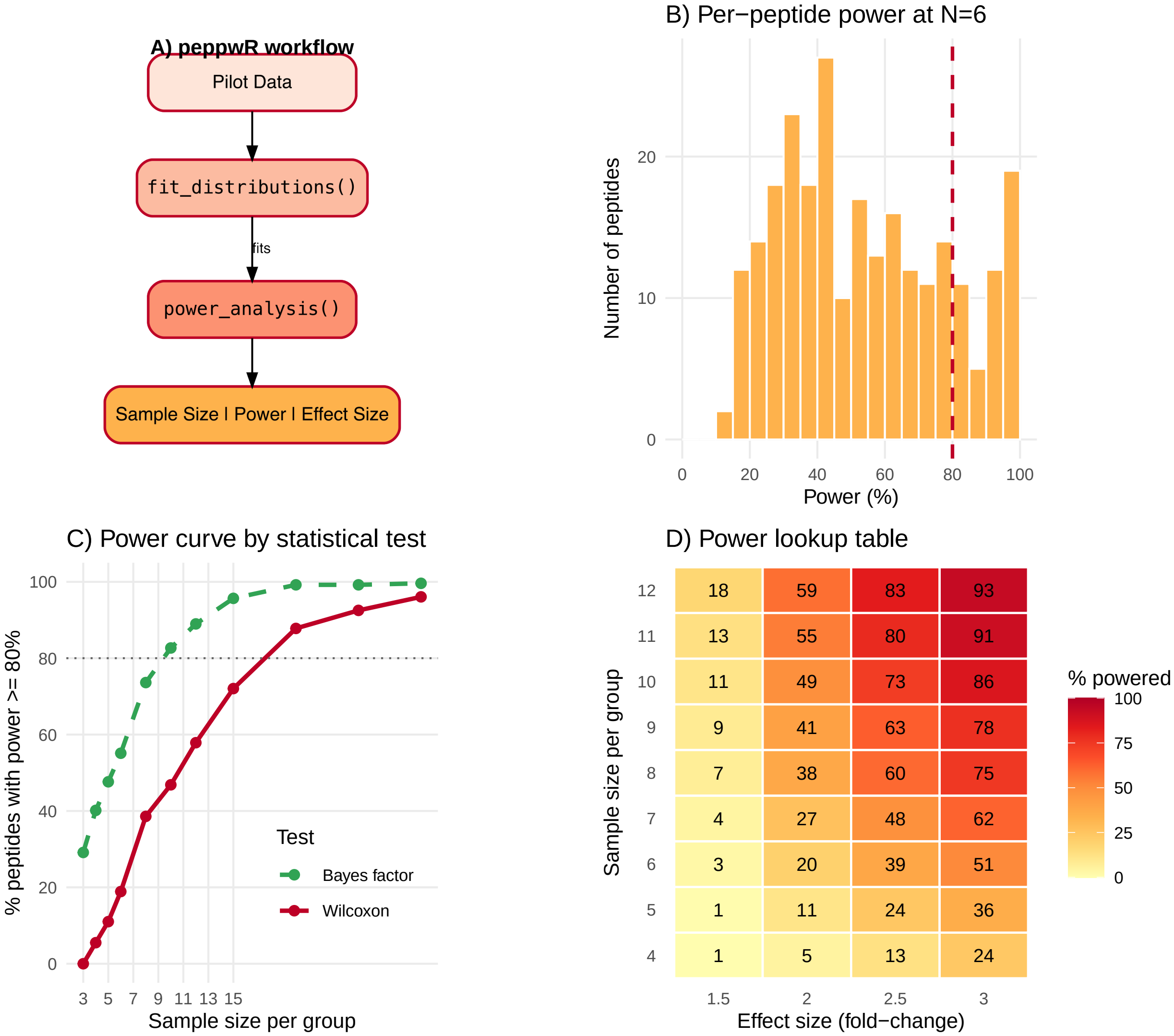

Figure 1: **Figure 1. peppwR workflow and example results.** (A) Two-stage workflow: fit distributions to pilot data, then run power simulations. (B) Distribution of per-peptide power at N=6 samples per group for detecting 2-fold changes. The dashed line indicates 80% power threshold. Power varies dramatically across peptides due to differences in abundance, variance, and missingness. (C) Power curves comparing Wilcoxon and Bayes factor tests. Y-axis shows the proportion of peptides that individually achieve 80% power at each sample size. Bayes factor reaches the 80% threshold (dotted line) at N=10 versus N=20 for Wilcoxon. (D) Power lookup table across sample sizes and effect sizes. Values indicate the percentage of peptides achieving 80% power for each combination.

Implementation

Workflow

peppwR implements a two-stage workflow (Figure 1A). First, `fit_distributions()` models the abundance distribution for each peptide in pilot data, selecting the best-fitting distribution from gamma, lognormal, normal, and inverse Gaussian candidates using AIC. Second, `power_analysis()` runs Monte Carlo simulations drawing from these fitted distributions to estimate power.

Two operational modes accommodate different use cases. In **aggregate mode**, users specify a single distribution and parameters, enabling power calculations without pilot data. In **per-peptide mode**, peppwR leverages pilot data to model heterogeneity across the peptidome, reporting the proportion of peptides achieving target power at each sample size.

Distribution fitting

For each peptide, peppwR fits candidate distributions using maximum likelihood estimation via the fitdistrplus (Delignette-Muller and Dutang 2015) and univariateML packages. The best-fitting distribution is selected by minimum AIC. When fitting fails (e.g., insufficient non-missing observations), users can choose to exclude the peptide, fall back to empirical bootstrap resampling, or use a moment-matched lognormal distribution.

Simulation engine

Power is estimated via Monte Carlo simulation. For each peptide and each candidate sample size:

1. Draw n samples from the fitted control distribution
2. Draw n samples from a treatment distribution scaled by the specified effect size
3. Apply the selected statistical test
4. Record whether the null hypothesis is rejected
5. Repeat for n_{sim} iterations

Power is the proportion of iterations achieving statistical significance. In per-peptide mode, peppwR reports the proportion of peptides exceeding target power at each sample size.

Statistical tests

peppwR supports three statistical tests. The **Wilcoxon rank-sum test** (Wilcoxon 1945) (default) is non-parametric and robust to distributional assumptions. The **bootstrap-t test** (Efron and Tibshirani 1994) handles non-normality through resampling. The **Bayes factor t-test** (Rouder et al. 2009) provides a Bayesian approach that quantifies evidence strength rather than dichotomous rejection, offering improved efficiency when parametric assumptions hold.

Missing data handling

peppwR tracks missingness at two levels. Per-peptide, it records the proportion of missing values. At the dataset level, it detects MNAR (missing-not-at-random) patterns by correlating mean abundance with missingness rate across peptides—a negative correlation indicates that low-abundance peptides are systematically more likely to be missing, the hallmark of detection-limit-driven MNAR common in mass spectrometry (Webb-Robertson et al. 2015). When `include_missingness = TRUE`, simulations incorporate observed per-peptide NA rates, providing power estimates that account for expected data loss.

FDR-aware power

For whole-peptidome studies, peppwR can simulate complete experiments with Benjamini-Hochberg FDR correction (Benjamini and Hochberg 1995). Setting `apply_fdr = TRUE` runs simulations across all peptides simultaneously, reporting the proportion of true positives discovered at the target FDR threshold.

Example Application

We demonstrate peppwR using simulated phosphoproteomics data mimicking typical pilot study characteristics: 500 peptides with heterogeneous gamma-distributed abundances and ~13% missingness with a realistic MNAR pattern (see Supplementary Material for complete code).

Per-peptide power analysis

Using `fit_distributions()` followed by `power_analysis()` with `find = "sample_size"`, we determined sample size requirements for detecting 2-fold changes using the Wilcoxon test. Because each peptide has its own distribution parameters, power is calculated separately for each peptide via simulation. We then ask: at what sample size do 80% of peptides individually achieve 80% power? This threshold-based approach acknowledges that not all peptides will be equally detectable. The answer: N = 20 samples per group.

Power heterogeneity

Figure 1B illustrates the central insight motivating per-peptide analysis: at N = 6 samples per group, power varies dramatically across peptides, ranging from near-zero to over 90%. This heterogeneity arises from differences in abundance, variance, and missingness. A single "average" power estimate would obscure this crucial variability and potentially mislead experimental design decisions.

Effect of statistical test choice

Comparing Wilcoxon and Bayes factor tests reveals substantial differences in sample size requirements (Figure 1C). The Bayes factor t-test achieves the 80% threshold at N = 10, compared to N = 20 for Wilcoxon—a 50% reduction in required samples. This efficiency gain reflects the Bayes factor's parametric assumptions, which are reasonable when the underlying distributions are well-characterized by pilot data. Users should match their power analysis to their planned statistical approach.

Power lookup table

Figure 1D presents a practical lookup table showing the proportion of peptides achieving 80% power across combinations of sample size and effect size. At N = 12 with 3-fold changes, 94% of peptides achieve adequate power; at N = 4 with 1.5-fold changes, only 1% do. Such tables enable researchers to make informed trade-offs between detectable effect sizes and experimental costs.

Discussion

peppwR provides the first dedicated power analysis tool for phosphoproteomics that accounts for per-peptide heterogeneity, non-normal distributions, and missingness patterns. The simulation-based approach accommodates the complex statistical properties of mass spectrometry data that violate assumptions of traditional power calculations.

Our example analysis demonstrates that power varies substantially across the peptidome, and that analytical choices—particularly the statistical test—meaningfully impact sample size requirements. By providing per-peptide power estimates, peppwR enables researchers to set realistic expectations about which peptides will be detectable at their planned sample size.

Limitations. Computational cost scales linearly with peptide count and number of simulations; typical analyses complete within minutes on modern hardware. Per-peptide mode requires pilot data from the same experimental system; when unavailable, aggregate mode with reasonable distributional assumptions provides useful guidance.

Future directions. Integration with the pepdiff package for differential abundance analysis would provide an end-to-end workflow from experimental design through analysis. Extension to TMT and iTRAQ labeling schemes, which have different variance structures, represents a natural next step.

Funding

This work was supported by the Gatsby Charitable Foundation.

Conflict of Interest

None declared.

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.

Button, Katherine S, John PA Ioannidis, Claire Mokrysz, Brian A Nosek, Jonathan Flint, Emma SJ Robinson, and Marcus R Munafò. 2013. "Power Failure: Why Small Sample Size Undermines the Reliability of Neuroscience." *Nature Reviews Neuroscience* 14 (5): 365–76. https://doi.org/10.1038/nrn3475.

Cohen, Jacob. 1988. "Statistical Power Analysis for the Behavioral Sciences." https://doi.org/10.4324/9780203771587.

Delignette-Muller, Marie Laure, and Christophe Dutang. 2015. "fitdistrplus: An R Package for Fitting Distributions." *Journal of Statistical Software* 64 (4): 1–34. https://doi.org/10.18637/jss.v064.i04.

Efron, Bradley, and Robert J. Tibshirani. 1994. *An Introduction to the Bootstrap.* 1st ed. Chapman; Hall/CRC. https://doi.org/10.1201/9780429246593.

Faul, Franz, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. "G*Power 3: A Flexible Statistical Power Analysis Program for the Social, Behavioral, and Biomedical Sciences." *Behavior Research Methods* 39 (2): 175–91. https://doi.org/10.3758/bf03193146.

Humphrey, Sean J, S Barzin 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.

Nyangoma, Stephen O, Stuart I Collins, Douglas G Altman, Philip Johnson, and Lucinda Billingham. 2012. "Sample Size Calculations for Designing Clinical Proteomic Profiling Studies Using Mass Spectrometry." *Statistical Applications in Genetics and Molecular Biology* 11 (3). https://doi.org/10.1515/1544-6115.1686.

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.

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.

Tarazona, Sonia, Leandro Balzano-Nogueira, David Gómez-Cabrero, Andreas Schmidt, Sara C Madeira, and Ana Conesa. 2020. "Harmonization of Quality Metrics and Power Calculation in Multi-Omic Studies." *Nature Communications* 11 (1): 3092. https://doi.org/10.1038/s41467-020-16937-8.

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, Jemima N Samuel, Karin D Rodland, Thomas RW Clauss, and Joel G Pounds. 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.