

Scoring the downstream chain of small-n label-free proteomics against known ratios: search engine, match-between-runs, imputation, differential testing and enrichment background on a public three-species benchmark

Byeongsoo Kang

shoo99@gmail.com

SYSOFT <https://orcid.org/0009-0007-2324-2351>

Research Article

Keywords: label-free quantification, benchmark, MaxQuant, FragPipe, DIA-NN, match-between-runs, missing value imputation, limma, DEqMS, enrichment analysis

Posted Date: October 15th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11370244/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: The authors declare potential competing interests as follows: The author is affiliated with SYSOFT and runs sbmlab.com, an advertising-supported website on which parts of these results were described. The author declares no other competing interests.

Abstract

Background.

Benchmarks of label-free proteomics usually compare search and quantification software. The result a biologist reads, however, is also shaped by later choices: whether identifications are transferred between runs, how missing values are handled, which test and threshold are applied, whether intensities are normalised again, and which background is used for enrichment. These choices are rarely scored against a known answer in the three-versus-three designs that are common in practice.

Methods.

I reanalysed a public hybrid-proteome benchmark (PRIDE PXD028735) in which human, yeast and *Escherichia coli* digests were mixed so that the true log₂ ratios between two conditions are 0, - 1 and + 2. Six Orbitrap DDA runs (three per condition) were processed with MaxQuant 2.8.1 and FragPipe 24, and six TripleTOF SWATH runs of the same samples with DIA-NN 2.3.2. Each downstream step was scored by the fraction of proteins whose ratio fell within ± 0.5 log₂ of the truth and by the number of unchanged human proteins called significant.

Results.

On the 3,866 proteins quantified by all three tools, the fraction within ± 0.5 for human / yeast / *E. coli* was 98.9 / 92.4 / 80.9% for MaxQuant LFQ, 99.7 / 99.3 / 97.7% for FragPipe and 100 / 99.4 / 97.7% for DIA-NN. For the 77 *E. coli* proteins MaxQuant quantified both with and without match-between-runs, 94.8% of ratios were within ± 0.5 with it and 62.3% without. With a strict valid-value filter and MBR, no imputation and five imputation methods found 84–92% of the true changes at 1.0–1.3% false discoveries; a loose filter raised the number of changes found but lowered the fraction found (for example 78% vs 92%) and roughly doubled the false discovery proportion, and k-nearest-neighbour imputation moved the median of a true + 2 change to + 0.02 in one setting. At BH-adjusted $p < 0.05$ without a fold-change threshold, t-test / limma / DEqMS called 68 / 104 / 284 unchanged human proteins on MaxQuant LFQ output (human median + 0.145), 15 / 43 / 37 on FragPipe (+ 0.014) and 106 / 167 / 188 on DIA-NN (- 0.010). A $|\log_2\text{FC}| \geq 0.5$ threshold reduced these to 2–22 for MaxQuant and DIA-NN but only to 7–35 for FragPipe, whose remaining false calls were large ratio errors in both directions. An additional per-sample median normalisation of DIA-NN output raised limma's false calls to 3,461 (59% of calls). Random lists of detected proteins returned a median of 103 GO and KEGG terms at FDR < 0.05 against STRING's whole-genome background.

Conclusions.

In this benchmark, single decisions after the search engine changed the number of false discoveries on the engines' output by up to 34-fold (adding a fold-change threshold: MaxQuant t-test, 68 to 2) or 25-fold (an extra median normalisation: DIA-NN t-test, 106 to 2,695). Pairing FDR with a fold-change threshold, using a strict valid-value filter, avoiding an extra median normalisation of FragPipe and DIA-NN output, and using the detected

proteins as enrichment background each removed much of the false signal. Scripts, settings and result tables are public (doi:10.5281/zenodo.23355028).

Background

Label-free quantitative proteomics has well-established benchmarks. Hybrid-proteome samples, in which digests of different species are mixed at known proportions, make it possible to score quantification against a known answer. Kuharev et al. used such samples to evaluate DIA software [1], Navarro et al. used them to compare software in a multicentre study [2], and Van Puyvelde et al. released a dataset of such samples measured with several acquisition strategies and instruments [3]. DIA software and workflows have since been benchmarked in detail [4, 5, 6], label-free workflows and imputation have been compared [7], and ProteoBench, a community platform, compares submitted workflows on subsets of the same benchmark dataset, including its Orbitrap DDA runs (at precursor and peptidoform level) [8].

Most of these studies end at the protein table. The conclusions a biologist draws, however, depend on a sequence of later steps: the intensity column (raw intensity, iBAQ or MaxLFQ [9]), whether identifications are transferred between runs (match-between-runs, MBR [10]), how missing values are handled [11], which statistical test and threshold are applied [12, 13, 14, 15], whether intensities are normalised again [16], and how enrichment is computed [17, 18, 19]. Each step has been studied on its own; their combined effect on the false discoveries of a small experiment is less often shown on one dataset with a known answer.

Small designs make this matter more. A comparison of three versus three is common, especially in core-facility and pilot work, and leaves little information for the variance estimates and normalisation that downstream methods rely on. Software also changes between versions in ways that can break analysis scripts without an error.

Here I scored the chain from search engine to enrichment on one public benchmark, using recent versions of three widely used tools, the same three-versus-three design and one scoring rule throughout. The aim is practical rather than a new finding: most individual effects shown here are known, and the contribution is to quantify them along one chain, on one dataset with one scoring rule, with reproducible scripts and a record of version-specific traps.

Methods

Data

PRIDE dataset PXD028735 [3] contains two hybrid samples: condition A with human, yeast and *E. coli* digests at 65 / 30 / 5% and condition B at 65 / 15 / 20%, so the true $\log_2(B/A)$ ratios are 0 (human), -1 (yeast) and +2 (*E. coli*). The two hybrid samples were prepared three consecutive times (labelled Alpha, Beta and Gamma) to include handling variability [3]. For each condition I used the three preparations, taking the first injection of each (injection 01), which gives three versus three. The design therefore contains preparation and technical variation but no biological variation.

MaxQuant and FragPipe were run on Orbitrap Q Exactive HF-X DDA data (120 min gradient). The dataset's Orbitrap data also include DIA acquisitions: all-ion fragmentation (AIF) with staggered 8 m/z windows and

narrow-window gas-phase-fractionated DIA runs for library building [3]. I did not analyse these, as the staggered windows require demultiplexing; for DIA-NN I used the SCIEX TripleTOF 6600 + SWATH runs of the same samples (99 variable windows). The DIA-NN results therefore differ from the DDA results in instrument and acquisition as well as software.

The protein database combined UniProtKB/Swiss-Prot reference proteomes for human (taxon 9606; 20,431 entries), yeast (559292; 6,733) and *E. coli* K-12 (83333; 4,531), downloaded on 6 October 2026.

Search and quantification

MaxQuant 2.8.1 [9, 20, 21] was run from the command line on Linux (.NET 8) with 16 threads: Trypsin/P, up to two missed cleavages, carbamidomethylation of cysteine as a fixed modification, oxidation of methionine and protein N-terminal acetylation as variable modifications, 1% PSM and protein FDR, LFQ with a minimum ratio count of 2, iBAQ, and MBR with a 0.4 min matching window and a 20 min alignment window, the values written by MaxQuantCmd --create in this version (MBR-specific FDR control not enabled). A second run with identical settings and MBR off was used for the MBR comparison. Protein groups flagged as decoy, potential contaminant (MaxQuant's built-in contaminant list) or only identified by site were removed.

FragPipe 24 (MSFragger 4.4.1 [22] and IonQuant 1.11.20 [23], installed separately and recorded in the FragPipe log) was run headless with the default LFQ-MBR workflow and 16 threads: strict trypsin, up to two missed cleavages, ± 20 ppm precursor and fragment tolerance, MSBooster [24] and Percolator [25] rescoring, Philosopher [26] filtering at 1% protein FDR (sequential, picked), and IonQuant MaxLFQ with FDR-controlled MBR and normalisation. The database was the same FASTA with Philosopher-added contaminants and reversed decoys. Thermo raw files were converted to mzML with ThermoRawFileParser 2.0.0 (GitHub release v2.0.0-dev) [27].

DIA-NN 2.3.2 [28] (Windows build, required for SCIEX .wiff files) was run with 24 threads in library-free mode. A spectral library was predicted from the same FASTA without contaminants (trypsin, one missed cleavage, peptide length 7–30, charge 2–4, precursor m/z 300–1800, fixed carbamidomethylation, up to one variable methionine oxidation), and the six runs were searched with MBR (--reanalyse), RT-dependent normalisation (--rt-profiling), 1% precursor q-value and automatic mass accuracy. The protein matrix (report.pg_matrix.tsv, MaxLFQ, DIA-NN's default matrix filtering) was used. The DIA-NN search space thus differed from the DDA searches (one instead of two missed cleavages, no N-terminal acetylation, no contaminant entries). For a no-MBR comparison, the first-pass quantification files were re-summarised with --use-quant; DIA-NN warns that this can differ from a fresh run when mass accuracy is set automatically, so this comparison is indicative only.

Contaminants were handled differently by the engines: MaxQuant flags them; in this FragPipe run, the 46 contaminant sequences Philosopher added to the database carried no contam_ prefix (for example sp|O77727|K1C15_SHEEP), so the protein table could not be filtered by prefix; and the DIA-NN database contained none. For the differential testing below, protein groups containing any accession from MaxQuant's contaminant list were removed from the FragPipe and DIA-NN tables (7 and 4 groups). In Fig. 1 they were not removed; removing them changed the scored counts by at most four human proteins and no percentage at one decimal place.

Species assignment and scoring

A protein group was assigned to a species only if all of its member proteins came from that species; mixed groups were excluded. For each protein, $\log_2(B/A)$ was the difference between the mean $\log_2$ intensities of the two conditions, requiring at least two of three values per condition. A ratio was scored as correct if it was within ± 0.5 of the truth; ± 0.3 and ± 1.0 were used as a sensitivity check. Engine comparisons (Fig. 1) and protein-level MBR comparisons (Fig. 2b, c) use uncentred ratios. Peptide-level MBR scores (Fig. 2a) and all imputation ratio metrics (medians and fractions within ± 0.5 , Fig. 3c) were centred on the median human ratio, which removes a global offset using only the unchanged background.

Match-between-runs

For MaxQuant, peptides.txt records for each run whether a value came from MS/MS or from matching. Peptides quantified in all six runs were grouped by how many of their six values came from matching (none, one or two, three or more) and scored with the same rule. To separate the effect of matching from that of abundance, peptides of each species were also binned into quartiles of mean $\log_2$ intensity and matched and fully MS/MS-based peptides were compared within each quartile. At protein level, MaxQuant LFQ ratios were compared between the MBR-on and MBR-off runs on the proteins scoreable in both.

Imputation

Starting from the MaxQuant LFQ table (with and without MBR), I applied two valid-value filters: strict (at least two of three values in both conditions) and loose (at least two of three in either condition). Missing values were left as they were ("none") or imputed with MinDet and MinProb (imputeLCMD 2.1; $q = 0.01$), a Perseus-style down-shifted normal distribution (width 0.3, shift 1.8 standard deviations, per sample) [29], QRILC (imputeLCMD) [11], or k-nearest neighbours (impute 1.76.0; $k = 10$, rowmax and colmax 0.9) [30], with a fixed random seed. Differential abundance was tested with limma [14] and called at BH-adjusted $p < 0.05$ [31] and $|\log_2FC| \geq 0.5$ (uncentred). Recall was the fraction of testable yeast and *E. coli* proteins called; the false discovery proportion was the fraction of calls that were human. With "none" under the loose filter, only proteins passing the strict filter can be tested.

Differential testing across engines

For each engine, proteins with all six values were tested with a Welch t-test, limma (moderated t) [14, 32] and DEqMS [15], and called at BH-adjusted $p < 0.05$, with and without $|\log_2FC| \geq 0.5$. DEqMS requires a count per protein: I used the minimum per-sample razor-plus-unique peptide count for MaxQuant (as in the DEqMS vignette), the minimum per-sample spectral count for FragPipe (no per-sample peptide count is reported) and the number of sequences for DIA-NN, each plus one. DEqMS results therefore differ between engines in their count input as well as in the quantities. True calls were counted regardless of sign; at most six true calls in any setting had the wrong sign. Each analysis was run on the engine output as is and after an additional per-sample median centring of $\log_2$ intensities, a frequently used normalisation step [16].

Enrichment background

The human proteins detected by DIA-NN were mapped to STRING v12.0 identifiers (5,686 identifiers). Twenty random lists of 300 detected proteins were each submitted to STRINGdb 2.14.3 [19] twice: with the default background (whole genome) and with the detected proteins as background. Terms at FDR < 0.05 in GO Biological Process, Cellular Component, Molecular Function and KEGG were counted; other categories returned by STRING were not counted. As a positive control, five lists were built from 15 detected proteins of the Reactome cholesterol biosynthesis gene set (msigdb 26.1.1) plus 285 random detected proteins.

Software environment and availability

MaxQuant, FragPipe and the R and Python analyses ran on one workstation (Intel Core i9-14900K, 128 GB RAM) under Linux in WSL2 with 64 GB available; DIA-NN ran natively on Windows on the same machine. Run times are therefore not directly comparable between DIA-NN (24 threads, native) and the DDA tools (16 threads, WSL2). R 4.3.3 (data.table 1.16.2, limma 3.58.1, DEqMS 1.20.0, imputeLCMD 2.1, impute 1.76.0, STRINGdb 2.14.3, msigdb 26.1.1) and Python 3.12.3 (pandas 3.0.6, matplotlib 3.11.2) were used for scoring and figures. Scripts, search parameter files, logs and result tables are available at <https://github.com/shoo99/sbmlab-benchmarks> and archived at Zenodo (doi:10.5281/zenodo.23355028). Raw data are available from PRIDE (PXD028735). An AI assistant (Claude Opus 5.5, Anthropic, used through Claude Code between 6 and 14 October 2026) was used under the author's direction to write and run analysis scripts, to draft the text and to check reported numbers against the result tables by independent recomputation; see Declarations.

Results

Search engines

MaxQuant LFQ scored 2,896 human, 907 yeast and 132 *E. coli* proteins; FragPipe MaxLFQ 4,605, 1,817 and 359; and DIA-NN 5,477, 2,374 and 446 (Fig. 1a). Over all scored proteins, the fraction within ± 0.5 of the truth was 98.8 / 92.3 / 81.1% (human / yeast / *E. coli*) for MaxQuant, 94.7 / 88.3 / 73.5% for FragPipe and 98.6 / 95.1 / 85.4% for DIA-NN (Fig. 1b). FragPipe's lower overall values come from the 2,905 proteins it scored that MaxQuant did not, which were less accurate (86.6 / 78.3 / 59.6%); on the 3,877 proteins scoreable in both, FragPipe was more accurate for every species (99.7 / 99.2 / 97.7% vs 98.9 / 92.2 / 80.9%). On the 3,866 proteins quantified by all three tools, the fractions were 98.9 / 92.4 / 80.9% (MaxQuant), 99.7 / 99.3 / 97.7% (FragPipe) and 100 / 99.4 / 97.7% (DIA-NN) (Fig. 1c), and DIA-NN's ratios were the tightest (interquartile range for human 0.044 vs 0.142 and 0.139). Because MaxQuant's ratios carried an offset, centring each engine's ratios on its own human median changed MaxQuant most: over all scored proteins its fractions became 99.3 / 96.9 / 82.6%, whereas FragPipe and DIA-NN changed by at most 0.5 points (Fig. 1b). On the 3,866 proteins quantified by all three tools, centring raised MaxQuant's yeast fraction from 92.4% to 97.1% (FragPipe 99.3%, DIA-NN 99.4%) but its *E. coli* fraction only from 80.9% to 82.4% (both others 97.7%): the offset explains part of the yeast gap and little of the *E. coli* gap. With a tolerance of ± 0.3 the differences between engines widened (MaxQuant 90.4 / 73.2 / 67.2%, FragPipe 97.7 / 94.7 / 91.6%, DIA-NN 99.8 / 98.4 / 94.7%); with ± 1.0 all engines exceeded 93%.

MaxQuant LFQ ratios were shifted: on the scored proteins the median human ratio was + 0.144 (+ 0.145 on the proteins with all six values used for testing), and the yeast ratio was compressed towards zero (median -

0.797). Raw MaxQuant intensity was less accurate (86 / 74 / 43% within ± 0.5) but scored more proteins (3,603 / 1,250 / 243); iBAQ gave between-sample ratios identical to raw intensity, as expected for a per-protein scaling factor.

Wall time for the six runs was 61.6 min for MaxQuant (52.0 min without MBR) and 8.9 min for FragPipe as reported by the pipeline (9.3 min wall clock), plus about 4 min for raw-to-mzML conversion; the DIA-NN search took 6 h 17 min plus 11 min for library prediction, of which reading the vendor .wiff files accounted for roughly 3.5 h (Fig. 1d).

Match-between-runs

In MaxQuant, 20,830 of 210,449 quantified peptide values (9.9%) came from matching. After centring, matched peptide values were noisier than values measured by MS/MS in all six runs, most clearly for low-abundance *E. coli* peptides: 88.1% of fully MS/MS-based *E. coli* peptides were within ± 0.5 , compared with 76.7% of those with one or two matched values and 61.5% of those with three or more (Fig. 2a). For human and yeast, peptides with one or two matched values were the least accurate (88.7% and 86.8%, compared with 95.0% and 93.3% for fully MS/MS-based peptides). Matched values were concentrated in low-intensity peptides (43–59% of peptides in the lowest intensity quartile of each species, 3–7% in the highest), but within the same intensity quartile, peptides with matched values were still less accurate in 11 of 12 species-by-quartile comparisons (for example, human lowest quartile 88.8% vs 94.6%); the exception was the highest *E. coli* quartile, with 8 matched peptides.

At protein level, in a separate comparison of the MBR-on and MBR-off runs, MBR improved the accuracy of proteins quantified in both runs: the fraction within ± 0.5 was 94.8% vs 62.3% for *E. coli* and 97.7% vs 95.0% for yeast (human 99.5% vs 99.1%) (Fig. 2b). MBR added 23% more scoreable human, 38% more yeast and 71% more *E. coli* proteins in MaxQuant, and 7%, 23% and 51% in DIA-NN (Fig. 2c). The 860 MaxQuant proteins scoreable only with MBR were less accurate (96 / 78 / 62% within ± 0.5) than those quantified either way.

Imputation

With MBR and the strict filter (3,935 proteins, 2.1% of values missing), every approach found between 868 and 959 of the 1,039 testable changed proteins (84–92%) at false discovery proportions of 1.0–1.3% (Fig. 3a). No imputation and kNN found the most. Without MBR (2.8% missing), the range was 586–701 of 733 (80–96%) at 1.0–1.7%.

The loose filter roughly doubled the false discovery proportion (2.3–2.9% with MBR; 3.3–3.7% without) (Fig. 3a, b). It increased the number of testable proteins (4,534 with MBR) and the number of true changes found (MinDet 1,102 vs 958 for no imputation on the strict subset), although the fraction of testable changes found fell (MinDet 1,102 of 1,415, 78%, vs 958 of 1,039, 92%), partly because the loose filter adds harder-to-detect proteins to the denominator. Ratio accuracy also fell: without MBR, MinDet placed 43% of *E. coli* proteins within ± 0.5 (centred), compared with 74% for the strict, unimputed set.

kNN imputation behaved differently from the left-censored methods. *E. coli* proteins are often missing in condition A because they are scarce there, which violates kNN's assumption that values are missing at random. Under the loose filter, consistent with kNN filling these gaps with values from similar, measured proteins, the median centred *E. coli* ratio fell from the true + 2 to + 0.02 without MBR (+ 1.05 with MBR) (Fig. 3c). QRILC moved the ratios in the opposite direction (median + 2.47 under the loose filter without MBR).

Differential testing

At BH-adjusted $p < 0.05$ without a fold-change threshold, the number of unchanged human proteins called significant differed by more than an order of magnitude between engines and tests (15 to 284) (Fig. 4a; Table 1). On MaxQuant LFQ, the t-test called 68, limma 104 and DEqMS 284 (false discovery proportions 11.8%, 10.5% and 24.1%), and 96–100% of these calls were in the direction of the + 0.145 offset of the human background. On FragPipe output, where the human median was + 0.014, the same tests called 15, 43 and 37. On DIA-NN output, where the human median was – 0.010, they called 106, 167 and 188, of which 65–74% were in the negative direction. DIA-NN ratios were less variable (human IQR 0.082 on all scored proteins, vs 0.142 for MaxQuant and 0.183 for FragPipe), which may let smaller deviations reach significance; I did not investigate which proteins deviated or why.

Table 1

Unchanged human proteins called significant (BH-adjusted $p < 0.05$), by engine, test and threshold. Proteins with all six values; contaminants removed. Changed proteins are yeast and *E. coli*; true calls are counted regardless of sign.

Engine	Proteins tested (changed)	Human median log2FC	Test	False calls, no FC threshold (FDP)	True calls	False calls, $ \log2FC \geq 0.5$ (FDP)	True calls
MaxQuant LFQ	3,532 (906)	+ 0.145	t-test	68 (11.8%)	507	2 (0.4%)	503
			limma	104 (10.5%)	891	10 (1.1%)	877
			DEqMS	284 (24.1%)	893	13 (1.5%)	880
FragPipe MaxLFQ	6,226 (1,905)	+ 0.014	t-test	15 (1.6%)	928	7 (0.7%)	927
			limma	43 (2.6%)	1,629	35 (2.1%)	1,625
			DEqMS	37 (2.2%)	1,628	34 (2.0%)	1,627
DIA-NN	7,747 (2,485)	–0.010	t-test	106 (4.6%)	2,178	5 (0.2%)	2,169
			limma	167 (6.5%)	2,415	21 (0.9%)	2,399
			DEqMS	188 (7.2%)	2,425	22 (0.9%)	2,409

Adding $|\log2FC| \geq 0.5$ reduced false calls to 2 / 10 / 13 (MaxQuant) and 5 / 21 / 22 (DIA-NN), with false discovery proportions at or below 1.5%, while losing at most 1.6% of the true calls. For FragPipe the threshold had little effect on limma and DEqMS (43 to 35 and 37 to 34; t-test 15 to 7): these remaining calls were human proteins with large ratio errors, split about evenly between the two directions, rather than consequences of an offset.

An extra per-sample median centring behaved differently by engine (Fig. 4b). It reduced MaxQuant's offset (to + 0.109; limma 53 false calls) but introduced one in FragPipe (+ 0.130; 105) and DIA-NN (+ 0.190; limma 3,461 false calls, 59% of all calls). In this benchmark a quarter to a third of the tested proteins change and most of the changed proteins (yeast) decrease, so the per-sample median no longer represents an unchanged background. Median centring removes any constant per-sample scaling that an engine applied, so the remaining offset is

largely the one centring itself induces, which depends on the fraction and composition of changed proteins in each protein set (26% changed for MaxQuant, 31% for FragPipe, 32% for DIA-NN). For MaxQuant this induced offset (+ 0.109) was smaller than the offset of its own LFQ output (+ 0.145); for FragPipe and DIA-NN it was larger than their near-zero starting offsets. This explanation is consistent with the numbers but was not tested directly, and the engines' normalisations are not purely constant shifts. Real experiments usually change fewer proteins, but the result shows that a second normalisation on top of an engine's own can create, rather than remove, a systematic offset.

Enrichment background

For 20 random lists of 300 detected human proteins, enrichment against STRING's default whole-genome background returned a median of 103 GO and KEGG terms at FDR < 0.05 (range 63–168). Thirty-five terms were significant in all 20 lists (17 GO Biological Process, 16 Cellular Component and 2 Molecular Function terms), such as cytosol, cytoplasm and nucleoplasm, reflecting what proteomics detects rather than what the lists contain. With the detected proteins as background, no list returned any term, which is expected because the lists were drawn from that background (Fig. 5a); the informative result is the inflation under the genome background. In the five planted lists, the cholesterol-biosynthesis signal was recovered in all ten enrichments (Fig. 5b); with the detected background it was among 9–14 significant GO and KEGG terms per list, compared with 110–228 with the genome background, where in one list the first sterol-related term ranked 21st of 110.

Software pitfalls

Several version-specific behaviours and setup traps affected the analysis (Table 2); five of them gave no error. The most consequential: the decoy column, called Reverse in older MaxQuant output, is named Decoy in the version used here (2.8.1), and a widely used base-R filter, `pg[pg$Reverse != "+", ]`, returns zero rows when the column is absent, so a script written for older output can silently produce an empty table.

Table 2

Version-specific behaviours and setup traps encountered. Rows marked * produced no error message.

Tool	Behaviour	Consequence	Check
MaxQuant 2.8.1*	Decoy column named Decoy (Reverse in older output)	Base-R filter <code>pg[pg\$Reverse != "+",]</code> returns 0 rows	Test that filters keep the expected number of rows
MaxQuant 2.8.1 (MaxQuantCmd --create)*	Experiment names truncated (e.g. Condition_A_Sample_Alpha_01 → A_Sample_Alph)	Column names do not match the design table	Set experiment names explicitly in mqpar.xml
ThermoRawFileParser 2.0.0 (v2.0.0-dev)*	-L selects MS levels (not log level)	-L 1 writes MS1 only; the search later finds 0 MS/MS scans	Do not pass -L 1 for MS/MS searches
FragPipe 24	Sub-tools use the java on PATH; headless mode requires diaTracer even for DDA	Java 8 on PATH fails with <code>UnsupportedClassVersionError</code>	Put a current Java first on PATH; install all tools
FragPipe 24 / Philosopher*	Added contaminant sequences carried no contam_ prefix in this run	Contaminants not removed by a prefix filter	Check the database headers; remove contaminants by accession
DIA-NN 2.3.2*	First-pass report has no protein quantities (PG.MaxLFQ = 0)	Protein matrix empty if only the first pass is used	Use the MBR pass or re-summarise with --use-quant

Discussion

In this three-versus-three benchmark, several steps after the search engine changed the number of false discoveries as much as, or more than, the choice of engine.

First, the false discoveries of the more powerful tests on MaxQuant LFQ output followed a small offset in the unchanged background. DEqMS called 284 unchanged proteins, nearly all in the direction of a + 0.145 offset; on FragPipe output, with almost no offset, it called 37. FDR control does not protect against a systematic offset, and here the tests with more power detected more of it on MaxQuant output, although not consistently on FragPipe output (limma 43, DEqMS 37). The MaxQuant offset is consistent with the LFQ normalisation's assumption that most proteins do not change [9], which this benchmark violates by design; IonQuant's MaxLFQ implementation, which also normalises, showed almost no offset, so the size of the effect depends on the implementation, and I did not test why. A fold-change threshold removed most offset-driven calls at little cost in power here, but it did not remove FragPipe's remaining false calls, which were individual proteins with large ratio errors.

Second, normalisation choices interact. An extra median centring improved MaxQuant output but damaged FragPipe and DIA-NN output, in which the engines' own normalisation had already placed the unchanged background near zero. Normalisation methods depend on assumptions about the fraction and direction of change [16], and applying a second method on top of an engine's own should be a deliberate decision, checked against proteins expected to be stable.

Third, choices that increase coverage trade accuracy. MBR improved the accuracy of proteins that would be quantified anyway while adding less accurate proteins; MaxQuant's MBR was run without its optional FDR control, and false transfers are a known risk [10]. Imputation after a loose filter found more changes in absolute terms but a smaller fraction of the testable changes, at about twice the false discovery proportion; with a strict filter the choice of imputation method mattered little. kNN, which assumes values are missing at random, pulled the median *E. coli* ratio to near zero in one setting where values are missing because a protein is scarce [11]. For enrichment, the whole-genome background turned random detected proteins into about 100 "significant" terms, consistent with long-standing warnings about background choice [17].

These observations suggest a short checklist for small label-free studies (Table 3): record the engine and version and check scripts after upgrades; use a strict valid-value filter unless the extra coverage is needed, and if imputing, match the method to the missingness mechanism; pair FDR with a fold-change threshold and check the median of proteins expected to be stable; avoid stacking normalisations; and use the detected proteins as enrichment background.

Table 3
Checklist for small label-free comparisons.

Step	Recommendation from this benchmark
Software	Record engine and version; re-check filters after upgrades
Intensity	Use MaxLFQ-type intensities for ratios; iBAQ does not change between-sample ratios
MBR	Keep on for coverage; flag proteins quantified only through matching
Missing values	Prefer a strict valid-value filter; if imputing, avoid methods that assume values are missing at random for low-abundance proteins
Testing	Pair FDR with a fold-change threshold; check the median of proteins expected to be stable
Normalisation	Do not stack a second normalisation on an engine's own without checking the stable background
Enrichment	Use the detected proteins as background

Limitations

The study has important limitations. It uses one benchmark dataset, one injection per sample and a three-versus-three design, so each number is a single realisation without a confidence interval; the dataset contains four Orbitrap DDA and three SWATH injections per preparation, and an analysis resampling across injections is planned for a revised version of this preprint. The samples are replicate preparations of two mixtures, so biological variation is absent. The low cost of the fold-change threshold is a property of this design, in which the true changes are -1 and $+2$; in experiments where many true changes are below $0.5 \log_2$, the same threshold

would discard real changes, so the recommendation is to pair FDR with a threshold suited to the expected effect sizes rather than to use 0.5 everywhere. DEqMS's count input differed between engines (peptides, spectra, sequences), so its cross-engine comparison mixes engine and input effects. The ± 0.5 tolerance is a convention; at ± 0.3 the differences between engines grew, and at ± 1.0 they mostly disappeared. The DIA-NN results come from a different instrument and acquisition mode, with a slightly different search space and no contaminant entries, and should not be read as a software comparison with the DDA workflows. Search settings were mostly defaults and were not tuned for each engine, and MaxQuant's MBR was run without its FDR option. The hybrid design, with a third of proteins changing and large fold changes, is more extreme than most biological experiments; in particular, the effect of median normalisation depends on the fraction and direction of change. Imputation was evaluated on MaxQuant output only, and enrichment on DIA-NN-detected proteins with STRING only. Finally, the DIA-NN no-MBR comparison re-used first-pass quantification files and is indicative only.

Conclusions

On a public benchmark with known ratios, single decisions after the search engine (extra normalisation, valid-value filtering, fold-change thresholds and enrichment background) changed the number of false discoveries in a small label-free experiment by up to 25- to 34-fold, while the choice of imputation method mattered little under a strict filter (under a loose filter, kNN shifted one species' ratios towards zero). Simple safeguards removed much of the false signal. Because the answer is known, the same scripts can be used to test other tools and settings; all are publicly available.

Declarations

Data and code availability. Raw data: PRIDE PXD028735. Scripts, parameter files and result tables: <https://github.com/shoo99/sbmlab-benchmarks>, archived at Zenodo, doi:10.5281/zenodo.23355028.

Use of AI tools. An AI assistant (Claude Opus 5.5, Anthropic, used through Claude Code between 6 and 14 October 2026) was used under the author's direction to write and run analysis scripts, to draft the text and to check reported numbers against the result tables by independent recomputation. The author designed the study, reviewed the analyses and the manuscript, and takes full responsibility for the content. The same statement is given in Methods.

Competing interests. The author is affiliated with SYSOFT and runs sbmlab.com, an advertising-supported website on which parts of these results were described. The author declares no other competing interests.

Prior dissemination. Parts of these results were described in blog posts on sbmlab.com in October 2026.

Funding. This work received no specific funding.

Acknowledgements. I thank Van Puyvelde et al. for making the benchmark dataset public.

References

1. Kuharev J, Navarro P, Distler U, Jahn O, Tenzer S (2015) In-depth evaluation of software tools for data-independent acquisition based label-free quantification. *Proteomics* 15:3140–3151. 10.1002/pmic.201400396
2. Navarro P, Kuharev J, Gillet LC et al (2016) A multicenter study benchmarks software tools for label-free proteome quantification. *Nat Biotechnol* 34:1130–1136. 10.1038/nbt.3685
3. Van Puyvelde B, Daled S, Willems S et al (2022) A comprehensive LFQ benchmark dataset on modern day acquisition strategies in proteomics. *Sci Data* 9:126. 10.1038/s41597-022-01216-6
4. Gotti C, Roux-Dalvai F, Joly-Beauparlant C et al (2021) Extensive and accurate benchmarking of DIA acquisition methods and software tools using a complex proteomic standard. *J Proteome Res* 20:4801–4814. 10.1021/acs.jproteome.1c00490
5. Fröhlich K, Brombacher E, Fahrner M et al (2022) Nat Commun 13:2622. 10.1038/s41467-022-30094-0. Benchmarking of analysis strategies for data-independent acquisition proteomics using a large-scale dataset comprising inter-patient heterogeneity
6. Lou R, Cao Y, Li S et al (2023) Benchmarking commonly used software suites and analysis workflows for DIA proteomics and phosphoproteomics. *Nat Commun* 14:94. 10.1038/s41467-022-35740-1
7. Välikangas T, Suomi T, Elo LL (2018) A comprehensive evaluation of popular proteomics software workflows for label-free proteome quantification and imputation. *Brief Bioinform* 19:1344–1355. 10.1093/bib/bbx054
8. Devreese R, Jachmann C, Van Puyvelde B et al (2025) ProteoBench: the community-curated platform for comparing proteomics data analysis workflows. *bioRxiv*. 10.64898/2025.12.09.692895
9. Cox J, Hein MY, Lubner CA, Paron I, Nagaraj N, Mann M (2014) Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol Cell Proteom* 13:2513–2526. 10.1074/mcp.M113.031591
10. Lim MY, Paulo JA, Gygi SP (2019) Evaluating false transfer rates from the match-between-runs algorithm with a two-proteome model. *J Proteome Res* 18:4020–4026. 10.1021/acs.jproteome.9b00492
11. Lazar C, Gatto L, Ferro M, Bruley C, Burger T (2016) Accounting for the multiple natures of missing values in label-free quantitative proteomics data sets to compare imputation strategies. *J Proteome Res* 15:1116–1125. 10.1021/acs.jproteome.5b00981
12. Pursiheimo A, Vehmas AP, Afzal S et al (2015) Optimization of statistical methods impact on quantitative proteomics data. *J Proteome Res* 14:4118–4126. 10.1021/acs.jproteome.5b00183
13. Kammers K, Cole RN, Tiengwe C, Ruczinski I (2015) Detecting significant changes in protein abundance. *EuPA Open Proteom* 7:11–19. 10.1016/j.euprot.2015.02.002
14. Ritchie ME, Phipson B, Wu D et al (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 43:e47. 10.1093/nar/gkv007
15. Zhu Y, Orre LM, Zhou Tran Y et al (2020) DEqMS: a method for accurate variance estimation in differential protein expression analysis. *Mol Cell Proteom* 19:1047–1057. 10.1074/mcp.TIR119.001646
16. Välikangas T, Suomi T, Elo LL (2018) A systematic evaluation of normalization methods in quantitative label-free proteomics. *Brief Bioinform* 19:1–11. 10.1093/bib/bbw095
17. Timmons JA, Szkop KJ, Gallagher IJ (2015) Multiple sources of bias confound functional enrichment analysis of global -omics data. *Genome Biol* 16:186. 10.1186/s13059-015-0761-7

18. Wijesooriya K, Jadaan SA, Perera KL, Kaur T, Ziemann M (2022) Urgent need for consistent standards in functional enrichment analysis. *PLoS Comput Biol* 18:e1009935. 10.1371/journal.pcbi.1009935
19. Szklarczyk D, Kirsch R, Koutrouli M et al (2023) The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res* 51:D638–D646. 10.1093/nar/gkac1000
20. Cox J, Mann M (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* 26:1367–1372. 10.1038/nbt.1511
21. Tyanova S, Temu T, Cox J (2016) The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat Protoc* 11:2301–2319. 10.1038/nprot.2016.136
22. Kong AT, Leprevost FV, Avtonomov DM, Mellacheruvu D, Nesvizhskii AI (2017) MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat Methods* 14:513–520. 10.1038/nmeth.4256
23. Yu F, Haynes SE, Nesvizhskii AI (2021) IonQuant enables accurate and sensitive label-free quantification with FDR-controlled match-between-runs. *Mol Cell Proteom* 20:100077. 10.1016/j.mcpro.2021.100077
24. Yang KL, Yu F, Teo GC et al (2023) MSBooster: improving peptide identification rates using deep learning-based features. *Nat Commun* 14:4539. 10.1038/s41467-023-40129-9
25. Käll L, Canterbury JD, Weston J, Noble WS, MacCoss MJ (2007) Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat Methods* 4:923–925. 10.1038/nmeth1113
26. da Veiga Leprevost F, Haynes SE, Avtonomov DM et al (2020) Philosopher: a versatile toolkit for shotgun proteomics data analysis. *Nat Methods* 17:869–870. 10.1038/s41592-020-0912-y
27. Hulstaert N, Shofstahl J, Sachsenberg T et al (2020) ThermoRawFileParser: modular, scalable, and cross-platform RAW file conversion. *J Proteome Res* 19:537–542. 10.1021/acs.jproteome.9b00328
28. Demichev V, Messner CB, Vernardis SI, Lilley KS, Ralser M (2020) DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nat Methods* 17:41–44. 10.1038/s41592-019-0638-x
29. Tyanova S, Temu T, Sinitcyn P et al (2016) The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat Methods* 13:731–740. 10.1038/nmeth.3901
30. Troyanskaya O, Cantor M, Sherlock G et al (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics* 17:520–525. 10.1093/bioinformatics/17.6.520
31. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B* 57:289–300. 10.1111/j.2517-6161.1995.tb02031.x
32. Smyth GK (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol.* ;3:Article 3. 10.2202/1544-6115.1027

Figures

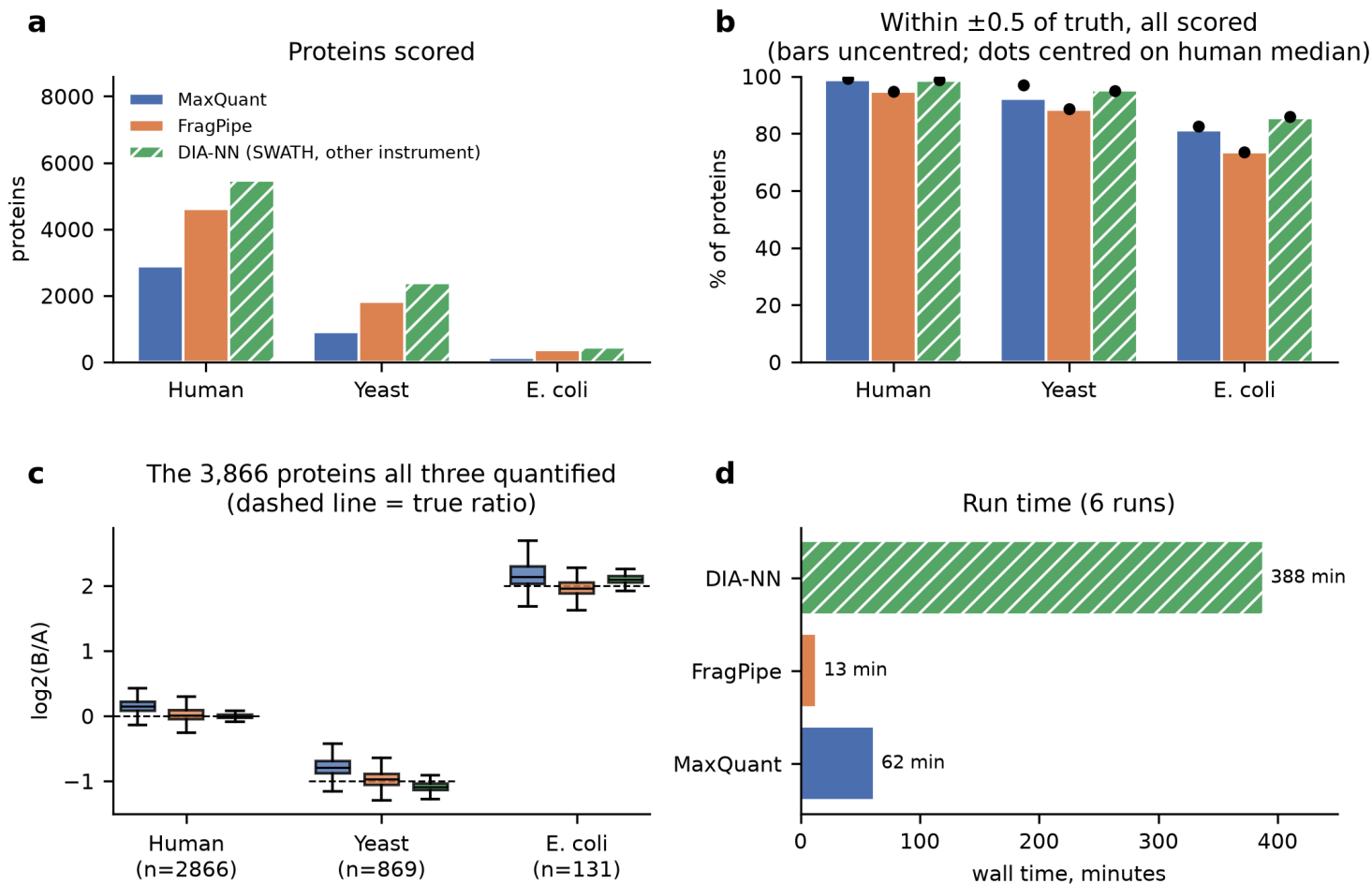


Figure 1

Search engines on the same samples. (a) Proteins scored (at least two of three values per condition) per species. (b) Fraction of scored proteins with $\log_2(B/A)$ within ± 0.5 of the truth: bars uncentred, dots after centring each engine's ratios on its human median. (c) Ratios of the 3,866 proteins quantified by all three tools; dashed lines show the true ratios (0, -1, +2); boxes show interquartile ranges, whiskers $1.5 \times \text{IQR}$, outliers not shown. (d) Wall time for six runs (FragPipe: 9.3 min wall clock plus about 4 min of raw-to-mzML conversion; DIA-NN includes library prediction). MaxQuant and FragPipe: Orbitrap DDA, 16 threads under WSL2; DIA-NN (hatched): TripleTOF 6600+ SWATH of the same samples, 24 threads on native Windows.

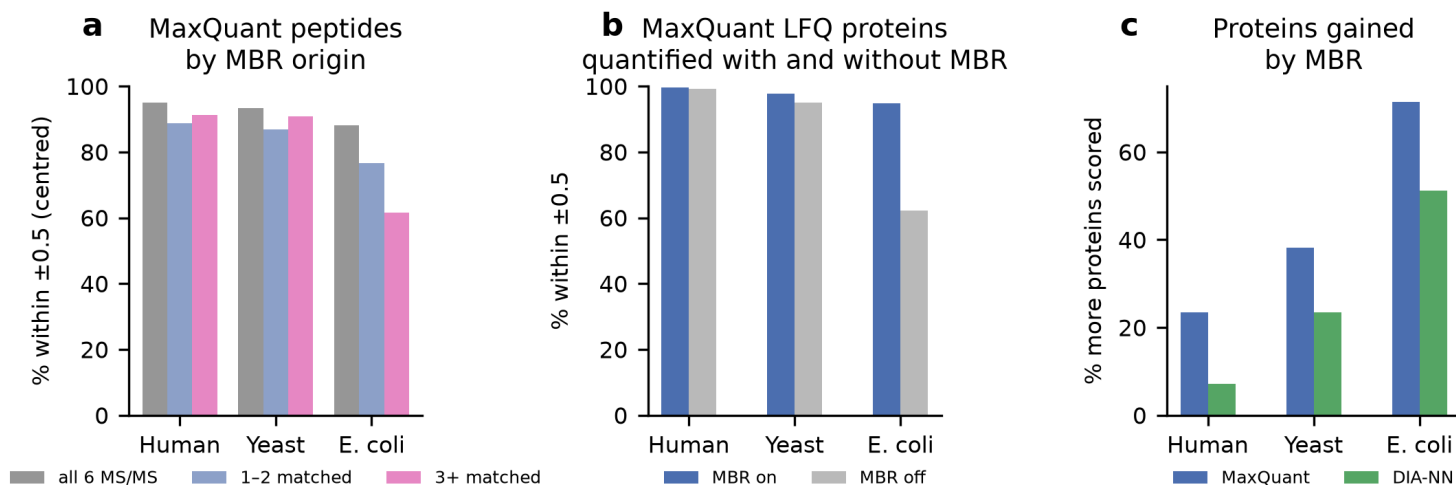


Figure 2

Match-between-runs. (a) MaxQuant peptides grouped by the number of their six values that came from matching, scored against the truth after centring on the human median. (b) MaxQuant LFQ proteins scoreable both with and without MBR: fraction within ± 0.5 of the truth in each run (uncentred). (c) Additional proteins scored with MBR, relative to the run without it, for MaxQuant and DIA-NN (DIA-NN no-MBR values re-summarised from first-pass files).

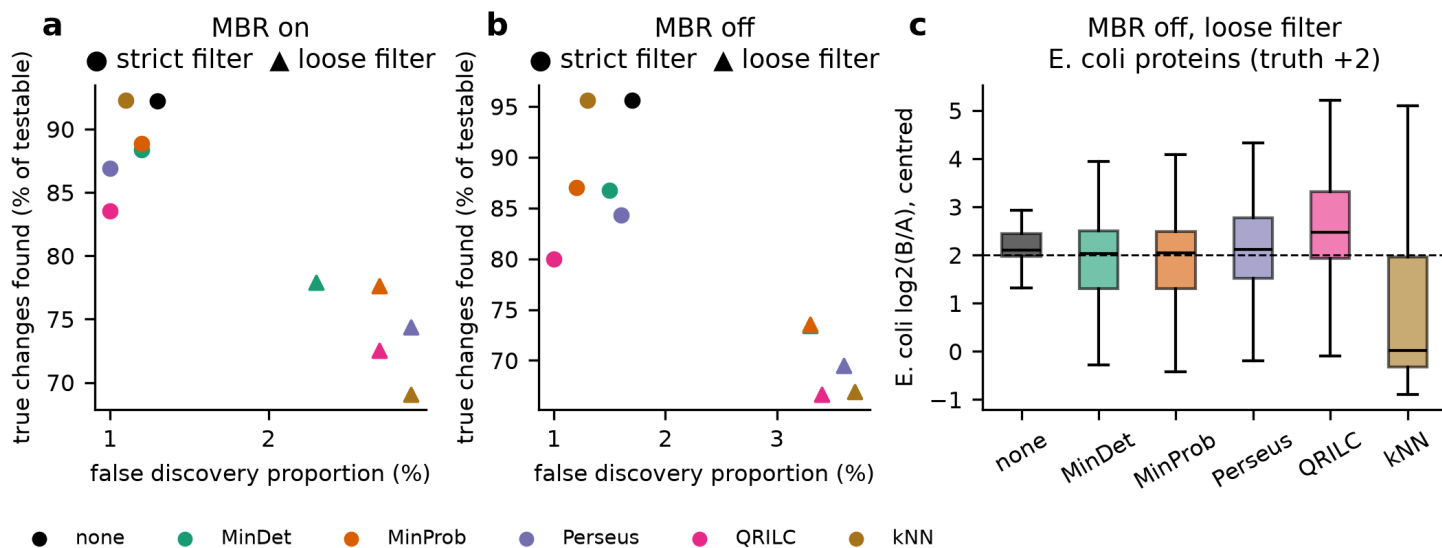


Figure 3

Imputation. (a, b) Fraction of testable changed proteins found versus false discovery proportion (limma, BH-adjusted $p < 0.05$ and $|\log_2 FC| \geq 0.5$) for each method under the strict (circles) and loose (triangles) filters, with MBR (a) and without (b). The number of testable changed proteins is larger under the loose filter (1,415 vs 1,039 with MBR; 1,163 vs 733 without). In (b), the MinDet and MinProb loose-filter points overlap (73.4% and 73.5% at 3.3%). (c) E. coli ratios under the loose filter without MBR, centred on the human median; the dashed line shows the truth (+2); "none" includes only proteins passing the strict filter.

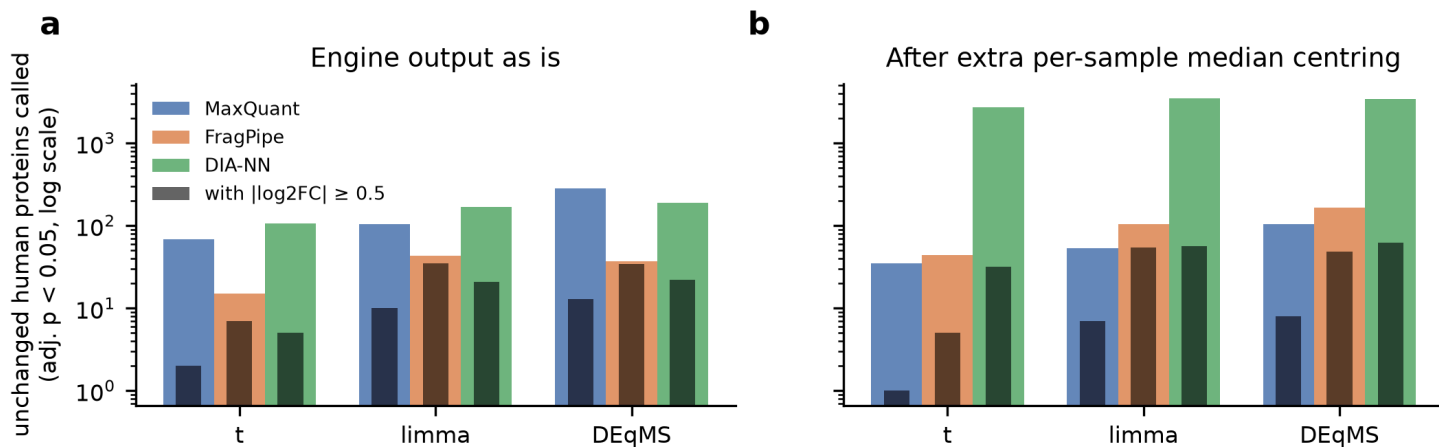


Figure 4

Differential testing across engines. Unchanged human proteins called significant (BH-adjusted $p < 0.05$; log scale) by t-test, limma and DEqMS, without (coloured bars) and with (dark bars) a $|\log_2\text{FC}| \geq 0.5$ threshold, on each engine's output as is (a) and after an additional per-sample median centring (b). Contaminants removed.

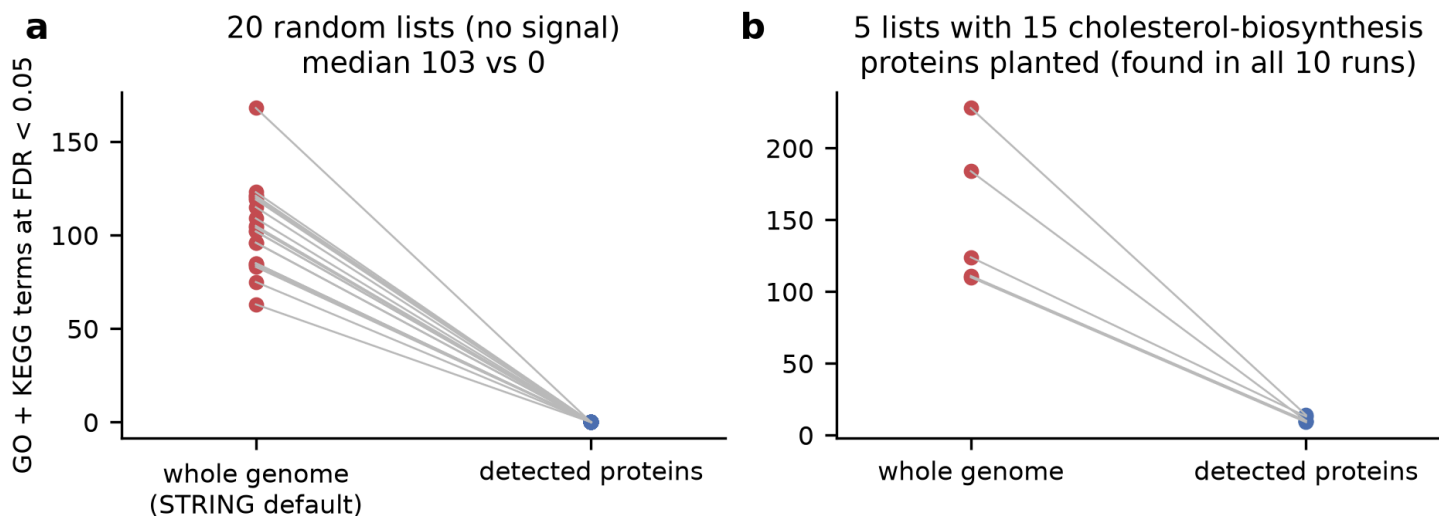


Figure 5

Enrichment background. GO and KEGG terms at $\text{FDR} < 0.05$ with STRING's default whole-genome background and with the detected proteins as background; lines connect the same list. (a) 20 random lists of 300 detected human proteins (zero with the detected background by construction). (b) Five lists of 15 detected cholesterol-biosynthesis proteins plus 285 random detected proteins; the planted pathway was significant with both backgrounds in every list.