Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

PERSEO: adaptive per-feature distribution modelling for differential analysis of heterogeneous omics data

Val Fernández-Lanza

valfernandez.vf@gmail.com

Ramón y Cajal Health Research Institute (IRYCIS) https://orcid.org/0000-0003-0870-9500

Sergio Olmos-Piñero

Ramón y Cajal Health Research Institute (IRYCIS)

Article

Keywords:

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11702308/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

PERSEO: adaptive per-feature distribution modelling for differential analysis of heterogeneous omics data

Sergio Olmos-Piñero[1,2], Val Fernández-Lanza[1,3*]

[1]*Translational Genomics and Bioinformatics Unit, Ramón y Cajal Institute for Health Research (IRYCIS), Madrid, Spain.*

[2]*Department of Biochemistry, University Hospital Ramón y Cajal-IRYCIS, Madrid, Spain*

[3]*CIBER de Enfermedades Infecciosas (CIBERINFEC), Instituto de Salud Carlos III, Madrid, Spain.*

Abstract

Differential analysis is fundamental to biomarker discovery across omics disciplines. However, existing statistical tools enforce a single probability distribution across all features, introducing systematic misspecification errors that compromise statistical power and false discovery control. Here we characterized 131 public datasets from nine omics profiles and found that no profile is governed by a single distribution, while standard normalizations merely displace empirical distributions rather than conforming them to downstream model assumptions. To overcome modality-specific constraints, we present PERSEO, an omics-agnostic R framework for feature-wise differential analysis across any high-throughput layer, irrespective of origin or platform. By operating within generalized additive models for location, scale and shape (GAMLSS) with Jacobian-corrected information criteria, PERSEO adaptively models each feature under its optimal distribution, evaluating counts, continuous intensities, proportions, and zero-adjusted measurements on a unified scale. In benchmarks spanning 43 profile-and-stress scenarios, PERSEO matched reference methods in bulk RNA-sequencing and outperformed specialized tools in metagenomics, proteomics, and metabolomics, where its F1-score advantage reached 0.19 under noise and outlier contamination. On 12 public cohorts, PERSEO resolved cross-method discordances, recovering verified differences while exposing pervasive false discoveries in field-specific tools. PERSEO is freely available at https://github.com/OPSergio/Perseo.

Introduction

Omics technologies quantify molecular entities such as transcripts, proteins, metabolites and microbial taxa across experimental conditions. RNA-sequencing, metagenomics, mass spectrometry-based proteomics and metabolomic profiling share a common quantitative structure: they assign measured signals to biological classes and generate abundance values that

reflect the relative representation of each entity in a sample.[1] One of the most common analytical tasks is the identification of entities that differ between groups, such as differentially expressed genes in transcriptomics, which can reveal regulatory processes, functional alterations or disease-associated mechanisms.[2]

Ideally, modelling an omics feature would begin by assessing its distribution to select an appropriate statistical family, such as a negative binomial for overdispersed counts, a zero-inflated model for excess zeros or a log-normal for continuous intensities. In practice, a vast ecosystem of differential analysis tools has emerged, but most were engineered for a specific omics layer, tailoring their statistical foundation to the assumed global properties of that data type. In bulk RNA-sequencing, frameworks such as DESeq2 and edgeR adopted negative binomial models with empirical Bayes shrinkage to capture count overdispersion under limited replication,[1–3] while limma-voom coupled precision weights with normal linear models.[4,5] In mass spectrometry-based proteomics, MSstats and limma handle continuous, transformed intensities under Gaussian assumptions,[6–8] and in metabolomics, standard pipelines rely on normal approximations via log-transformed Welch's t-tests or limma.[9] In metagenomics, tools such as ANCOM-BC2 and ALDEx2 were devised to accommodate compositionality, library-size variation and pervasive zeros through log-ratio transformations or linear models with bias correction.[10–12] Yet, despite addressing distinct technological paradigms, these field-specific methods share a common constraint: each assumes that all features within an assay follow the same single probabilistic distribution after normalization. These approaches perform well when the data fit that assumption, but real omics datasets show substantial feature-wise heterogeneity: some features follow overdispersed counts,[2] others carry structural zeros,[13,14] and metagenomic data present compositional effects that have motivated beta-binomial modelling.[11,15] Zero inflation has been discussed across RNA-sequencing, single-cell and metagenomic studies, where excess zeros may reflect biological absence or technical capture variability.[11,16] Because overdispersion, zero inflation and baseline variance differ in origin and statistical behaviour, forcing one distribution on all features leads to model misspecification, and differential abundance methods built on such assumptions have been shown not to control the false discovery rate.[17]

Differential analysis across high-throughput modalities also faces well-documented technical challenges that extend far beyond bulk transcriptomics.[18–20] The detection of truly differential features is strongly contingent on experimental design, sample size, and the treatment of low-abundance signals, where noisy fold-change estimates and extreme dropout severely hinder robust ranking and classification.[2,14,21] While early transcriptomic frameworks relied on Poisson

models that underestimated biological variability and inflated false discoveries,[1,22] negative binomial formulations, such as those implemented in edgeR and DESeq2, addressed overdispersion by borrowing information across features via empirical Bayes shrinkage.[2,3,23–26] Similar variance-stabilization and moderation strategies had been developed for microarrays, where limma moderates feature-wise variances toward a global trend to preserve power under small sample sizes,[5,6,8] whereas MSstats, used in quantitative proteomics, relies on linear mixed models.[7] In metagenomics, where compositional constraints and library-size variation distort relative abundances, methods like ALDEx2 and ANCOM-BC2 introduce Dirichlet-multinomial sampling, log-ratio transformations, or bias-corrected estimating equations to account for sampling zeros and uneven depth.[10–12] Yet, across all these modalities, linear and generalized linear frameworks remain fundamentally sensitive to outliers, asymmetric tail behaviours, and mixtures of biological and structural zeros.[2,14,27] Consequently, while empirical moderation and specialized *ad hoc* heuristics alleviate small-sample instability, they do not resolve the systematic bias caused by fitting an inflexible, misspecified distributional form to an inherently heterogeneous feature space.

We developed PERSEO, a feature-wise differential analysis framework based on generalized additive models for location, scale and shape (GAMLSS).[28] Instead of imposing one family on all features, PERSEO selects the distribution of each feature by information criteria corrected with the Jacobian of the transformation, so that the same procedure applies to counts, continuous intensities, proportions and zero-adjusted measurements from any omics layer. The study proceeded in three steps. First, we characterized the distribution families of 131 public datasets spanning bulk and single-cell transcriptomics, metagenomics, proteomics and metabolomics, at the raw scale and after the normalizations of each field. Second, we simulated nine omics profiles from the family mixtures observed in the first step, under five scenarios that each add one source of noise or bias, and compared PERSEO with the reference method of each discipline, run on the input its guidelines recommend. Third, we compared the same methods on 12 public datasets, interpreted their disagreements through the simulation and used label permutation to measure false discoveries on real data.

Results

Study design and datasets

Throughout the study, a dataset denotes one public study, that is, one feature-by-sample matrix with its metadata, and a sample denotes one biological specimen within it; in single-cell data the

samples are cells, which are aggregated into one pseudobulk profile per donor for differential analysis. The characterization corpus comprised 135 datasets from five omics layers (Supplementary Table 1). Bulk RNA sequencing contributed 21 datasets, analysed as two profiles because they differ in scale and origin: 10 consortium cohorts from the Genotype-Tissue Expression project (GTEx) and The Cancer Genome Atlas (TCGA), with 524–2,931 samples each (median 794; consortium RNA-seq), and 11 independent studies from individual laboratories deposited in the Sequence Read Archive (SRA), with 12–80 samples each (independent-study RNA-seq). Single-cell RNA sequencing contributed 23 datasets from three protocols: 8 droplet-based with unique molecular identifiers (UMI; 5,183–49,300 cells), 7 plate-based with UMI (1,977–51,840 cells) and 8 plate-based full-length (288–3,514 cells). Metagenomics contributed 39 16S rRNA amplicon datasets (16S metagenomics; 24–2,466 samples, median 78) and 20 whole-genome shotgun metagenomes (WGS metagenomics; 32–292 samples, median 150). Proteomics contributed 12 tumour proteomes from the Clinical Proteomic Tumor Analysis Consortium (CPTAC) (110–230 samples) and metabolomics 20 Metabolomics Workbench studies (90–1,014 samples), 16 of which provided data at the raw scale. The 131 datasets with raw-scale data defined nine omics profiles: two of bulk RNA sequencing, three of single-cell RNA sequencing, two of metagenomics, one of proteomics and one of metabolomics.

For the benchmark on real data we selected 12 of these datasets, two per omics layer, at the extremes of the effective sample size available in the corpus (Supplementary Table 2; numbers are cases versus controls): bulk RNA sequencing of quadriceps muscle from a mouse model of Huntington's disease (SRP074887; 10 versus 10 samples) and of breast tumours against normal tissue from TCGA-BRCA (1,095 versus 113 samples); single-cell RNA sequencing of pancreatic alpha cells from donors with type 2 diabetes (4 versus 6 donors samples) and of bone marrow plasma cells from patients with multiple myeloma (14 versus 11 donors samples); 16S metagenomes of *Clostridioides difficile* infection (17 versus 23 samples) and colorectal cancer (117 versus 146 samples); WGS metagenomes of schizophrenia (16 versus 16 samples) and liver cirrhosis (123 versus 114 samples); proteomes of glioblastoma (100 versus 10 samples) and pancreatic ductal adenocarcinoma (98 versus 45 samples); and metabolomes of A375 melanoma cells treated with ranolazine (8 versus 8 samples) and of plasma from men with and without prostate cancer (267 versus 313 samples).

No single family describes omics data

We first identified, for every feature of the 131 datasets with raw-scale data, the GAMLSS family selected by the Bayesian information criterion (Fig. 1a, Supplementary Table 3; family codes in

Supplementary Table 4). No omics profile was described by a single family. In bulk RNA sequencing, the Poisson–inverse Gaussian (PIG) family was modal in all 10 consortium cohorts (59% of genes, with the negative binomial (NBI) at 32%), whereas among the 11 independent studies PIG was modal in six, NBI in four and the zero-inflated negative binomial (ZINBI) in one, and the Poisson (PO) share rose to 22%. Single-cell RNA sequencing followed the same overdispersed families, with a zero-inflated Poisson (ZIP) component of 8% and 11% of genes in the droplet and plate UMI protocols. In metagenomics, 16S data combined NBI (36%), ZIP (34%) and PIG (21%), with ZIP modal in 11 of 39 datasets, whereas WGS data were almost exclusively NBI (89%; modal in all 20). The proteomes split between Student's *t* (TF; 48%) and Gaussian (NO; 46%) families, each modal in six of the 12 datasets. Metabolomics was the most heterogeneous layer: eight families exceeded 5% of features, the modal family varied between studies, and the zero-adjusted inverse Gaussian (ZAIG) and zero-adjusted gamma (ZAGA) families together accounted for 34% of metabolites. A principal component analysis of the raw family compositions separated the datasets by profile (Fig. 1b; 41% and 24% of the variance on the first two components), with the largest loadings for NBI, PIG, ZIP, NO and TF.

Normalization relocated these families rather than producing the family assumed by the downstream test (Fig. 1c, Supplementary Table 3). In RNA-sequencing, TMM scaling moved most genes to a beta law (44% of genes in consortium and 42% in independent-study RNA-seq), DESeq2 size factors to ZAIG (57% and 35%) and the variance-stabilizing transformation to TF (50% and 33%), with a Gaussian fit for only 5% and 17% of genes. In 16S metagenomics, TMM produced zero-adjusted families (ZAIG 61% and ZAGA 39%) and rarefied proportions a beta law in every dataset. Variance-stabilizing normalization brought proteomic data closest to a single family (NO 87%), but in metabolomics it produced mainly TF (54%). In single-cell data, log-normalization produced zero-adjusted families, whereas pseudobulk aggregation returned PIG and NBI and reduced the zero-inflated component of droplet data below 5%. These observations motivate modelling each feature under the family that fits it, and they provided the generative mixtures of the synthetic benchmark (Supplementary Table 5).

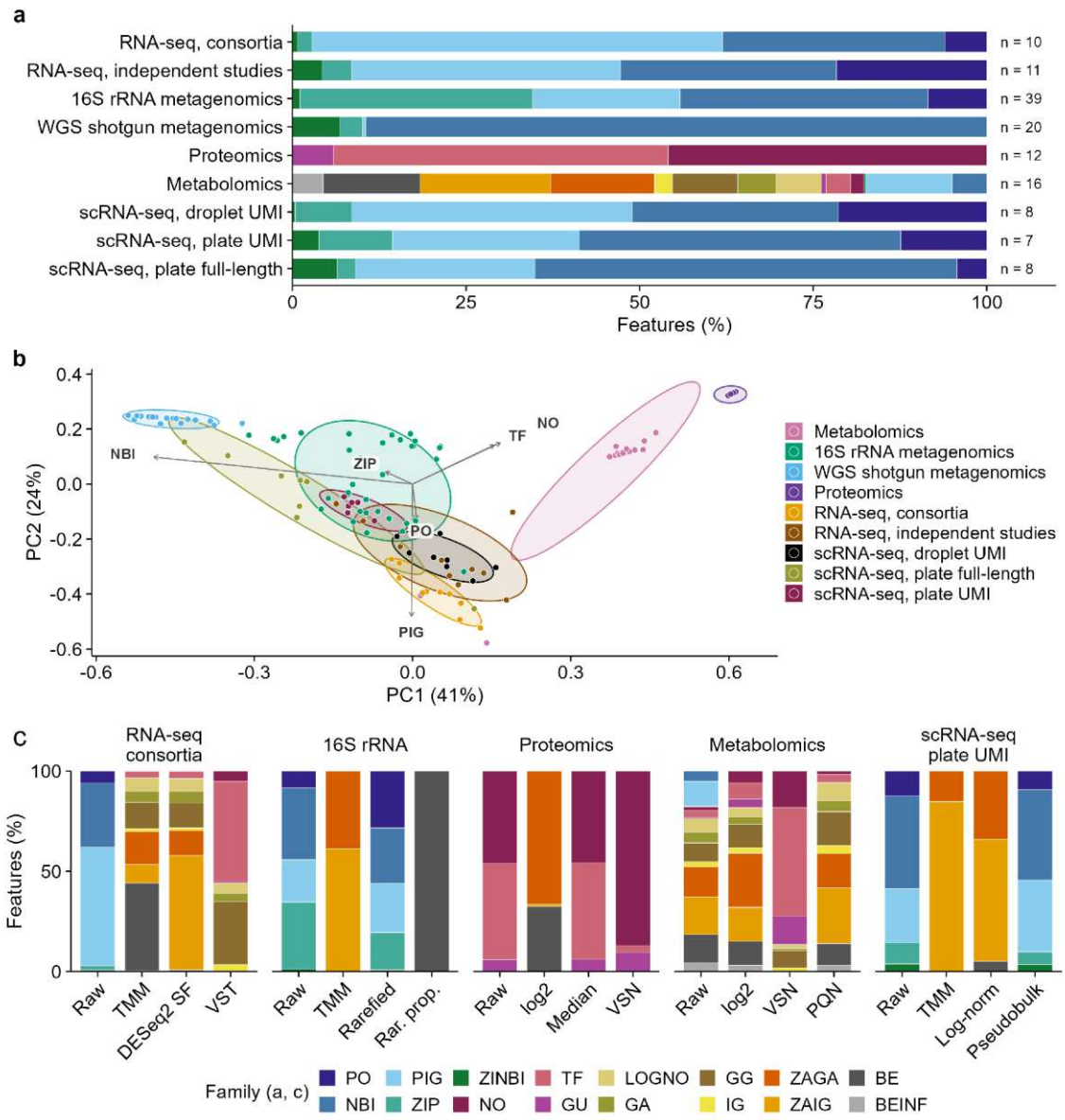


Fig. 1 | Distributional heterogeneity of public omics data. a, Families selected per feature at the raw scale, pooled over the datasets of each profile; *n*, number of datasets (studies). **b**, Principal component analysis of the 131 datasets described by their raw family compositions; arrows show the six families with the largest loadings; ellipses, 68% normal contours of each profile. **c**, Family composition after the normalizations used by the reference methods, for five representative profiles (all profiles in Supplementary Table 3). Panels **a** and **c** share the family colour key shown below **c**; family codes are defined in Supplementary Table 4.

Simulated gains beyond RNA-seq

We simulated the nine profiles from their raw family mixtures (Supplementary Table 5) with 80 samples per group, 30% of features with a low effect (fold change 1.5 or a shift of 0.5 standard deviations) and 5% with a high effect (fold change 3.0 or a shift of 1.5 standard deviations). Single-

cell profiles were simulated with a hierarchical donor-to-cell model, with 20 donors per group and 250 cells per donor, and analysed at the pseudobulk level. Each profile was simulated under five scenarios that each add one source of difficulty: a baseline with low dispersion; a noisy scenario with high dispersion, heavier tails and a stronger zero process; an outlier scenario in which 2% of the matrix entries were replaced by extreme values; a zero-inflation scenario with a stronger excess of zeros, applied to the seven profiles whose mixture contains a zero-inflated or zero-adjusted family; and a batch scenario with a two-level batch factor partially confounded with the group (60% of controls and 40% of cases in the first batch). PERSEO was compared with the reference methods of each layer, each run on the input its guidelines recommend (Methods). In the baseline scenario PERSEO had the highest F1-score in five of the six bulk profiles and tied with edgeR in consortium RNA-seq (0.926 each; Fig. 2a and Supplementary Table 6). The margin was largest for the continuous profiles, metabolomics (0.852 versus 0.755 for Welch's *t*-test) and proteomics (0.837 versus 0.769 for limma), and was 0.017–0.018 in metagenomics, where ANCOM-BC2 was the strongest competitor, and 0.006 in independent-study RNA-seq. In the three single-cell profiles DESeq2 had the highest F1-score, 0.020–0.058 above PERSEO.

Across the 43 profile-by-scenario combinations, PERSEO exceeded an observed false discovery rate (FDR) of 0.05 in nine, DESeq2 in 12 of 24, edgeR in 11 of 24 and ANCOM-BC2 in 7 of 10, whereas limma-voom, ALDEx2, LEfSe, MSstats and Welch's *t*-test never did and limma did so once (Fig. 2b). Eight of the nine exceedances of PERSEO occurred in metabolomics (four, with a maximum of 0.116 in the zero-inflation scenario) and in single-cell data (four, at most 0.065); the ninth was the baseline 16S profile (0.060). Values for every scenario are given in Supplementary Tables 6–10.

PERSEO performed best when it received depth information: in the baseline 16S profile its F1-score was 0.924 with the log size factor as a covariate, 0.901 on integer-rounded TMM counts, 0.880 on TMM-scaled counts and 0.367 on rarefied counts (0.934, 0.917, 0.910 and 0.436 in WGS metagenomics; Supplementary Table 11), consistent with the loss of information caused by rarefaction.[29]

Gains grow under stress scenarios

Stress increased the advantage of PERSEO outside consortium RNA-seq (Fig. 2a). Under added noise, the difference in F1-score from the best competitor reached 0.184 in metabolomics (0.683 versus 0.499), 0.152 in WGS metagenomics (0.774 versus 0.622 for ANCOM-BC2), 0.065 in proteomics and 0.049 in 16S metagenomics. Outlier contamination separated the methods most in proteomics (0.815 versus 0.683 for limma). In consortium RNA-seq the F1-scores were tied, but

the difference appeared in error control: under outliers edgeR reached an FDR of 0.103 and DESeq2 0.074, against 0.044 for PERSEO, at F1-scores of 0.761, 0.767 and 0.766.

The zero-inflation scenario, applied to the seven profiles whose mixture contains a zero-inflated or zero-adjusted family, reproduced the ordering in bulk profiles: PERSEO led in independent-study RNA-seq (0.850 versus 0.819), 16S metagenomics (0.842 versus 0.806, with ANCOM-BC2 at an FDR of 0.101), WGS metagenomics (0.776 versus 0.675) and metabolomics (0.729 versus 0.543), although in metabolomics at an FDR of 0.116. With a batch factor partially confounded with the group, PERSEO fitted with a batch term matched or exceeded the best covariate-aware competitor in every bulk profile, with differences in F1-score from 0.001 in consortium RNA-seq to 0.125 in metabolomics.

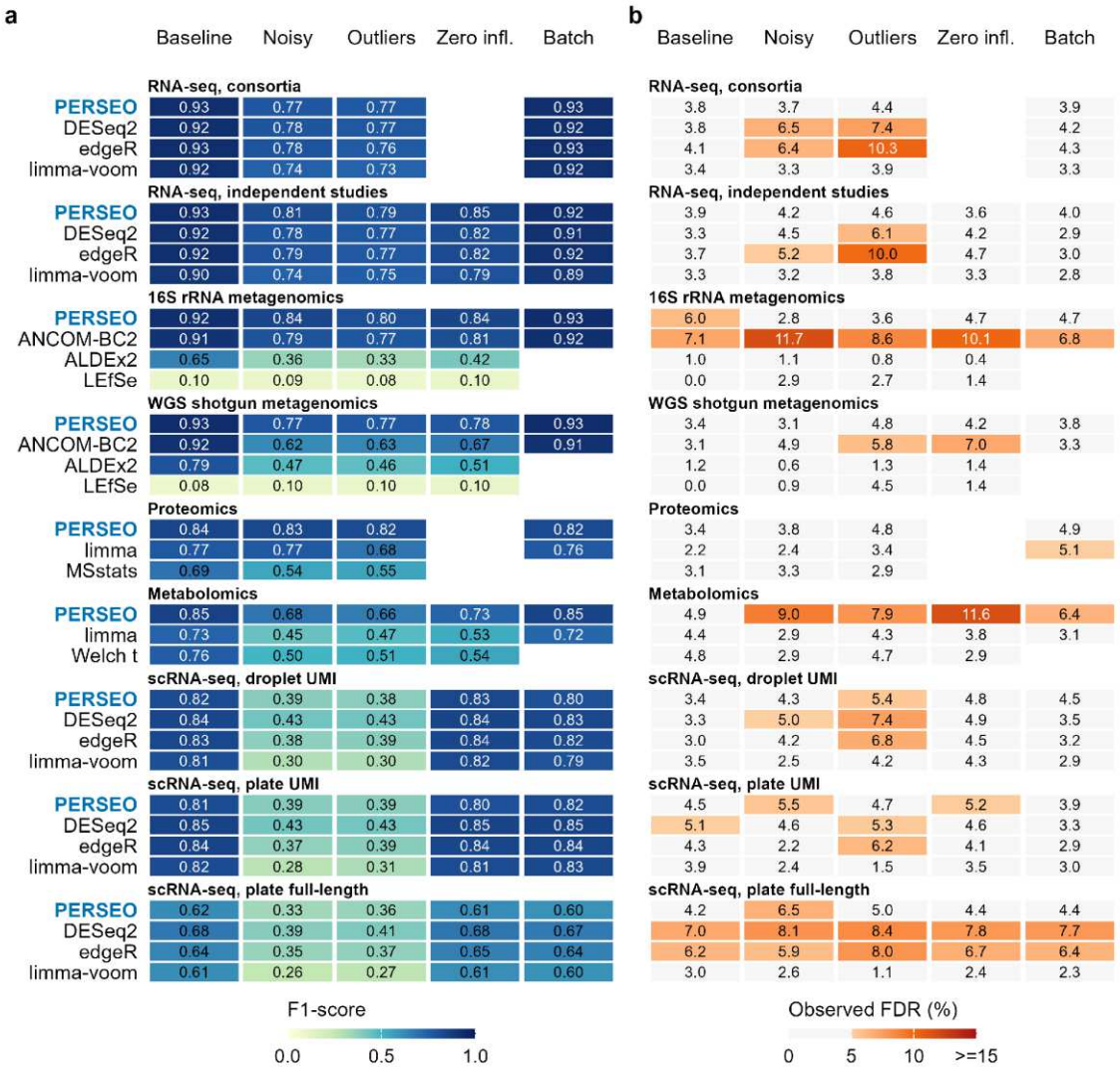


Fig. 2 | Performance on synthetic data across profiles and scenarios. a, F1-score at $q \leq 0.05$ for each method, profile and scenario. **b**, Observed FDR (%) of the same analyses; cells above the nominal 5% are coloured from orange to red. Single-cell profiles at the pseudobulk level; empty cells, scenario not simulated for the profile or method not run. Each profile and scenario was simulated once (80 samples per group; 20 donors per group with 250 cells each in single-cell profiles). Values are given in Supplementary Tables 6–10.

Performance stratified by underlying distribution family

To determine the origin of these performance differences, we evaluated detection sensitivity stratified by the true underlying distribution family of each feature (Fig. 3a). When features satisfied the distributional assumptions of canonical methods—such as NBI, PIG, and PO in bulk RNA-sequencing or LOGNO and BE in metabolomics—all tools yielded comparable sensitivities within a few percentage points. Pronounced performance gaps emerged specifically where these assumptions were violated. In independent-study RNA-seq, PERSEO detected 86% of ZIP and 55%

of ZINBI features, compared to 22–39% and 19–31% for DESeq2, edgeR, and limma-voom. In metabolomics, it identified 71% of ZAGA features, versus 7% for limma and 16% for Welch's t-test, while in proteomics it detected 66% of TF and 72% of Gumbel (GU) features, compared with 51% and 48% for limma and 39% and 36% for MSstats. In 16S metagenomics, PERSEO captured 85% of ZIP features compared with 4% for ALDEx2, whereas ANCOM-BC2 achieved 90%; similarly, ANCOM-BC2 proved more sensitive for ZINBI features in WGS metagenomics (65% versus 50%). In both metagenomic assays, however, the higher sensitivity of ANCOM-BC2 was accompanied by an inflated family-specific false discovery rate (9.6% and 14.2%, versus 5.5% and 3.5% for PERSEO; Fig. 3b). In contrast, ZAIG features in metabolomics remained largely undetectable across all methods (2–8% sensitivity). Finally, in single-cell pseudobulk data, where aggregation naturally shifts features into standard count families, PERSEO trailed DESeq2 by 3–6 percentage points across all families.

The family selected by PERSEO matched the generating family for 68–82% of features in the bulk profiles and for 33–49% in the single-cell profiles (Fig. 3c). In the real-data analyses below, features are labelled with the family selected by PERSEO, and the simulation was stratified in the same way so that both can be compared.

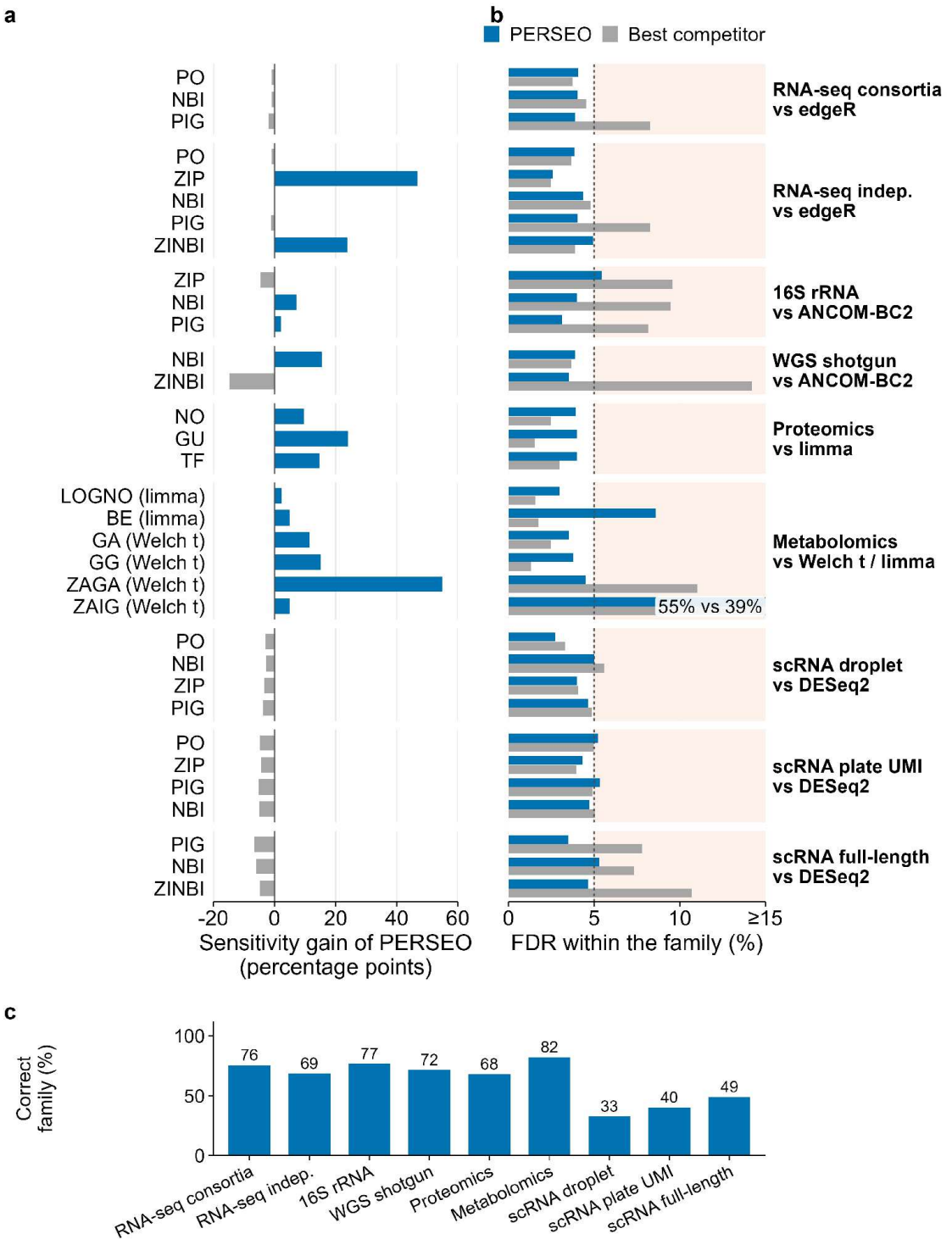


Fig. 3 | Sensitivity and false discoveries by generating family. a, Sensitivity of PERSEO minus that of the best competitor (the most sensitive reference method, given on the right of each profile, or next to the family where it differs) for each generating family; blue, PERSEO more sensitive; grey, competitor more sensitive. **b**, FDR within the family (false calls among the calls of the family) of PERSEO (blue) and of the best competitor (grey); dashed line and shading, FDR above the nominal 5%; values above 15% are given in the panel. Rows are grouped by profile; single-cell profiles at the

pseudobulk level. **c**, Fraction of features for which PERSEO selected the generating family in each profile. All panels pool the baseline, noisy, outlier and zero-inflation scenarios at $q \leq 0.05$; families are those of the simulated mixture of each profile with at least 50 differential features (Supplementary Table 5); LEfSe omitted. Values are provided as Source Data.

Pseudobulk controls single-cell FDR

Cell-level evaluations that did not account for donor-level variation led to severe false discovery rate (FDR) inflation across all single-cell profiles and scenarios: 0.35–0.56 for DESeq2, 0.36–0.54 for PERSEO, 0.36–0.56 for edgeR, 0.32–0.67 for limma-voom, and 0.34–0.53 for the Wilcoxon test (Supplementary Fig. 1a), in agreement with previously described pseudoreplication biases in single-cell differential expression.[30] Incorporating a donor random intercept via a negative binomial mixed model mitigated but failed to eliminate this inflation (0.11–0.32). In contrast, aggregating measurements into donor-level pseudobulk profiles fully restored FDR control, yielding nominal error rates across methods (0.034–0.065 for PERSEO, 0.033–0.084 for DESeq2, 0.022–0.080 for edgeR, and 0.011–0.043 for limma-voom).

Under pseudobulk aggregation, scenarios introducing high noise or outlier contamination at the donor level reduced overall F1-scores to 0.26–0.43, with DESeq2 achieving the highest F1-score across all 15 profile-by-scenario combinations (Supplementary Fig. 1b). Notably, pseudobulk aggregation condensed feature distributions such that PERSEO selected either negative binomial (NBI) or Poisson–inverse Gaussian (PIG) families for 95–100% of genes, regardless of their simulated generative distribution (Supplementary Fig. 1c). This distributional convergence largely eliminates the feature-wise heterogeneity observed at the single-cell level, leaving limited opportunity for adaptive family selection to confer an advantage over conventional count-based models.

Agreement and calibration on real data

We next compared the methods on the 12 real datasets (effective sample size 4.8–288; Fig. 4a). Without ground truth, we measured agreement, false discoveries after a balanced permutation of the group labels and the correspondence with simulation. The Spearman correlation between the *P* values of PERSEO and those of the competitors depended more on the layer than on sample size (Fig. 4b and Supplementary Fig. 2): it was 0.98–0.99 in proteomics, 0.75–0.84 in TCGA-BRCA and 0.71–0.73 in myeloma single-cell data, 0.26–0.62 in metagenomics and 0.20–0.23 in prostate metabolomics.

With permuted labels, every call is a false discovery (Fig. 4c). ANCOM-BC2 kept most of its calls in the metagenomic datasets (74 of 111 in *Clostridioides difficile* infection, 510 with permuted against 497 with true labels in colorectal cancer, 16 of 45 in schizophrenia and 69 of 145 in cirrhosis), and LEfSe made 31 calls with permuted labels against 37 with true labels in colorectal cancer. In TCGA-BRCA, DESeq2 and edgeR made 1,465 and 1,897 calls with permuted labels (6.7% and 8.7% of their 21,731 and 21,777 calls with true labels), in line with previous reports of false discoveries by these methods in large human cohorts,[31] whereas PERSEO made 220 (1.1% of 19,677) and limma-voom none. PERSEO made more calls with permuted than with true labels in prostate metabolomics (207 versus 120).

Across the 66 method-by-family combinations with at least 20 calls, the FDR observed in simulation correlated with the permutation FDR on real data (Spearman's ρ = 0.41; Fig. 4d). The agreement was close for DESeq2 and edgeR in PIG genes of TCGA-BRCA (simulated 6.9% and 8.3%, permutation 7.9% and 10.5%), whereas the simulation underestimated the false discoveries of ANCOM-BC2 in metagenomics (simulated 3–10%, permutation 38% or more) and did not anticipate those of PERSEO in the ZAGA and GA families of prostate metabolomics.

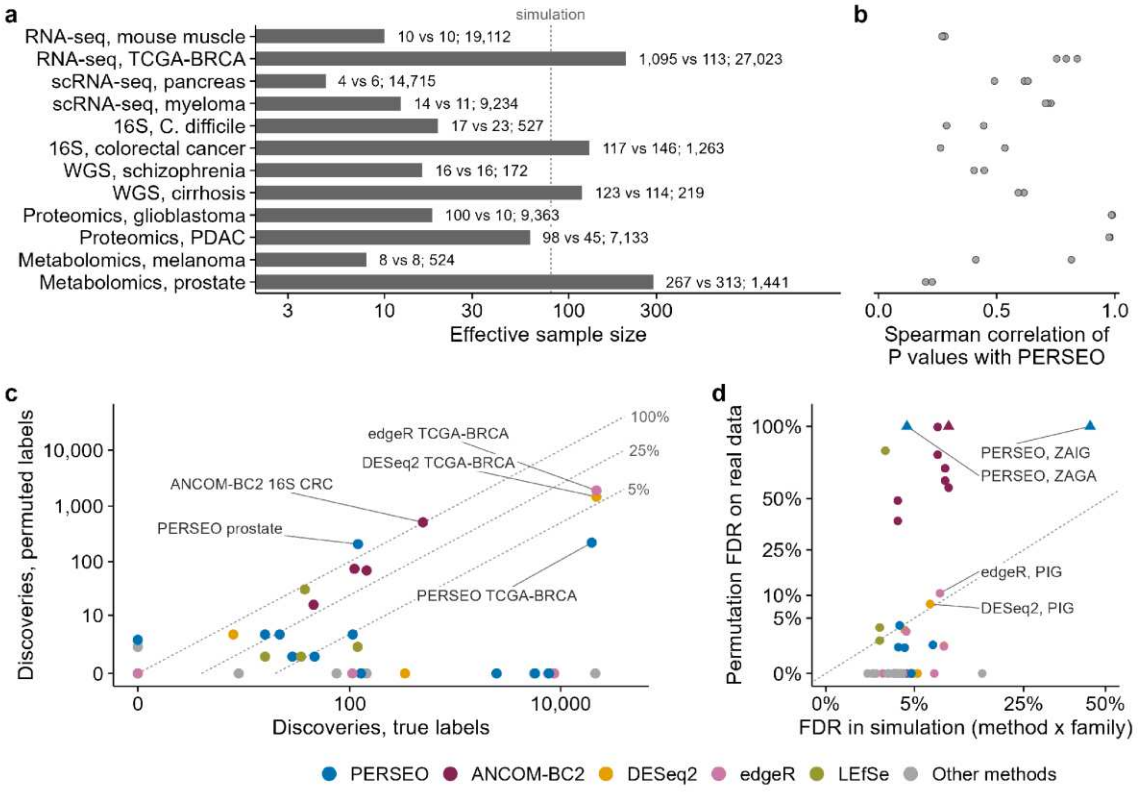


Fig. 4 | Concordance and calibration on 12 public datasets. a, Effective sample size (bars; $2n_0n_1/(n_0 + n_1)$) of the 12 datasets; labels give case versus control samples and features tested; the dashed line marks the simulated design (80

samples per group). **b**, Spearman correlation between the *P* values of PERSEO and those of each competitor (one point per competitor); rows are aligned with **a**. **c**, Calls with true labels against calls after a balanced permutation of the labels, where every call is false, for each method and dataset; dashed lines mark permutation FDRs of 5%, 25% and 100%. **d**, FDR observed in simulation against the permutation FDR on the matched real dataset, for each method and PERSEO-selected family with at least 20 calls with true labels; triangles mark values capped at 100%; labelled points are PERSEO in prostate metabolomics and DESeq2 and edgeR in TCGA-BRCA; dashed line, identity.

Reading disagreements via simulation

In simulation, features called by both PERSEO and a competitor were true positives in 96–99% of cases, and features called by PERSEO only were true more often than those called by the competitor only in every profile except the single-cell ones (Fig. 5a). By PERSEO-selected family, PERSEO-only calls were the more reliable in 26 of 40 family-by-competitor pairs (Fig. 5b). We used these fractions as an *a priori* reading of the real-data disagreements and checked each reading against the permutation (Fig. 6).

The reading identified probable false negatives of the competitors. In proteomics, 160–207 proteins per dataset were called by PERSEO but not by limma or MSstats, mostly TF features for which the simulated fraction of true positives was 88–94%; in metagenomics, the features called by PERSEO but not by ALDEx2 or LEfSe had simulated fractions of 92–96%; and in TCGA-BRCA, 1,247–1,364 genes, mostly PIG, were called by PERSEO but not by DESeq2 or edgeR (87%). Examples are SCAPER, a TF protein in pancreatic ductal adenocarcinoma ($q = 4.6 \times 10^{-6}$ for PERSEO and 0.47 for limma), and *Romboutsia ilealis* in cirrhosis, detected in 29% and 17% of control and cirrhosis samples ($q = 0.0063$ for PERSEO, 0.94 for ANCOM-BC2 and 0.16 for ALDEx2).

The reading also identified probable false positives of the competitors. The 100 and 472 operational taxonomic units called by ANCOM-BC2 alone in the 16S datasets had a simulated fraction of true positives of 39–47%, and ANCOM-BC2 did not withstand the permutation in these families. In WGS metagenomics the simulation assigned 65% to the NBI features called by ANCOM-BC2 alone, but the permutation contradicted it. The roughly 3,400 genes called by DESeq2 or edgeR but not by PERSEO in TCGA-BRCA were mixed (54–61% in PIG, about 1,300–1,500 expected false positives), the same order as the 1,307 and 1,750 PIG genes these methods called with permuted labels. In single-cell myeloma data, by contrast, the genes called by DESeq2, edgeR or limma-voom but not by PERSEO were probable true positives (84–91%), that is, false negatives of PERSEO.

The same reading identified the calls of PERSEO with low expected precision. In prostate metabolomics, 73 of its 120 calls were ZAIG metabolites, for which the simulation predicted about

half false positives, and the permutation produced 136 calls in this family; nicotinate, absent in 82% and 81% of control and cancer samples, received $q = 1.4 \times 10^{-185}$ and was also called with permuted labels. The remaining 47 calls belonged to families classified as reliable by the simulation that did not withstand the permutation.

Two conditions absent from the simulation affected PERSEO on real data (Supplementary Fig. 3). In the mouse muscle RNA-seq study (SRP074887), sequencing depth followed the group ($R^2 = 0.85$ for log size factor on group). Because PERSEO receives the log size factor as a covariate rather than as an offset, the covariate absorbed part of the group effect: the correlation between the effect estimates of PERSEO and DESeq2 was 0.50, against 0.85 in TCGA-BRCA, 18% of PERSEO calls had the opposite sign to DESeq2 (5% in TCGA-BRCA), and PERSEO made 2,478 calls against 8,614–8,701 for the competitors, none of which made calls with permuted labels. In prostate metabolomics, with 580 samples, PERSEO called with permuted labels 18% of ZAIG, 14% of ZAGA and 22% of GA metabolites, against 0.1% of all metabolites for limma.

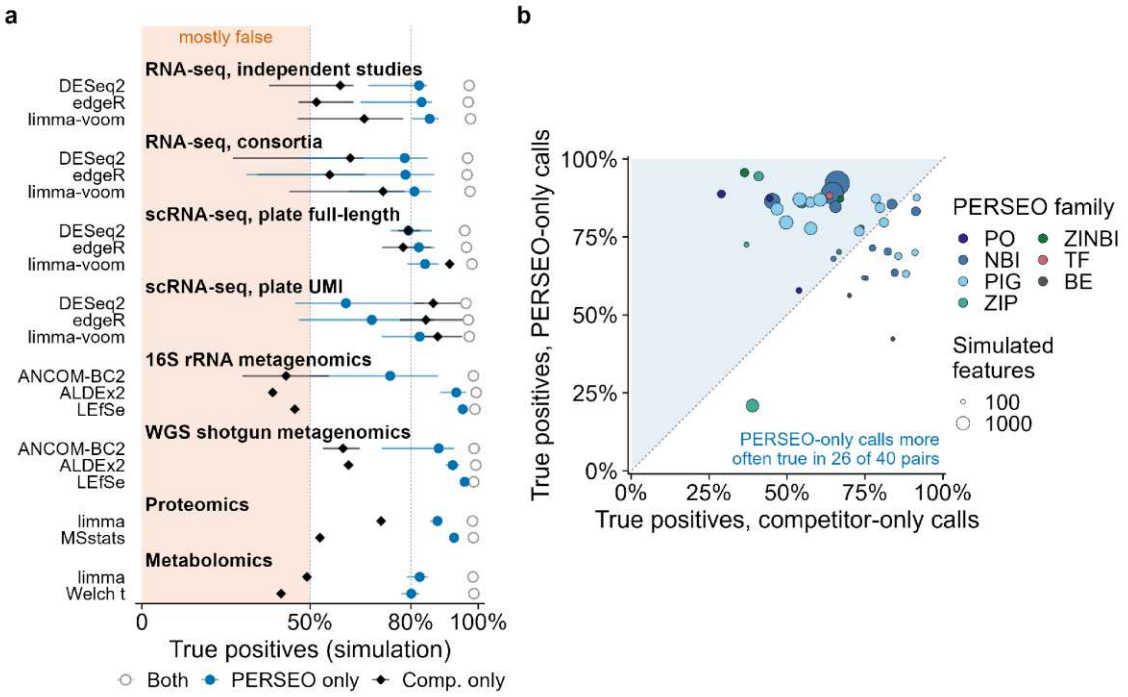


Fig. 5 | The simulation key used to read disagreements between methods. a, Fraction of true positives in simulation among features called by both methods (open circles), by PERSEO only (blue) and by the competitor only (black), for each simulated profile and competitor. Points pool the baseline, noisy, outlier and zero-inflation scenarios; horizontal lines span the individual scenarios with at least 20 features; the shaded area marks sets that are mostly false. **b**, The same comparison by PERSEO-selected family: fraction of true positives among PERSEO-only calls against competitor-

only calls (pairs with at least 20 simulated features in both sets); point area is proportional to the number of simulated features.

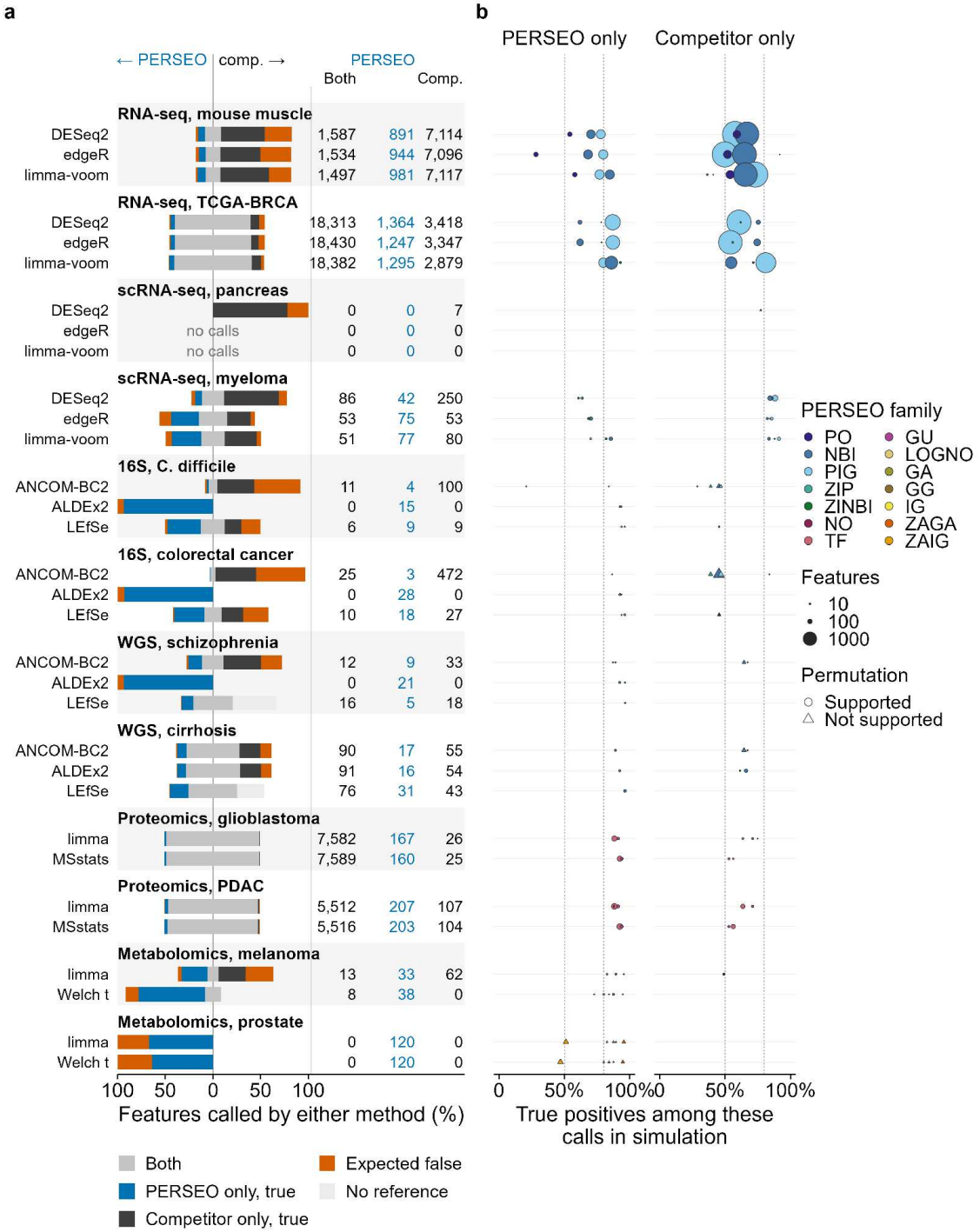


Fig. 6 | Disagreements between PERSEO and the reference methods on 12 public datasets, read through the simulation. a, Features called by both methods (grey, centred), by PERSEO only (left) or by the competitor only (right), as a percentage of the features called by either method. Each exclusive set is split into expected true positives (blue for

PERSEO, dark grey for the competitor) and expected false positives (orange), using the fraction of true positives observed for the same set and PERSEO family in simulation; light grey marks features without a simulation reference. Counts are given on the right. **b**, Fraction of true positives in simulation for each exclusive set of **a**, by PERSEO family (colour); point area is proportional to the number of real features. Triangles mark sets whose calls do not withstand the permutation (at least three calls with permuted labels and more than 25% of the calls with true labels). Rows are aligned with **a**.

Discussion

Real omics data are distributionally heterogeneous. Across 131 public datasets no profile was described by a single family, at the raw scale or after the normalizations of each field, extending descriptive surveys of omics data[18] to the level of the probability law of each feature. The normalizations prescribed by the reference methods relocated the families without producing the family each test assumes: variance-stabilized counts were better described by Student's *t* than by a Gaussian law, and depth-normalized counts by zero-adjusted or beta families.

Modelling each feature under its own family improved detection where the fixed assumption fails and cost nothing where it holds. When PERSEO selects the negative binomial for a feature, the GAMLSS model reduces to a parametrization equivalent to that of DESeq2 or edgeR under the same link,[28,32] and the methods tied on bulk RNA-sequencing counts. The gains concentrated in zero-inflated counts of independent RNA-seq studies, zero-adjusted metabolites and heavy-tailed or skewed proteins, where fitting a Gaussian or negative binomial model inflates the residual variance and attenuates the test statistic. Crucially, whether these observed zeros reflect true biological absence or technical artifacts—such as capture inefficiency, dropouts, or detection limits—is strictly contingent on the upstream experimental and technological pipeline. Regardless of their mechanistic origin, such zeros fundamentally alter the data structure and must be treated with dedicated statistical formulations to ensure valid differential inference. By targeting the whole conditional distribution, distributional regression properly accommodates these zero processes while registering differences expressed through dispersion, tail behaviour, or zero inflation itself.[33]

Error control distinguished the methods as much as detection. DESeq2 and edgeR exceeded the nominal FDR under outliers in simulation and made 1,465 and 1,897 false discoveries in a permuted TCGA-BRCA cohort, consistent with the exaggerated false positives reported for these methods in population-scale data.[31] ANCOM-BC2 kept most of its calls after permutation in all four metagenomic datasets, in line with benchmarks in which several differential abundance methods failed to control false discoveries[17,34] and with the disagreement between methods

observed across microbiome datasets.[19] ALDEx2 and LEfSe, conversely, controlled false discoveries at the cost of detecting few features.

In single-cell data, the unit of analysis mattered more than the distributional model. At the cell level every method that ignored donors inflated the FDR, as described previously,[30] whereas at the pseudobulk level, the appropriate unit for between-subject inference, the FDR of all methods remained close to the nominal level (at most 0.084). Aggregation left mainly negative binomial and Poisson–inverse Gaussian features, for which PERSEO fits models close to those of DESeq2 and edgeR, and the methods differed by at most 0.07 in F1-score. This agrees with the absence of excess zeros in UMI-based data beyond those expected from low means,[16] although whether zeros in sequencing data are primarily technical or biological remains debated.[27,35]

Reading real-data disagreements through the simulation and checking each reading against label permutation provides an interpretable way to decide which calls to trust when methods disagree. Simulated and permutation FDRs correlated across method-by-family combinations, and the reading identified probable false negatives of the competitors in proteomics, metagenomics and TCGA-BRCA, as well as probable false positives of ANCOM-BC2 in the 16S datasets.

The study has limitations. The simulation drew data from the GAMLSS families among which PERSEO selects, which favours PERSEO by construction, although the mixtures were estimated from 131 real datasets; it used a single replicate per scenario, and the real-data benchmark, with one permutation per dataset, provides the test without that advantage. The simulation underestimated the false discoveries of ANCOM-BC2 and did not include group-dependent sequencing depth, which PERSEO absorbs through its depth covariate, or large zero-rich metabolomic cohorts, in which zero-adjusted fits called features with permuted labels. The *P* values of PERSEO are conditional on the selected family, and an implementation error affected MSstats in synthetic proteomics (Methods). PERSEO is slower than the reference methods (11 h for TCGA-BRCA with 20 workers, against 31–294 s), and its performance depends on depth information, which rarefaction removes.[29] Support for an offset, a minimum number of non-zero values per group before fitting zero-adjusted families, replicate simulations and signal implantation into real profiles[34] are natural next steps. Nevertheless, despite these operational and computational trade-offs, PERSEO consistently outperforms canonical tools across heterogeneous omics layers, providing a principled, distribution-adaptive framework that substantially reduces false discoveries and recovers genuine biological signals that fixed-model methods systematically miss.

In summary, PERSEO provides a single, principled framework for feature-wise differential analysis across count, continuous, and zero-adjusted omics data. By tailoring models to individual feature behavior, it equals specialized reference methods within their intended domains while systematically recovering biologically relevant signals that fixed-distribution models miss.

Methods

Public datasets

We characterized 135 public datasets spanning bulk and single-cell transcriptomics, proteomics, metabolomics and metagenomics. Although noise, excess zeros and heavy tails are widely recognized in omics data, a systematic quantification of the probability families that govern individual features across modalities has, to our knowledge, not been reported; previous work described sparsity, skewness, kurtosis and missingness.[18] A dataset is one public study, that is, one feature-by-sample matrix with its metadata. The corpus comprised 21 bulk RNA-sequencing datasets (10 consortium cohorts from GTEx and TCGA and 11 independent studies from the SRA), 23 single-cell RNA-sequencing datasets (8 droplet-based with unique molecular identifiers (UMI), 7 plate-based with UMI and 8 plate-based full-length), 12 CPTAC proteomes, 20 Metabolomics Workbench studies, 39 16S rRNA amplicon datasets and 20 shotgun metagenomes (Supplementary Table 1). Family assignments at the raw scale were available for 131 datasets; for four metabolomic studies only the normalized conditions were available. The two metagenomic assays (16S and WGS metagenomics), the two bulk transcriptomic sources (consortium and independent-study RNA-seq) and the three single-cell protocols were analysed as separate profiles throughout.

Bulk transcriptomes were obtained as raw gene-level read counts through recount3[36] and TCGAbiolinks,[37] and single-cell count matrices through the scRNAseq Bioconductor package.[38] CPTAC proteomes were supplied as log2 tandem-mass-tag ratios normalized at source,[39] so the unprocessed proteomic condition is itself a pre-normalized matrix. Metabolomic intensities were retrieved from Metabolomics Workbench,[40] 16S rRNA amplicon count tables from the collection of Nearing et al.,[19] and shotgun species-level MetaPhlAn abundances from curatedMetagenomicData,[41] rescaled to a common total of 10^6 per sample and rounded to integer pseudo-counts.

All outputs were converted to a features-by-samples representation paired one-to-one with sample metadata. Samples were matched strictly to metadata, unmatched entities were discarded, and features uniformly zero or uniformly missing were removed. For every dataset all applicable normalizations were analysed: for RNA-sequencing, unprocessed counts, the trimmed mean of M-values (TMM),[42] DESeq2 size factors[1,2] and the variance-stabilizing transformation (VST); for proteomics, unprocessed values, log2(x + 1), per-sample median centring and variance-stabilizing normalization (VSN);[6] for metabolomics, unprocessed values, log2(x + 1), probabilistic

quotient normalization (PQN)[43] and VSN; for the metagenomic assays, unprocessed counts, rarefied counts, rarefied proportions and TMM;[29] and for single-cell data, unprocessed cell counts, log-normalized counts (counts scaled to 10,000 per cell and log-transformed), TMM-scaled counts and pseudobulk profiles summed per donor and cell type. Single-cell datasets with more than 3,000 cells were subsampled to 3,000 cells, stratified by annotated cell type, donor or batch, with a fixed seed.

Datasets of the real-data benchmark

To test whether the behaviour observed in simulation holds in real data, PERSEO was compared with the same reference methods on 12 public datasets from the characterization corpus, two per omics layer (Supplementary Table 2). Datasets were required to provide a two-group contrast with accessible metadata, one sample per subject and a scale consistent with the layer. Because group imbalance limits power through the smaller group, size was measured as the effective sample size, $n_{\mathrm{eff}} = 2n_0n_1/(n_0 + n_1)$, and for each layer the two datasets at the extremes of the available range were selected. Real data carry no ground truth, so the benchmark measured agreement between methods, false discoveries under label permutation, and the correspondence between the behaviour of each method on real data and in simulation.

The bulk RNA-sequencing datasets were quadriceps muscle of R6/2 and wild-type mice from the vehicle arm of a Huntington's disease study (SRP074887),[44] retrieved through recount3, and primary breast tumours against solid tissue normal samples from TCGA-BRCA,[45] retaining one primary tumour aliquot per patient. The single-cell datasets were pancreatic alpha cells from donors with type 2 diabetes and healthy donors[46] and bone marrow plasma cells from patients with multiple myeloma and controls,[47] both obtained through the scRNAseq package and analyzed as pseudobulk profiles that sum all cells of a donor and cell type; donors with fewer than ten cells of the cell type were excluded. The 16S datasets compared *Clostridioides difficile* infection with healthy controls[48] and colorectal cancer with healthy controls,[49] from the collection of Nearing et al..[19] The WGS metagenomic datasets compared the oropharyngeal metagenome of individuals with schizophrenia and controls[50] and the gut metagenome of patients with liver cirrhosis and controls,[51] retrieved from curatedMetagenomicData with phenotypes from its curation tables and one sample per subject. The proteomic datasets compared glioblastoma and pancreatic ductal adenocarcinoma with normal tissue from CPTAC,[52,53] with one sample per case. The metabolomic datasets were a lipidomic profile of A375 melanoma cells treated for 24 h with ranolazine or vehicle (Metabolomics Workbench ST003945) and plasma metabolomes of men with and without a diagnosis of prostate cancer (ST002498).

The PERSEO method

PERSEO is an R package for feature-wise modelling with generalized additive models for location, scale and shape (GAMLSS), built on the GAMLSS framework and its R implementation.[28,32] It was evaluated in three stages: an empirical characterization of the distribution families of public omics data, a synthetic benchmark grounded in that characterization, and a benchmark on real datasets. Estimation formulas are summarized in Supplementary Table 12 and family codes in Supplementary Table 4.

Data integration, support detection and common mask

PERSEO operates on a feature matrix *Y* (*p* features by *n* samples) and a design matrix *X* (*n* samples by *q* covariates), with a one-to-one correspondence between the columns of *Y* and the rows of *X*. It requires no specific normalization and handles raw data or data normalized at the experiment level. For each feature PERSEO can detect its empirical support (non-negative integers, the unit interval, positive continuous or real values) and restrict model comparison to families consistent with that support; when this option is disabled, the data are adapted to the domain of each candidate family. A common mask of valid observations is built for each feature, excluding missing or out-of-domain values, so that all information criteria are computed on the same data.

Variable transformation and Jacobian correction

Some families require an intermediate scale $z = g(y)$ to match their domain. PERSEO applies reversible transformations and records, for each observation, the term $\log|g'(y)|$, so that the log-likelihood used for model selection adds this Jacobian term to the log-density on the transformed scale. AIC, BIC and GAIC values are therefore comparable on the original scale, and no family is favoured by the transformation.

$$logL(\theta;\, y) \;=\; logL(\theta;\, z) \;+\; \sum_{i} \quad log\left|\frac{\partial z_i}{\partial y_i}\right|$$

Family screening by bootstrap

To reduce the search space, PERSEO first screens families on random subsets of features, with the number of features sampled and the number of bootstrap iterations as parameters. For each sampled feature, models are fitted across the families compatible with its support and compared by information criteria and Jacobian-corrected log-likelihoods. The output is a set of frequently selected families, globally or by support, that defines the candidate pool of the fitting step.

Feature-wise fitting and model selection

PERSEO fits a GAMLSS model for each feature and candidate family, modelling the location parameter as a function of the design matrix. All models of a feature are fitted on its common mask, and the family is selected by an information criterion (by default GAIC with penalty k = log n, equivalent to BIC). The effective sample size, the Jacobian-corrected log-likelihood and the criterion value are retained. Users may restrict or predefine the candidate families.

$$IC \;=\; -2\left[logL(\hat{\theta};\, z) + \sum_{i} \quad log|J_i| \right] + k\, df, \qquad k \in \{\, 2,\; logn,\; k_{GAIC} \,\}$$

Contrast testing and multiple-testing correction

Once the family is selected, PERSEO tests the location coefficients with Wald tests.[54] For two-level factors the coefficient of interest is tested directly with a t statistic on the residual degrees of freedom. For factors with three or more levels, an optional hierarchical strategy first applies an omnibus test (a multi-degree-of-freedom Wald test or a likelihood-ratio test of the full against the reduced model), and only features passing it proceed to pairwise contrasts, computed as linear combinations of the coefficients with standard errors obtained by variance propagation. *P* values are adjusted across features for each term and contrast, by default with the Benjamini–Hochberg procedure.[55]

$$\hat{V} = (X^{\top} W X)^{-1}, \qquad t_j = \frac{\hat{\beta}_j}{\sqrt{\hat{V}_{jj}}}$$

$$\hat{\delta} = C\hat{\beta}, \qquad Var(\hat{\delta}) = diag\big(C\,\hat{V}\,C^{\top}\big), \qquad z = \hat{\delta} \,/\, se(\hat{\delta})$$

$$W = \hat{\beta}_S^{\top}\, \hat{V}_S^{-1}\, \hat{\beta}_S \;\sim\; \chi^2(df)$$

The arguments of run_perseo() control the bootstrap screening (enabled by default; number of features sampled and number of iterations), the number of dominant families carried to the fitting step or an explicit candidate set, the restriction of candidates to the inferred support, the selection criterion (GAIC with k = log n by default, with AIC and BIC as alternatives), the multiple-testing method, the optional omnibus gate and the evidence thresholds for zero and one inflation. By design only the location parameter depends on the covariates; the scale and shape parameters are fitted as intercepts.

Output report

PERSEO writes a self-contained interactive HTML report that records the configuration, the number of features tested and called, the distribution of selected families, a searchable table of results exportable to CSV or Excel, and, for each comparison, volcano and MA plots, boxplots of the top features and an UpSet diagram across contrasts.

Characterization of distribution families

For each dataset and normalization, every feature was fitted with PERSEO across a panel of GAMLSS families and the family was selected by BIC. The panel comprised the discrete families PO, NBI and PIG and their zero-inflated forms (ZIP and ZINBI), the continuous families NO, TF, LOGNO, GA, GG, IG, GU, BE and BEINF, and the zero-adjusted families ZAGA and ZAIG. Assignments were stabilized by bootstrap resampling (three iterations; 2% of features per iteration, at least ten) with support-grouped aggregation and a strict transformation mode. The workflow ran in R on Bioconductor[56] with fixed seeds.

Synthetic benchmark

PERSEO was evaluated on synthetic data whose generative structure was taken from the characterization. Raw-scale data were simulated for each omics profile and then normalized with the procedures applied to the corresponding real data, so that the families observed after normalization arose from the transformation rather than being imposed. Two properties make the design valid. The generating family of a feature was drawn independently of its differential status, so the family structure carries no information about the truth. Library size varied between samples, as in real data, so depth normalization had real variation to correct.

Nine profiles were simulated: consortium RNA-seq (GTEx and TCGA), independent-study RNA-seq (SRA), 16S and WGS metagenomics, proteomics, metabolomics, and single-cell RNA-sequencing from droplet UMI, plate UMI and plate full-length protocols. For each profile the mean proportion of every family across its datasets was computed at the raw scale, families below 5% were discarded and the remainder rescaled to sum to one (Supplementary Table 5). The zero-inflated Poisson family of the independent-study RNA-seq profile (0.042) was retained below the threshold to preserve the zero inflation of that profile. The proteomic mixture was estimated on the 12 CPTAC proteomes and the metabolomic mixture on 14 of the 16 metabolomic studies, after excluding two quantified on an integer scale whose discrete families would otherwise have entered a continuous profile. Depth heterogeneity, expressed as the standard deviation of log library size, was set for each count profile to the median measured across its datasets

(Supplementary Table 5). WGS metagenomic matrices are rescaled to a constant total, so their depth heterogeneity could not be measured and was set to 0.35.

Within each profile, 30% of features were assigned a low effect, 5% a high effect and the remaining 65% no effect. Differential features were split equally between up- and down-regulation, and each feature received a generating family sampled from the profile mixture independently of its class. Baseline means were drawn per feature, log-uniformly between 0.5 and 800 for count families and uniformly within family-specific ranges for continuous families (Supplementary Table 13). The effect was applied on the natural scale of each family: multiplicatively for count and positive families, additively on the log scale for the log-normal family, additively on the logit scale for unit-interval families and as a shift in standard-deviation units for real-valued families. The low effect was a fold change of 1.5 or a shift of 0.5 standard deviations, and the high effect a fold change of 3.0 or a shift of 1.5 standard deviations. Count features were multiplied by a per-sample log-normal depth factor. Each bulk dataset contained 80 samples per group and 12,000 features for the sequencing profiles or 2,000 for proteomics and metabolomics.

The three single-cell profiles were simulated with a hierarchical model in which cells are nested in donors and the condition label belongs to the donor. Each dataset contained 20 donors per group, 250 cells per donor representing a single cell type, and 5,000 features. The mean of feature g in cell c of donor d was $\mu_{gc} = \mu_{0g} \cdot \mathrm{FC}_g^{z_d} \cdot \gamma_{gd} \cdot \ell_c \cdot k$, where z_d indicates the treated group, γ_{gd} is a donor effect shared by all cells of a donor, ℓ_c is a per-cell log-normal depth factor with the standard deviation measured for each protocol, and k is a multiplier calibrated by bisection so that the fraction of zeros matched the value measured in real data after the standard pseudobulk filter of muscat[57] (Supplementary Table 5). The donor effect of count families was drawn from a gamma distribution with unit mean and dispersion φ_g, sampled per feature from a scenario-specific range. The cell-level dispersion of the negative binomial, zero-inflated negative binomial and Poisson–inverse Gaussian families was drawn from the interquartile range measured on real data for each protocol. Ten per cent of the features, drawn among those without effect, were simulated as sporadically detected genes with low baseline means, a class that passes standard filters but is detected in few cells. Pseudobulk profiles were obtained by summing the counts of each donor.[30]

Five scenarios each applied a single stressor while the effect structure, sample size and number of features were held fixed. Stress acted through per-feature dispersion, tail and zero-process parameters drawn from ranges that widen with the stress level (Supplementary Table 13). The

baseline scenario used low dispersion. The noisy scenario used high dispersion, combining overdispersion, heavier tails and a stronger zero process. The outlier scenario used high dispersion and replaced 2% of the matrix entries with spikes: additive shifts of four to eight row standard deviations with random sign for real-valued families, and multiplication by a log-normal factor with a log-scale standard deviation of 1.2 for count and positive families. The zero-inflation scenario sharpened the zero process (for example, 40–65% structural zeros for the zero-inflated Poisson family and 30–55% for the zero-adjusted families) and was applied to the seven profiles whose mixture contains a zero-inflated or zero-adjusted family: 16S and WGS metagenomics, metabolomics, independent-study RNA-seq and the three single-cell profiles. The batch scenario used low dispersion and added a two-level batch factor partially confounded with the group, with 60% of the control samples and 40% of the treated samples in the first batch, and a batch effect of magnitude 1.5 on the natural scale of each family. In the single-cell profiles the stressors acted at the donor level: the dispersion of the donor effect was 0.05–0.25 in the baseline, zero-inflation and batch scenarios and 0.25–0.80 in the noisy and outlier scenarios, outliers multiplied 2% of the feature-by-donor means by the same log-normal factor, and the batch was assigned to donors.

Each competing method received the input recommended by its own guidelines. In the bulk sequencing profiles, DESeq2[2] used poscounts size factors and the Wald test; edgeR[2] used TMM normalization and the likelihood-ratio test; and limma-voom[4] used TMM normalization, voom and empirical Bayes moderation, all on raw counts. ANCOM-BC2,[10] without a prevalence filter, and ALDEx2,[12] with 64 Monte Carlo instances, the centred log-ratio transformation against all features and Welch's *t*-test, received raw metagenomic counts. LEfSe,[58] run through its R implementation lefser,[59] received rarefied proportions rescaled to relative abundances. In proteomics, MSstats[7] received log2(x + 1)-transformed intensities, to which its dataProcess function applied a second log2 transformation, with one feature per protein and no normalization, and limma[8] the variance-stabilized matrix (VSN).[6] In metabolomics, Welch's *t*-test[9] was applied on the log2 scale and limma on the VSN matrix. PERSEO received only depth normalizations and never a variance-stabilizing or logarithmic transformation: raw counts with the logarithm of the DESeq2 poscounts size factor as a nuisance covariate, because PERSEO does not support an offset, and, as additional inputs, TMM-scaled and integer-rounded TMM counts in the sequencing profiles and rarefied counts in the metagenomic profiles. On the continuous profiles PERSEO received raw intensities. In the batch scenario only methods able to fit a group-plus-batch model were run (DESeq2, edgeR, limma-voom, limma, ANCOM-BC2 and PERSEO).

The single-cell profiles were analyzed at two levels. At the pseudobulk level, PERSEO, DESeq2, edgeR and limma-voom were applied to the donor-summed counts, with edgeR using the quasi-likelihood F-test.[60] At the cell level, the same four methods were applied to raw cell counts, DESeq2 with the settings recommended for single-cell data (likelihood-ratio test with *t*-distributed statistics, a minimum mean of 10^{-6} and no outlier replacement). Two cell-level references were added: a negative binomial mixed model with a donor random intercept and a log library-size offset, fitted with glmmTMB,[61] and the Wilcoxon rank-sum test[62] on log-normalized counts. MAST was not run because it was not available on the computing system.

PERSEO (version 1.0.0) was run with BIC family selection, bootstrap family screening on 5% of the features (at least 50) over five iterations, retention of the eight most frequent families as the candidate pool, support-grouped screening, the strict transformation mode and Benjamini–Hochberg adjustment. The test of interest was the group coefficient of the location model.

Every output was reduced to a binary call at a Benjamini–Hochberg-adjusted threshold of $q \leq 0.05$,[55] recomputed within each dataset over all features; features without a *P* value were counted as not called, and LEfSe calls were used as returned. From the confusion counts we computed sensitivity, specificity, precision, the observed false discovery rate (FDR), the F1-score and the Matthews correlation coefficient, overall, by generating family and by effect level. Each scenario and profile was simulated once with a fixed seed; no replicate simulations were performed. The single-cell profiles were simulated and analysed in a separate run of the same pipeline. Software versions are listed in Supplementary Table 14.

Benchmark on real datasets

A single feature filter was applied per layer, and every method received the same filtered features. RNA-sequencing and single-cell pseudobulk counts were filtered with filterByExpr[63] with a minimum count of 10. Metagenomic features were retained if present in at least 10% of the samples, the default prevalence cut-off of ANCOM-BC2. Proteins were retained if quantified in every sample and variable across samples. Metabolites were retained if they had at least three non-zero values and non-zero variance; features with a large fraction of zeros, which metabolomic workflows commonly remove, were kept so that zero-adjusted families could be compared on equal terms. Supplementary Table 2 gives the number of features before and after filtering.

The method panel and inputs followed the synthetic benchmark, with changes imposed by the data. edgeR used the quasi-likelihood F-test in every layer, as currently recommended.[63] The proteomic data were analyzed as supplied, as log2 tandem-mass-tag ratios normalized at source:

limma and PERSEO received these ratios without variance-stabilizing normalization, which is designed for linear intensities, and MSstats received 2 raised to the ratio plus a constant that made every intensity greater than one, so that its internal log2 transformation returned the ratios up to a constant. The WGS metagenomic matrices are rescaled to a constant total, so the depth covariate of PERSEO was omitted. ALDEx2 used 128 Monte Carlo instances. ANCOM-BC2 was run without prevalence or library-size filters and without structural-zero detection; taxa with zero variance that prevented the fit were removed and counted as not called. PERSEO was run with the settings used in simulation and its default panel of candidate families. The *P* values of every method were adjusted within each dataset with the Benjamini–Hochberg procedure over all filtered features, and a feature was called at $q \leq 0.05$. LEfSe, which returns calls without *P* values, was included only in call-based comparisons.

Agreement between PERSEO and each competitor was summarized by the Spearman correlation of *P* values over features with a *P* value from both methods, by the concordance-at-the-top curve, which gives the fraction of the *k* top-ranked features shared by two methods,[64] and by the overlap of calls. For each pair formed by PERSEO and a competitor, features were classified as called by both methods, by PERSEO only or by the competitor only, and each feature was labelled with the family that PERSEO selected for it on that dataset.

False discoveries were measured by repeating the complete analysis once per dataset after a balanced permutation of the group labels, in which each pseudo-group received the same proportion of true controls and cases. The real effect then cancels and every call is false. Label permutation has previously been used to expose false discoveries of differential expression methods in population-scale data.[31] For each method and PERSEO family we recorded the number of calls with permuted labels and the permutation FDR, defined as the number of calls with permuted labels divided by the number of calls with true labels.

Disagreements were interpreted through the simulation. Each real dataset was matched to the simulated profile of its layer (Supplementary Table 2). In the simulated data of that profile, pooled over the baseline, noisy, outlier and zero-inflation scenarios, features were classified with the same rules as the real data, by the family that PERSEO selected and by agreement category, and the fraction of true positives in each category was computed against the simulated truth. The batch scenario was excluded because the real datasets were analyzed without batch covariates. Multiplying this fraction by the number of real features in the same category gives the expected numbers of true and false positives. When a family contributed fewer than ten simulated features to a category, the fraction pooled over all families of the profile was used. A set was classified as

probably true when the simulated fraction was at least 0.8, as probably false when it was below 0.5 and as mixed otherwise; a fraction below 0.5 that rested only on the pooled value was classified as mixed. This classification was then checked against the permutation. A method was considered not to withstand the permutation in a family when it kept at least three calls there with permuted labels and more than 25% of its calls with true labels. In that case a set classified as probably true was reported as discordant with the simulation, and a mixed set as probably false; for features called by both methods, both had to fail the check. The same simulated data provided, for each method and PERSEO family, the FDR observed in simulation, which was compared with the permutation FDR on real data by Spearman correlation over the method-by-family combinations with at least 20 calls in the real dataset.

Analyses ran on the Linux server and R installation used for the synthetic benchmark, with 20 parallel workers for PERSEO and ANCOM-BC2 (versions in Supplementary Table 14).

Statistics and reproducibility

In the synthetic benchmark, each profile and scenario was simulated once with a fixed seed, with 80 samples per group in the bulk profiles and 20 donors per group, each contributing 250 cells, in the single-cell profiles, where the donor is the experimental unit. Performance was computed over all features of each simulated dataset (12,000, 5,000 or 2,000). Because no replicate simulations were run, differences between methods are reported as point estimates without confidence intervals. In the real-data benchmark the unit is the subject, with one sample per subject and one pseudobulk profile per donor in single-cell data, and sample sizes were those of the public datasets (Supplementary Table 2); no statistical method was used to predetermine sample size. Randomization and blinding do not apply to this computational study.

All tests were two-sided. PERSEO tested the group coefficient with a Wald t statistic on the residual degrees of freedom; DESeq2 used the Wald test in bulk data and the likelihood-ratio test at the single-cell level; edgeR used the likelihood-ratio test in the synthetic bulk profiles and the quasi-likelihood F-test in single-cell pseudobulk and real data; limma and limma-voom used moderated t statistics; ANCOM-BC2 tested its bias-corrected log fold change with a Wald-type statistic; ALDEx2 reported the expected P value of Welch's t-test over Monte Carlo instances; and glmmTMB tested the group coefficient with a Wald z statistic. P values were adjusted with the Benjamini–Hochberg procedure within each dataset and method, and features were called at $q \leq 0.05$. Features without a P value were counted as not called; this affected DESeq2 at the cell level in the plate full-length profile, which left 56–68% of genes untested, and glmmTMB, which did not converge for 8–15% of genes in the UMI profiles.

Observed FDR was defined as FP/(TP + FP) and set to zero when a method made no calls, sensitivity as TP/(TP + FN) and the F1-score as 2TP/(2TP + FP + FN); the Matthews correlation coefficient was computed from the four cells of the confusion matrix. In real data, agreement was measured by the Spearman correlation of *P* values over features tested by both methods, and the correspondence between simulated and permutation FDRs by the Spearman correlation over the 66 method-by-family combinations with at least 20 calls; these correlations are descriptive. One balanced label permutation was run per dataset. Seeds were fixed for the simulation, rarefaction, the bootstrap family screening of PERSEO and the permutation.

The synthetic runs correspond to code commits 320b651 (bulk profiles) and ef33ada (single-cell profiles). Their configuration files, manifests with package versions, per-method outputs and metrics, and the per-dataset outputs and permutation results of the real-data benchmark, are archived with the code (Code availability).

Software

All analyses used R 4.5.3 with PERSEO 1.0.0, gamlss.dist 6.1.1, DESeq2 1.48.2, edgeR 4.6.3, limma 3.66.0, vsn 3.76.0, ANCOMBC 2.13.2, ALDEx2 1.40.0, lefser 1.20.2, MSstats 4.21.0 and glmmTMB 1.1.14 (Supplementary Table 14). The characterization was run with R 4.3 or later on Bioconductor 3.18–3.22.

Data availability

All datasets analysed are publicly available. Accessions of the characterization datasets are listed in Supplementary Table 1. The 12 datasets of the real-data benchmark are SRA study SRP074887 (through recount3; http://rna.recount.bio), TCGA-BRCA at the Genomic Data Commons (https://portal.gdc.cancer.gov), the Segerstolpe pancreas (ArrayExpress E-MTAB-5061) and Ledergor myeloma datasets through the scRNAseq Bioconductor package, the 16S datasets cdi_vincent and crc_baxter from Nearing et al.,[19] the curatedMetagenomicData studies Castro-NallarE_2015 and QinN_2014, the CPTAC studies PDC000204 and PDC000270 at the Proteomic Data Commons (https://pdc.cancer.gov), and the Metabolomics Workbench studies ST003945 and ST002498.[65,66] The synthetic datasets can be regenerated exactly from the code, the archived configuration and the seed; the simulated ground truth and all method outputs are archived with the code.

Code availability

PERSEO is freely available at https://github.com/OPSergio/Perseo. The code for the characterization, the synthetic benchmark (simulation, scoring and figures) and the real-data benchmark, together with the archived configurations and results, is available at https://github.com/OPSergio/benchmarking.

References

1. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).
2. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
3. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
4. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15**, R29 (2014).
5. Smyth, G. K. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* **3**, 3 (2004).
6. Huber, W., von Heydebreck, A., Sültmann, H., Poustka, A. & Vingron, M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **18**, S96–S104 (2002).
7. Choi, M. et al. MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. *Bioinformatics* **30**, 2524–2526 (2014).
8. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
9. Welch, B. L. The generalization of 'Student's' problem when several different population variances are involved. *Biometrika* **34**, 28–35 (1947).
10. Lin, H. & Peddada, S. D. Multigroup analysis of compositions of microbiomes with covariate adjustments and repeated measures. *Nat. Methods* **21**, 83–91 (2024).
11. Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V. & Egozcue, J. J. Microbiome datasets are compositional: and this is not optional. *Front. Microbiol.* **8**, 2224 (2017).
12. Fernandes, A. D. et al. Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome* **2**, 15 (2014).
13. Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S. & Vert, J.-P. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat. Commun.* **9**, 284 (2018).

14. Van den Berge, K. et al. Observation weights unlock bulk RNA-seq tools for zero inflation and single-cell applications. *Genome Biol.* **19**, 24 (2018).

15. Martin, B. D., Witten, D. & Willis, A. D. Modeling microbial abundances and dysbiosis with beta-binomial regression. *Ann. Appl. Stat.* **14**, 94–115 (2020).

16. Svensson, V. Droplet scRNA-seq is not zero-inflated. *Nat. Biotechnol.* **38**, 147–150 (2020).

17. Hawinkel, S., Mattiello, F., Bijnens, L. & Thas, O. A broken promise: microbiome differential abundance methods do not control the false discovery rate. *Brief. Bioinform.* **20**, 210–221 (2019).

18. Brombacher, E., Schilling, O. & Kreutz, C. Characterizing the omics landscape based on 10,000+ datasets. *Sci. Rep.* **15**, 3189 (2025).

19. Nearing, J. T. et al. Microbiome differential abundance methods produce different results across 38 datasets. *Nat. Commun.* **13**, 342 (2022).

20. Peng, H., Wang, H., Kong, W., Li, J. & Goh, W. W. B. Optimizing differential expression analysis for proteomics data via high-performing rules and ensemble inference. *Nat. Commun.* **15**, 3922 (2024).

21. Boer, J. M. et al. Identification and classification of differentially expressed genes in renal cell carcinoma by expression profiling on a global human 31,500-element cDNA array. *Genome Res.* **11**, 1861–1870 (2001).

22. Robinson, M. D. & Smyth, G. K. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics* **23**, 2881–2887 (2007).

23. Wu, H., Wang, C. & Wu, Z. A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. *Biostatistics* **14**, 232–243 (2013).

24. Hardcastle, T. J. & Kelly, K. A. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics* **11**, 422 (2010).

25. Van De Wiel, M. A. et al. Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors. *Biostatistics* **14**, 113–128 (2013).

26. McCarthy, D. J., Chen, Y. & Smyth, G. K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* **40**, 4288–4297 (2012).

27. Jiang, R., Sun, T., Song, D. & Li, J. J. Statistics or biology: the zero-inflation controversy about scRNA-seq data. *Genome Biol.* **23**, 31 (2022).

28. Rigby, R. A. & Stasinopoulos, D. M. Generalized additive models for location, scale and shape. *J. R. Stat. Soc. C* **54**, 507–554 (2005).

29. McMurdie, P. J. & Holmes, S. Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput. Biol.* **10**, e1003531 (2014).

30. Squair, J. W. et al. Confronting false discoveries in single-cell differential expression. *Nat. Commun.* **12**, 5692 (2021).

31. Li, Y., Ge, X., Peng, F., Li, W. & Li, J. J. Exaggerated false positives by popular differential expression methods when analyzing human population samples. *Genome Biol.* **23**, 79 (2022).

32. Stasinopoulos, D. M. & Rigby, R. A. Generalized additive models for location scale and shape (GAMLSS) in R. *J. Stat. Softw.* **23**, 1–46 (2007).

33. Stasinopoulos, D. M., Kneib, T., Klein, N., Mayr, A. & Heller, G. Z. *Generalized Additive Models for Location, Scale and Shape: A Distributional Regression Approach, with Applications* (Cambridge Univ. Press, Cambridge, 2024).

34. Wirbel, J., Essex, M., Forslund, S. K. & Zeller, G. A realistic benchmark for differential abundance testing and confounder adjustment in human microbiome studies. *Genome Biol.* **25**, 247 (2024).

35. Sarkar, A. & Stephens, M. Separating measurement and expression models clarifies confusion in single-cell RNA sequencing analysis. *Nat. Genet.* **53**, 770–777 (2021).

36. Wilks, C. et al. recount3: summaries and queries for large-scale RNA-seq expression and splicing. *Genome Biol.* **22**, 323 (2021).

37. Colaprico, A. et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.* **44**, e71 (2016).

38. Risso, D. & Cole, M. scRNAseq: collection of public single-cell RNA-seq datasets. Bioconductor https://doi.org/10.18129/B9.bioc.scRNAseq (2017).

39. Edwards, N. J. et al. The CPTAC Data Portal: a resource for cancer proteomics research. *J. Proteome Res.* **14**, 2707–2713 (2015).

40. Sud, M. et al. Metabolomics Workbench: an international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. *Nucleic Acids Res.* **44**, D463–D470 (2016).

41. Pasolli, E. et al. Accessible, curated metagenomic data through ExperimentHub. *Nat. Methods* **14**, 1023–1024 (2017).

42. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, R25 (2010).

43. Dieterle, F., Ross, A., Schlotterbeck, G. & Senn, H. Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in ^{1}H NMR metabonomics. *Anal. Chem.* **78**, 4281–4290 (2006).

44. Bondulich, M. K. et al. Myostatin inhibition prevents skeletal muscle pathophysiology in Huntington's disease mice. *Sci. Rep.* **7**, 14275 (2017).

45. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).

46. Segerstolpe, Å. et al. Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. *Cell Metab.* **24**, 593–607 (2016).

47. Ledergor, G. et al. Single cell dissection of plasma cell heterogeneity in symptomatic and asymptomatic myeloma. *Nat. Med.* **24**, 1867–1876 (2018).

48. Vincent, C. et al. Reductions in intestinal Clostridiales precede the development of nosocomial Clostridium difficile infection. *Microbiome* **1**, 18 (2013).

49. Baxter, N. T., Ruffin, M. T., Rogers, M. A. M. & Schloss, P. D. Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions. *Genome Med.* **8**, 37 (2016).

50. Castro-Nallar, E. et al. Composition, taxonomy and functional diversity of the oropharynx microbiome in individuals with schizophrenia and controls. *PeerJ* **3**, e1140 (2015).

51. Qin, N. et al. Alterations of the human gut microbiome in liver cirrhosis. *Nature* **513**, 59–64 (2014).

52. Wang, L.-B. et al. Proteogenomic and metabolomic characterization of human glioblastoma. *Cancer Cell* **39**, 509–528 (2021).

53. Cao, L. et al. Proteogenomic characterization of pancreatic ductal adenocarcinoma. *Cell* **184**, 5031–5052 (2021).

54. Wald, A. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Trans. Am. Math. Soc.* **54**, 426–482 (1943).

55. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300 (1995).

56. Huber, W. et al. Orchestrating high-throughput genomic analysis with Bioconductor. *Nat. Methods* **12**, 115–121 (2015).

57. Crowell, H. L. et al. muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data. *Nat. Commun.* **11**, 6077 (2020).

58. Segata, N. et al. Metagenomic biomarker discovery and explanation. *Genome Biol.* **12**, R60 (2011).

59. Khleborodova, A. et al. lefser: implementation of metagenomic biomarker discovery tool, LEfSe, in R. *Bioinformatics* **40**, btae707 (2024).

60. Lund, S. P., Nettleton, D., McCarthy, D. J. & Smyth, G. K. Detecting differential expression in RNA-sequence data using quasi-likelihood with shrunken dispersion estimates. *Stat. Appl. Genet. Mol. Biol.* **11**, 8 (2012).

61. Brooks, M. E. et al. glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *R J.* **9**, 378–400 (2017).

62. Mann, H. B. & Whitney, D. R. On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.* **18**, 50–60 (1947).

63. Chen, Y., Lun, A. T. L. & Smyth, G. K. From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. *F1000Research* **5**, 1438 (2016).

64. Irizarry, R. A. et al. Multiple-laboratory comparison of microarray platforms. *Nat. Methods* **2**, 345–350 (2005).

65. Metabolomics Workbench. Lipidomics characterization of A375 and A375-VR melanoma cells treated with ranolazine and sampled longitudinally (project PR002472, study ST003945). Metabolomics Workbench https://doi.org/10.21228/M83R8F (2024).

66. Metabolomics Workbench. Plasma metabolomics profiling of 580 patients from the Weill Cornell Medicine Early Detection Research Network prostate cancer cohort (project PR001613, study ST002498). Metabolomics Workbench https://doi.org/10.21228/M86H7K (2022).

67. Grossman, R. L. et al. Toward a shared vision for cancer genomic data. *N. Engl. J. Med.* **375**, 1109–1112 (2016).

68. Duvallet, C., Gibbons, S. M., Gurry, T., Irizarry, R. A. & Alm, E. J. Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. *Nat. Commun.* **8**, 1784 (2017).

Acknowledgements

Author contributions

Competing interests

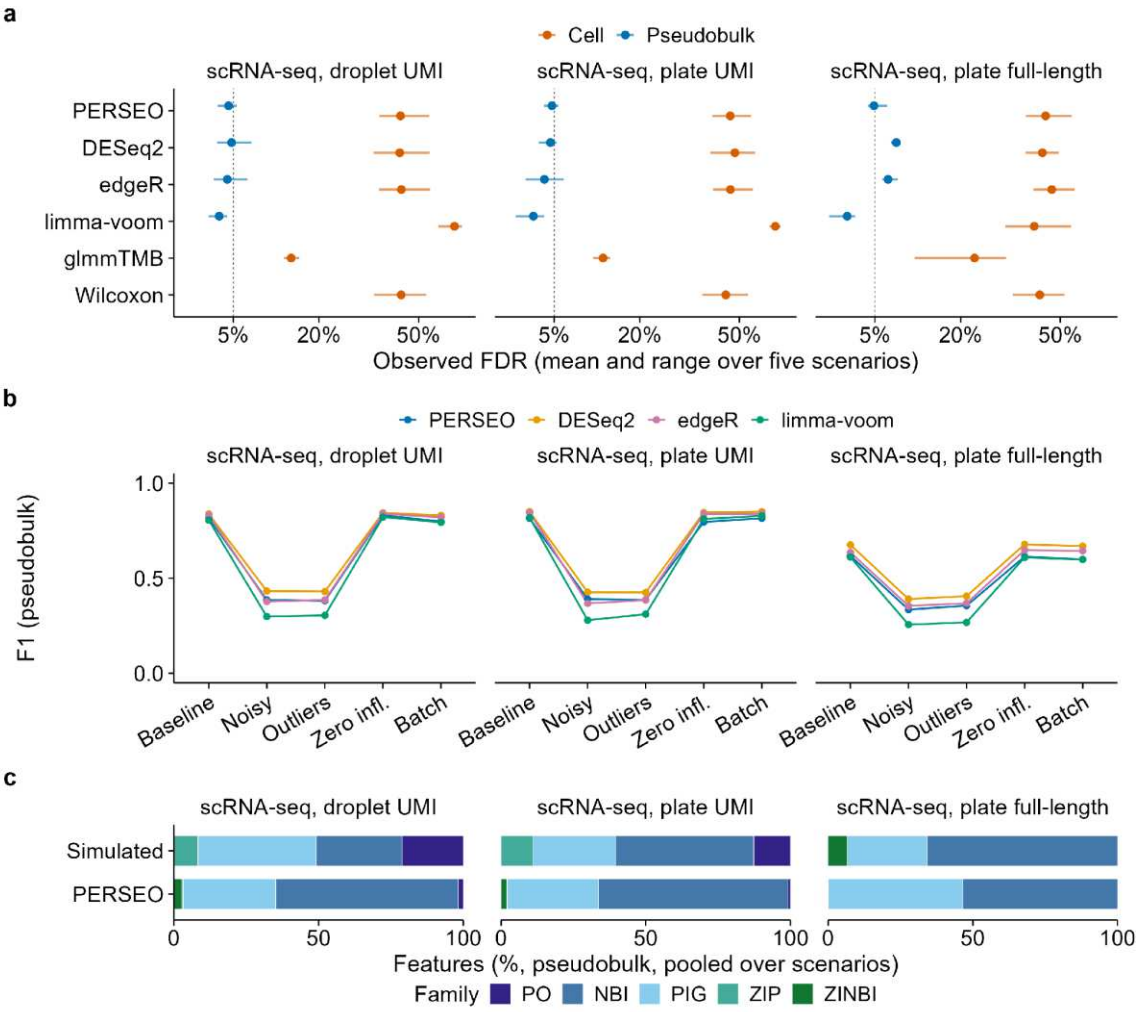
The authors declare no competing interests.

Additional information

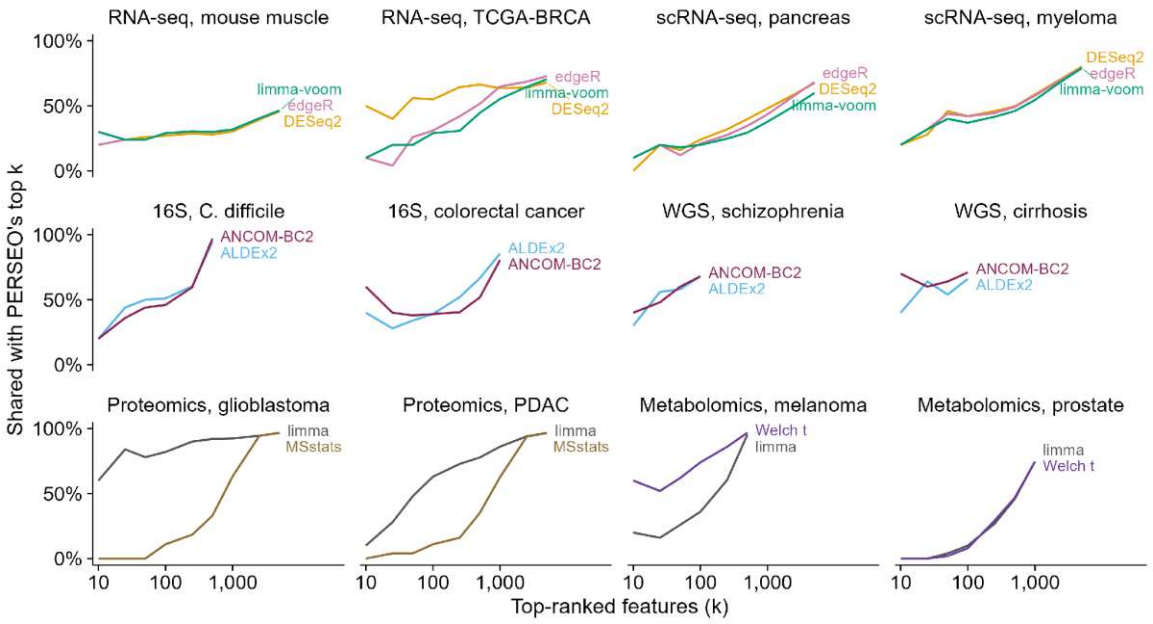
Supplementary Information is available for this paper.

Correspondence and requests for materials should be addressed to Val Fernández-Lanza.

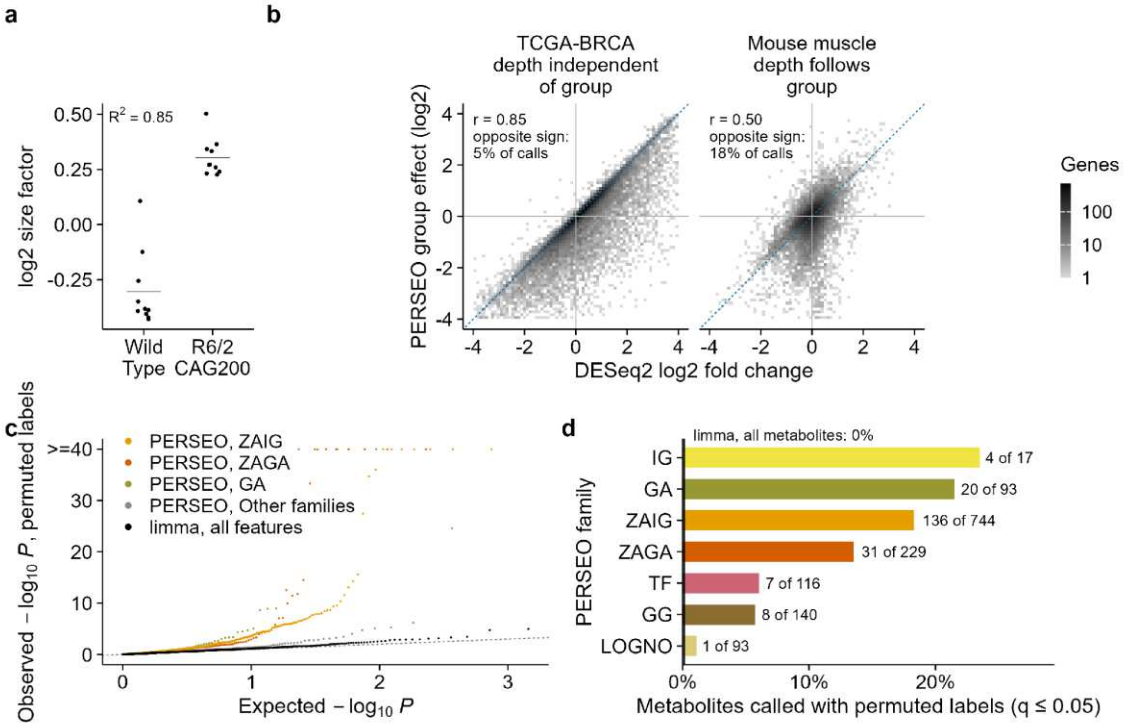
Supplementary Material



Supplementary Fig. 1 | Single-cell profiles. a, Observed FDR of cell-level and pseudobulk analyses (mean and range over the five scenarios); dashed line, 0.05. Cell-level methods were applied to raw counts, except the Wilcoxon test (log-normalized counts); glmmTMB fitted a donor random intercept. **b**, F1-score at the pseudobulk level by scenario and method. **c**, Families of the simulated features and families selected by PERSEO at the pseudobulk level, pooled over scenarios.



Supplementary Fig. 2 | Concordance at the top between PERSEO and each competitor. Fraction of the *k* top-ranked features (by *P* value) shared between PERSEO and each competitor, for each of the 12 real datasets.



Supplementary Fig. 3 | Two situations not covered by the simulation. a, log2 DESeq2 size factors in the mouse muscle RNA-seq study (SRP074887) by group (R^2 of size factor on group). **b**, Group effect estimated by PERSEO (log2) against the DESeq2 log2 fold change for every gene in TCGA-BRCA (left), where depth is independent of the group, and in the mouse muscle study (right), where it follows the group; genes beyond ±4 are not drawn. **c**, Quantile–quantile plot of *P*

values with permuted labels in prostate cancer metabolomics (580 samples), for PERSEO by selected family and for limma. **d**, Fraction of metabolites called with permuted labels ($q \leq 0.05$) by PERSEO family; the vertical line marks limma.

Supplementary Table 1 | Public datasets used for the family characterization. Datasets with family assignments at the raw scale (131); four further metabolomic studies had only normalized conditions. Retrieval tools: recount3,[36] TCGAbiolinks[37] over the Genomic Data Commons,[67] the scRNAseq Bioconductor package,[38] curatedMetagenomicData via ExperimentHub,[41] MicrobiomeHD[68] as aggregated by Nearing et al.,[19] the Proteomic Data Commons[39] and Metabolomics Workbench.[40] UMI, unique molecular identifier; TMT, tandem mass tag.

Layer	Profile	Repository (retrieval tool)	Datasets	Native data type
Transcriptomics	Consortium cohorts (GTEx, TCGA)	recount3; TCGAbiolinks (GDC)	10	Gene-level read counts
Transcriptomics	Independent studies (SRA)	recount3	11	Gene-level read counts
Transcriptomics	Single-cell, droplet UMI	scRNAseq (Bioconductor)	8	UMI counts
Transcriptomics	Single-cell, plate UMI	scRNAseq (Bioconductor)	7	UMI counts
Transcriptomics	Single-cell, plate full-length	scRNAseq (Bioconductor)	8	Read counts
Metagenomics	16S rRNA metagenomics	MicrobiomeHD[19]	39	Amplicon sequence variant counts
Metagenomics	WGS metagenomics	curatedMetagenomicData (ExperimentHub)	20	MetaPhlAn abundances as integer pseudo-counts
Proteomics	CPTAC (mass spectrometry)	Proteomic Data Commons	12	log2 TMT ratios, normalized at source
Metabolomics	MS and NMR	Metabolomics Workbench (REST)	16	Intensities

Supplementary Table 2 | Datasets of the real-data benchmark. n_{eff}, effective sample size, $2n_0n_1/(n_0 + n_1)$; in single-cell datasets n counts donors. Features are given before and after the feature filter of each layer. The simulated profile is the one used to read the real-data disagreements through the simulation. T2D, type 2 diabetes; CDI, *Clostridioides difficile* infection; CRC, colorectal cancer; GBM, glioblastoma; PDAC, pancreatic ductal adenocarcinoma; FL, full-length.

Layer	Dataset (accession)	Contrast	Control	Case	n_{eff}	Features	Simulated profile
RNA-seq	Mouse muscle (SRP074887)	R6/2 vs wild type	10	10	10.0	55,421 → 19,112	RNA-seq, independent studies
RNA-seq	TCGA-BRCA	Tumour vs normal tissue	113	1,095	204.9	60,660 → 27,023	RNA-seq, consortia
scRNA-seq	Alpha cells (E-MTAB-5061)	T2D vs healthy	6	4	4.8	26,179 → 14,715	Plate FL
scRNA-seq	Plasma cells (Ledergor)	Myeloma vs control	11	14	12.3	57,874 → 9,234	Plate UMI
16S metagenomics	cdi_vincent	CDI vs healthy	23	17	19.6	760 → 527	16S metagenomics
16S metagenomics	crc_baxter	CRC vs healthy	146	117	129.9	18,413 → 1,263	16S metagenomics
WGS metagenomics	Castro-NallarE_2015	Schizophrenia vs control	16	16	16.0	266 → 172	WGS metagenomics

Layer	Dataset (accession)	Contrast	Control	Case	n_{eff}	Features	Simulated profile
WGS metagenomics	QinN_2014	Cirrhosis vs control	114	123	118.3	654 → 219	WGS metagenomics
Proteomics	GBM (PDC000204)	Tumour vs normal tissue	10	100	18.2	10,977 → 9,363	Proteomics
Proteomics	PDAC (PDC000270)	Tumour vs normal tissue	45	98	61.7	10,116 → 7,133	Proteomics
Metabolomics	A375 lipidome (ST003945)	Ranolazine vs vehicle, 24 h	8	8	8.0	527 → 524	Metabolomics
Metabolomics	Prostate plasma (ST002498)	Cancer vs no cancer	313	267	288.2	1,482 → 1,441	Metabolomics

946

947 **Supplementary Table 3 | Distribution families by omics profile and normalization.** Families reaching at least 5% of
948 features, pooled over the datasets of each profile, at the raw scale and after the normalizations of each layer. Family
949 codes are defined in Supplementary Table 4.

Profile	Normalization	Datasets	Families (≥ 5% of features)
RNA-seq, consortia (GTEx, TCGA)	Raw	10	PIG 59%, NBI 32%, PO 6%
	TMM	10	BE 44%, ZAGA 16%, GG 13%, ZAIG 9%, LOGNO 7%, GA 6%
	DESeq2 size factors	10	ZAIG 57%, GG 13%, ZAGA 12%, LOGNO 6%, GA 6%
	VST	10	TF 50%, GG 32%, LOGNO 5%, NO 5%
RNA-seq, independent studies (SRA)	Raw	11	PIG 39%, NBI 31%, PO 22%
	TMM	11	BE 42%, ZAIG 11%, ZAGA 9%, GA 9%, NO 8%, GG 7%, IG 6%
	DESeq2 size factors	11	ZAIG 35%, ZAGA 22%, GA 10%, NO 9%, GG 7%, IG 6%
	VST	11	TF 33%, GG 23%, NO 17%, GU 14%, IG 6%, LOGNO 5%
16S rRNA metagenomics	Raw	39	NBI 36%, ZIP 34%, PIG 21%, PO 8%
	TMM	39	ZAIG 61%, ZAGA 39%
	Rarefied counts	39	PO 28%, NBI 28%, PIG 25%, ZIP 19%
	Rarefied proportions	39	BE 100%
WGS shotgun metagenomics	Raw	20	NBI 89%, ZINBI 7%
	TMM	20	ZAGA 56%, ZAIG 44%
	Rarefied counts	20	NBI 89%, ZINBI 7%
	Rarefied proportions	20	BE 100%
Proteomics (CPTAC)	Raw	12	TF 48%, NO 46%, GU 6%
	log2(x + 1)	12	ZAGA 66%, BE 33%
	Median centring	12	TF 48%, NO 46%, GU 6%

Profile	Normalization	Datasets	Families (≥ 5% of features)
	VSN	12	NO 87%, GU 9%
Metabolomics	Raw	16	ZAIG 19%, ZAGA 15%, BE 14%, PIG 12%, GG 9%, LOGNO 7%, GA 6%, NBI 5%
	log2(x + 1)	20	ZAGA 27%, ZAIG 17%, BE 12%, GG 12%, TF 8%, NO 6%
	PQN	20	ZAIG 28%, ZAGA 17%, GG 17%, BE 11%, LOGNO 9%, GA 6%
	VSN	20	TF 54%, NO 18%, GU 14%, GG 9%
scRNA-seq, droplet UMI	Raw	8	PIG 40%, NBI 30%, PO 21%, ZIP 8%
	TMM	8	ZAIG 78%, ZAGA 22%
	Log-normalized	8	ZAGA 75%, ZAIG 23%
	Pseudobulk	8	PIG 57%, NBI 30%, PO 7%
scRNA-seq, plate UMI	Raw	7	NBI 46%, PIG 27%, PO 12%, ZIP 11%
	TMM	7	ZAIG 85%, ZAGA 15%
	Log-normalized	7	ZAIG 61%, ZAGA 34%, BE 5%
	Pseudobulk	5	NBI 45%, PIG 36%, PO 9%, ZIP 6%
scRNA-seq, plate full-length	Raw	8	NBI 61%, PIG 26%, ZINBI 6%
	TMM	7	ZAIG 51%, ZAGA 46%
	Log-normalized	8	ZAGA 58%, BE 38%
	Pseudobulk	4	NBI 67%, PIG 18%, ZINBI 11%

Supplementary Table 4 | GAMLSS families. Codes as in gamlss.dist.

Code	Distribution	Type
PO	Poisson	Discrete
NBI	Negative binomial type I	Discrete, overdispersed
PIG	Poisson–inverse Gaussian	Discrete, overdispersed
ZIP	Zero-inflated Poisson	Discrete
ZINBI	Zero-inflated negative binomial	Discrete
BI	Binomial	Discrete, bounded counts
BB	Beta-binomial	Discrete, overdispersed bounded counts
NO	Gaussian (normal)	Continuous
TF	Student's t	Continuous, heavy-tailed

Code	Distribution	Type
GU	Gumbel	Continuous
LOGNO	Log-normal	Continuous, positive
GA	Gamma	Continuous, positive
GG	Generalized gamma	Continuous, positive
IG	Inverse Gaussian	Continuous, positive
ZAGA	Zero-adjusted gamma	Continuous, ≥ 0
ZAIG	Zero-adjusted inverse Gaussian	Continuous, ≥ 0
ZILN	Zero-inflated log-normal	Continuous, ≥ 0
BE	Beta	Continuous on (0, 1)
BEo	Beta (mean–precision form)	Continuous on (0, 1)
BEZI	Zero-inflated beta	Continuous on [0, 1)
BEINF0	Beta inflated at zero	Continuous on [0, 1)
BEINF	Inflated beta	Continuous on [0, 1]

Supplementary Table 5 | Generative family mixtures and design of each simulated profile. For every profile the mean proportion of each family was computed across its datasets (each dataset weighted equally); families with a mean proportion of at least 5% were retained and rescaled to sum to one. ZIP in the independent-study RNA-seq profile (0.042) was retained below the threshold to preserve its zero inflation. The metabolomic mixture excludes two studies quantified on an integer scale. Depth s.d., standard deviation of log library size (per sample in bulk profiles, per cell in single-cell profiles); for WGS metagenomics it was assumed, because the matrices are rescaled to a constant total. Single-cell profiles: target fraction of zeros after filtering 0.84, 0.79 and 0.61, and cell-level dispersion of the negative binomial-type families drawn from 1.25–3.33, 1.40–5.22 and 2.38–24.58 for droplet UMI, plate UMI and plate full-length, respectively. Family codes are defined in Supplementary Table 4.

Profile	Datasets	Generative family mixture	Features	Samples	Depth s.d.
RNA-seq, consortia (GTEx, TCGA)	10	PIG 0.607, NBI 0.330, PO 0.063	12,000	80 per group	0.31
RNA-seq, independent studies (SRA)	11	PIG 0.386, NBI 0.303, PO 0.218, ZINBI 0.051, ZIP 0.042	12,000	80 per group	0.25
16S rRNA metagenomics	39	NBI 0.453, ZIP 0.351, PIG 0.196	12,000	80 per group	0.47
WGS shotgun metagenomics	20	NBI 0.907, ZINBI 0.093	12,000	80 per group	0.35
Proteomics (CPTAC)	12	TF 0.478, NO 0.463, GU 0.059	2,000	80 per group	–
Metabolomics	14	ZAIG 0.251, ZAGA 0.241, BE 0.161, GG 0.127, GA 0.117, LOGNO 0.103	2,000	80 per group	–

Profile	Datasets	Generative family mixture	Features	Samples	Depth s.d.
scRNA-seq, droplet UMI	8	PIG 0.405, NBI 0.298, PO 0.215, ZIP 0.082	5,000	20 donors × 250 cells	0.65
scRNA-seq, plate UMI	7	NBI 0.478, PIG 0.281, PO 0.131, ZIP 0.110	5,000	20 donors × 250 cells	0.62
scRNA-seq, plate full-length	8	NBI 0.654, PIG 0.276, ZINBI 0.070	5,000	20 donors × 250 cells	1.21

962

963 **Supplementary Table 6 | Performance in the baseline scenario.** Methods on their canonical input at $q \leq 0.05$; single-
964 cell profiles at the pseudobulk level. TPR high and TPR low, sensitivity for high-effect (fold change 3.0 or 1.5 s.d.) and
965 low-effect (fold change 1.5 or 0.5 s.d.) features; best F1-score per profile in bold. MSstats values are affected by a
966 double log2 transformation (Methods).

Profile	Method	Input	F1	MCC	FDR	TPR high	TPR low
RNA-seq, consortia (GTEx, TCGA)	PERSEO	Raw counts + log size factor	**0.926**	0.890	0.038	0.997	0.876
	DESeq2	Raw counts	0.924	0.887	0.038	0.997	0.871
	edgeR	Raw counts	**0.926**	0.889	0.041	0.997	0.878
	limma-voom	Raw counts	0.920	0.882	0.034	0.997	0.858
RNA-seq, independent studies (SRA)	PERSEO	Raw counts + log size factor	**0.927**	0.891	0.039	1.000	0.878
	DESeq2	Raw counts	0.918	0.880	0.033	0.997	0.854
	edgeR	Raw counts	0.921	0.884	0.037	1.000	0.863
	limma-voom	Raw counts	0.896	0.852	0.033	0.997	0.808
16S rRNA metagenomics	PERSEO	Raw counts + log size factor	**0.924**	0.885	0.060	0.998	0.894
	ANCOM-BC2	Raw counts	0.907	0.860	0.071	0.978	0.871
	ALDEx2	Raw counts	0.649	0.609	0.010	0.810	0.428
	LEfSe	Rarefied proportions	0.095	0.182	0.000	0.198	0.025
WGS shotgun metagenomics	PERSEO	Raw counts + log size factor	**0.934**	0.901	0.034	0.998	0.889
	ANCOM-BC2	Raw counts	0.916	0.878	0.031	0.992	0.849
	ALDEx2	Raw counts	0.787	0.735	0.012	0.905	0.612
	LEfSe	Rarefied proportions	0.081	0.167	0.000	0.140	0.026
Proteomics (CPTAC)	PERSEO	Raw intensities	**0.837**	0.781	0.034	1.000	0.695
	limma	VSN	0.769	0.713	0.022	1.000	0.573
	MSstats	log2(x + 1)	0.692	0.638	0.031	0.860	0.484
Metabolomics	PERSEO	Raw intensities	**0.852**	0.793	0.049	0.850	0.758
	limma	VSN	0.734	0.671	0.044	0.770	0.567

Profile	Method	Input	F1	MCC	FDR	TPR high	TPR low
	Welch t-test	log2(x + 1)	0.755	0.690	0.048	0.820	0.593
scRNA-seq, droplet UMI	PERSEO	Pseudobulk counts + log size factor	0.818	0.759	0.034	1.000	0.660
	DESeq2	Pseudobulk counts	**0.838**	0.782	0.033	1.000	0.695
	edgeR	Pseudobulk counts	0.831	0.776	0.030	1.000	0.682
	limma-voom	Pseudobulk counts	0.806	0.746	0.035	1.000	0.641
scRNA-seq, plate UMI	PERSEO	Pseudobulk counts + log size factor	0.815	0.753	0.045	1.000	0.662
	DESeq2	Pseudobulk counts	**0.850**	0.791	0.051	1.000	0.731
	edgeR	Pseudobulk counts	0.845	0.787	0.043	1.000	0.715
	limma-voom	Pseudobulk counts	0.819	0.759	0.039	1.000	0.665
scRNA-seq, plate full-length	PERSEO	Pseudobulk counts + log size factor	0.617	0.569	0.042	1.000	0.365
	DESeq2	Pseudobulk counts	**0.675**	0.607	0.070	0.996	0.453
	edgeR	Pseudobulk counts	0.635	0.576	0.062	0.992	0.395
	limma-voom	Pseudobulk counts	0.611	0.569	0.030	1.000	0.353

Supplementary Table 7 | Performance in the noisy (high dispersion) scenario. Methods on their canonical input at $q \leq 0.05$; single-cell profiles at the pseudobulk level. TPR high and TPR low, sensitivity for high-effect (fold change 3.0 or 1.5 s.d.) and low-effect (fold change 1.5 or 0.5 s.d.) features; best F1-score per profile in bold. MSstats values are affected by a double log2 transformation (Methods).

Profile	Method	Input	F1	MCC	FDR	TPR high	TPR low
RNA-seq, consortia (GTEx, TCGA)	PERSEO	Raw counts + log size factor	0.774	0.713	0.037	0.993	0.589
	DESeq2	Raw counts	0.778	0.707	0.065	0.988	0.613
	edgeR	Raw counts	**0.781**	0.711	0.064	0.995	0.616
	limma-voom	Raw counts	0.742	0.683	0.033	0.995	0.537
RNA-seq, independent studies (SRA)	PERSEO	Raw counts + log size factor	**0.815**	0.754	0.042	0.993	0.661
	DESeq2	Raw counts	0.778	0.713	0.045	0.987	0.601
	edgeR	Raw counts	0.786	0.719	0.052	0.995	0.617
	limma-voom	Raw counts	0.742	0.683	0.032	0.953	0.542
16S rRNA metagenomics	PERSEO	Raw counts + log size factor	**0.837**	0.782	0.028	0.995	0.692
	ANCOM-BC2	Raw counts	0.788	0.700	0.117	0.972	0.668
	ALDEx2	Raw counts	0.357	0.387	0.011	0.592	0.156
	LEfSe	Rarefied proportions	0.091	0.172	0.029	0.168	0.028

Profile	Method	Input	F1	MCC	FDR	TPR high	TPR low
WGS shotgun metagenomics	PERSEO	Raw counts + log size factor	**0.774**	0.715	0.031	0.993	0.587
	ANCOM-BC2	Raw counts	0.622	0.570	0.049	0.965	0.378
	ALDEx2	Raw counts	0.472	0.472	0.006	0.782	0.231
	LEfSe	Rarefied proportions	0.097	0.182	0.009	0.187	0.028
Proteomics (CPTAC)	PERSEO	Raw intensities	**0.832**	0.774	0.038	1.000	0.688
	limma	VSN	0.767	0.710	0.024	1.000	0.570
	MSstats	log2(x + 1)	0.543	0.514	0.033	0.920	0.287
Metabolomics	PERSEO	Raw intensities	**0.683**	0.606	0.090	0.730	0.517
	limma	VSN	0.447	0.445	0.029	0.400	0.272
	Welch t-test	log2(x + 1)	0.499	0.483	0.029	0.550	0.300
scRNA-seq, droplet UMI	PERSEO	Pseudobulk counts + log size factor	0.386	0.396	0.043	0.912	0.130
	DESeq2	Pseudobulk counts	**0.432**	0.426	0.050	0.976	0.164
	edgeR	Pseudobulk counts	0.377	0.391	0.042	0.952	0.115
	limma-voom	Pseudobulk counts	0.298	0.340	0.025	0.816	0.069
scRNA-seq, plate UMI	PERSEO	Pseudobulk counts + log size factor	0.390	0.395	0.055	0.908	0.135
	DESeq2	Pseudobulk counts	**0.427**	0.424	0.046	0.932	0.165
	edgeR	Pseudobulk counts	0.368	0.391	0.022	0.916	0.111
	limma-voom	Pseudobulk counts	0.279	0.327	0.024	0.792	0.058
scRNA-seq, plate full-length	PERSEO	Pseudobulk counts + log size factor	0.335	0.352	0.065	0.824	0.101
	DESeq2	Pseudobulk counts	**0.391**	0.385	0.081	0.896	0.140
	edgeR	Pseudobulk counts	0.354	0.368	0.059	0.868	0.110
	limma-voom	Pseudobulk counts	0.256	0.309	0.026	0.712	0.053

Supplementary Table 8 | Performance in the outlier (2% contaminated entries) scenario. Methods on their canonical input at $q \leq 0.05$; single-cell profiles at the pseudobulk level. TPR high and TPR low, sensitivity for high-effect (fold change 3.0 or 1.5 s.d.) and low-effect (fold change 1.5 or 0.5 s.d.) features; best F1-score per profile in bold. MSstats values are affected by a double log2 transformation (Methods).

Profile	Method	Input	F1	MCC	FDR	TPR high	TPR low
RNA-seq, consortia (GTEx, TCGA)	PERSEO	Raw counts + log size factor	0.766	0.702	0.044	0.993	0.580
	DESeq2	Raw counts	**0.767**	0.693	0.074	0.988	0.600
	edgeR	Raw counts	0.761	0.676	0.103	0.992	0.605

Profile	Method	Input	F1	MCC	FDR	TPR high	TPR low
	limma-voom	Raw counts	0.735	0.674	0.039	0.993	0.528
RNA-seq, independent studies (SRA)	PERSEO	Raw counts + log size factor	**0.795**	0.731	0.046	0.990	0.629
	DESeq2	Raw counts	0.768	0.698	0.061	0.982	0.594
	edgeR	Raw counts	0.766	0.682	0.100	0.990	0.613
	limma-voom	Raw counts	0.748	0.687	0.038	0.958	0.554
16S rRNA metagenomics	PERSEO	Raw counts + log size factor	**0.801**	0.741	0.036	0.993	0.634
	ANCOM-BC2	Raw counts	0.773	0.695	0.086	0.972	0.620
	ALDEx2	Raw counts	0.326	0.366	0.008	0.558	0.135
	LEfSe	Rarefied proportions	0.083	0.164	0.027	0.162	0.024
WGS shotgun metagenomics	PERSEO	Raw counts + log size factor	**0.766**	0.700	0.048	0.990	0.582
	ANCOM-BC2	Raw counts	0.629	0.572	0.058	0.968	0.389
	ALDEx2	Raw counts	0.459	0.460	0.013	0.775	0.220
	LEfSe	Rarefied proportions	0.097	0.175	0.045	0.160	0.033
Proteomics (CPTAC)	PERSEO	Raw intensities	**0.815**	0.752	0.048	1.000	0.665
	limma	VSN	0.683	0.629	0.034	1.000	0.450
	MSstats	log2(x + 1)	0.546	0.519	0.029	0.920	0.290
Metabolomics	PERSEO	Raw intensities	**0.660**	0.590	0.079	0.790	0.468
	limma	VSN	0.473	0.458	0.043	0.520	0.280
	Welch t-test	log2(x + 1)	0.512	0.486	0.047	0.700	0.292
scRNA-seq, droplet UMI	PERSEO	Pseudobulk counts + log size factor	0.381	0.389	0.054	0.904	0.128
	DESeq2	Pseudobulk counts	**0.430**	0.416	0.074	0.948	0.169
	edgeR	Pseudobulk counts	0.386	0.387	0.068	0.936	0.128
	limma-voom	Pseudobulk counts	0.305	0.339	0.042	0.852	0.069
scRNA-seq, plate UMI	PERSEO	Pseudobulk counts + log size factor	0.386	0.395	0.047	0.876	0.137
	DESeq2	Pseudobulk counts	**0.425**	0.420	0.053	0.928	0.165
	edgeR	Pseudobulk counts	0.385	0.389	0.062	0.920	0.129
	limma-voom	Pseudobulk counts	0.311	0.353	0.015	0.800	0.082
scRNA-seq, plate full-length	PERSEO	Pseudobulk counts + log size factor	0.356	0.373	0.050	0.840	0.115
	DESeq2	Pseudobulk counts	**0.406**	0.394	0.084	0.892	0.155
	edgeR	Pseudobulk counts	0.368	0.370	0.080	0.880	0.122
	limma-voom	Pseudobulk counts	0.268	0.323	0.011	0.716	0.061

977

978 **Supplementary Table 9 | Performance in the zero-inflation scenario.** Methods on their canonical input at $q \leq 0.05$;
979 single-cell profiles at the pseudobulk level. TPR high and TPR low, sensitivity for high-effect (fold change 3.0 or 1.5 s.d.)
980 and low-effect (fold change 1.5 or 0.5 s.d.) features; best F1-score per profile in bold. MSstats values are affected by a
981 double log2 transformation (Methods).

Profile	Method	Input	F1	MCC	FDR	TPR high	TPR low
RNA-seq, independent studies (SRA)	PERSEO	Raw counts + log size factor	**0.850**	0.795	0.036	0.982	0.724
	DESeq2	Raw counts	0.816	0.755	0.042	0.958	0.669
	edgeR	Raw counts	0.819	0.757	0.047	0.970	0.677
	limma-voom	Raw counts	0.790	0.730	0.033	0.943	0.623
16S rRNA metagenomics	PERSEO	Raw counts + log size factor	**0.842**	0.783	0.047	0.987	0.716
	ANCOM-BC2	Raw counts	0.806	0.725	0.101	0.970	0.690
	ALDEx2	Raw counts	0.417	0.433	0.004	0.610	0.206
	LEfSe	Rarefied proportions	0.096	0.181	0.014	0.145	0.035
WGS shotgun metagenomics	PERSEO	Raw counts + log size factor	**0.776**	0.713	0.042	0.947	0.604
	ANCOM-BC2	Raw counts	0.675	0.607	0.070	0.973	0.455
	ALDEx2	Raw counts	0.514	0.500	0.014	0.762	0.278
	LEfSe	Rarefied proportions	0.098	0.182	0.014	0.182	0.030
Metabolomics	PERSEO	Raw intensities	**0.729**	0.639	0.116	0.770	0.595
	limma	VSN	0.527	0.500	0.038	0.420	0.353
	Welch t-test	log2(x + 1)	0.543	0.516	0.029	0.510	0.355
scRNA-seq, droplet UMI	PERSEO	Pseudobulk counts + log size factor	0.831	0.770	0.048	1.000	0.693
	DESeq2	Pseudobulk counts	**0.844**	0.784	0.049	1.000	0.718
	edgeR	Pseudobulk counts	0.838	0.779	0.045	1.000	0.705
	limma-voom	Pseudobulk counts	0.822	0.761	0.043	1.000	0.673
scRNA-seq, plate UMI	PERSEO	Pseudobulk counts + log size factor	0.796	0.730	0.052	1.000	0.633
	DESeq2	Pseudobulk counts	**0.846**	0.787	0.046	1.000	0.719
	edgeR	Pseudobulk counts	0.837	0.779	0.041	1.000	0.701
	limma-voom	Pseudobulk counts	0.812	0.752	0.035	1.000	0.651
scRNA-seq, plate full-length	PERSEO	Pseudobulk counts + log size factor	0.613	0.566	0.044	0.992	0.361
	DESeq2	Pseudobulk counts	**0.678**	0.606	0.078	0.996	0.459
	edgeR	Pseudobulk counts	0.648	0.585	0.067	0.996	0.413

Profile	Method	Input	F1	MCC	FDR	TPR high	TPR low
	limma-voom	Pseudobulk counts	0.609	0.570	0.024	0.996	0.351

Supplementary Table 10 | Performance in the batch (covariate-aware methods only) scenario. Methods on their canonical input at $q \leq 0.05$; single-cell profiles at the pseudobulk level. TPR high and TPR low, sensitivity for high-effect (fold change 3.0 or 1.5 s.d.) and low-effect (fold change 1.5 or 0.5 s.d.) features; best F1-score per profile in bold. MSstats values are affected by a double log2 transformation (Methods).

Profile	Method	Input	F1	MCC	FDR	TPR high	TPR low
RNA-seq, consortia (GTEx, TCGA)	PERSEO	Raw counts + log size factor	**0.927**	0.891	0.039	0.997	0.879
	DESeq2	Raw counts	0.924	0.886	0.042	0.995	0.875
	edgeR	Raw counts	0.926	0.889	0.043	0.998	0.880
	limma-voom	Raw counts	0.919	0.881	0.033	0.998	0.856
RNA-seq, independent studies (SRA)	PERSEO	Raw counts + log size factor	**0.924**	0.887	0.040	0.998	0.873
	DESeq2	Raw counts	0.914	0.876	0.029	0.993	0.842
	edgeR	Raw counts	0.920	0.882	0.030	0.998	0.854
	limma-voom	Raw counts	0.891	0.846	0.028	0.990	0.794
16S rRNA metagenomics	PERSEO	Raw counts + log size factor	**0.932**	0.897	0.047	0.998	0.896
	ANCOM-BC2	Raw counts	0.922	0.881	0.068	0.990	0.900
WGS shotgun metagenomics	PERSEO	Raw counts + log size factor	**0.935**	0.902	0.038	0.997	0.894
	ANCOM-BC2	Raw counts	0.914	0.875	0.033	0.985	0.847
Proteomics (CPTAC)	PERSEO	Raw intensities	**0.822**	0.760	0.049	1.000	0.678
	limma	VSN	0.765	0.698	0.051	1.000	0.580
Metabolomics	PERSEO	Raw intensities	**0.848**	0.784	0.064	0.790	0.772
	limma	VSN	0.723	0.666	0.031	0.640	0.567
scRNA-seq, droplet UMI	PERSEO	Pseudobulk counts + log size factor	0.798	0.735	0.045	1.000	0.633
	DESeq2	Pseudobulk counts	**0.830**	0.773	0.035	1.000	0.683
	edgeR	Pseudobulk counts	0.820	0.763	0.032	1.000	0.663
	limma-voom	Pseudobulk counts	0.794	0.735	0.029	1.000	0.616
scRNA-seq, plate UMI	PERSEO	Pseudobulk counts + log size factor	0.815	0.755	0.039	1.000	0.659
	DESeq2	Pseudobulk counts	**0.850**	0.796	0.033	1.000	0.719
	edgeR	Pseudobulk counts	0.838	0.783	0.029	1.000	0.693
	limma-voom	Pseudobulk counts	0.828	0.772	0.030	1.000	0.676

Profile	Method	Input	F1	MCC	FDR	TPR high	TPR low
scRNA-seq, plate full-length	PERSEO	Pseudobulk counts + log size factor	0.599	0.554	0.044	0.988	0.344
	DESeq2	Pseudobulk counts	**0.669**	0.598	0.077	0.996	0.446
	edgeR	Pseudobulk counts	0.644	0.582	0.064	0.996	0.407
	limma-voom	Pseudobulk counts	0.598	0.561	0.023	1.000	0.336

Supplementary Table 11 | PERSEO performance by input representation. F1-score and, in parentheses, observed FDR at $q \leq 0.05$ for PERSEO on each input of the sequencing profiles. The first row of each profile is the canonical input used in the main comparisons; –, scenario not simulated for the profile.

Profile	Input	Baseline	Noisy	Outliers	Zero infl.	Batch
RNA-seq, consortia (GTEx, TCGA)	Raw counts + log size factor	0.926 (0.038)	0.774 (0.037)	0.766 (0.044)	–	0.927 (0.039)
	TMM counts, integer-rounded	0.910 (0.034)	0.764 (0.031)	0.753 (0.038)	–	0.901 (0.033)
	TMM-scaled counts	0.897 (0.043)	0.722 (0.041)	0.712 (0.057)	–	0.894 (0.047)
RNA-seq, independent studies (SRA)	Raw counts + log size factor	0.927 (0.039)	0.815 (0.042)	0.795 (0.046)	0.850 (0.036)	0.924 (0.040)
	TMM counts, integer-rounded	0.905 (0.027)	0.754 (0.031)	0.780 (0.036)	0.837 (0.037)	0.905 (0.027)
	TMM-scaled counts	0.895 (0.043)	0.773 (0.038)	0.754 (0.078)	0.810 (0.065)	0.894 (0.044)
16S rRNA metagenomics	Raw counts + log size factor	0.924 (0.060)	0.837 (0.028)	0.801 (0.036)	0.842 (0.047)	0.932 (0.047)
	TMM counts, integer-rounded	0.901 (0.066)	0.813 (0.028)	0.784 (0.041)	0.824 (0.072)	0.908 (0.050)
	TMM-scaled counts	0.880 (0.085)	0.799 (0.041)	0.752 (0.100)	0.801 (0.075)	0.887 (0.082)
	Rarefied counts	0.367 (0.031)	0.393 (0.056)	0.497 (0.058)	0.441 (0.044)	0.560 (0.038)
WGS shotgun metagenomics	Raw counts + log size factor	0.934 (0.034)	0.774 (0.031)	0.766 (0.048)	0.776 (0.042)	0.935 (0.038)
	TMM counts, integer-rounded	0.917 (0.029)	0.765 (0.036)	0.757 (0.044)	0.771 (0.047)	0.910 (0.033)
	TMM-scaled counts	0.910 (0.034)	0.736 (0.042)	0.730 (0.070)	0.751 (0.046)	0.902 (0.042)
	Rarefied counts	0.436 (0.032)	0.280 (0.060)	0.379 (0.066)	0.299 (0.044)	0.313 (0.031)

Supplementary Table 12 | PERSEO estimation: formulas and parameters. Only the location parameter μ depends on the covariates; the scale (σ) and shape (ν, τ) parameters are fitted as intercepts. Parametrizations follow gamlss.dist.

Quantity	Expression	Parameters
Transformed log-likelihood	$\log L(\theta; y) = \log L(\theta; z) + \Sigma_i \log\lvert \partial z_i/\partial y_i \rvert$	θ, distribution parameters; y, native scale; z, transformed scale
Jacobian term	$\log\lvert J_i \rvert = -\log(b - a)$ (unit interval); $-\log s$ (real-valued); 0 (count, positive)	a, b, feature minimum and maximum; s, feature standard deviation

Quantity	Expression	Parameters
Selection criterion	$IC = -2[\log L(\hat{\theta}; z) + \Sigma_i \log\lvert J_i \rvert] + k \cdot df$	k = 2 (AIC), log n (BIC) or k_{GAIC}; df, effective parameters; n, valid observations
Family choice	Family with minimum IC; ties within Δ = 2 broken by goodness of fit	Goodness of fit, correlation of sorted quantile residuals with normal quantiles
Coefficient test (Wald)	$\hat{V} = (X^T W X)^{-1}$; $t_j = \hat{\beta}_j / \sqrt{\hat{V}_{jj}}$	β, location coefficients; X, design; W, IRLS weights
Contrast	$\hat{\delta} = C\hat{\beta}$; $Var(\hat{\delta}) = diag(C\hat{V}C^T)$; $z = \hat{\delta}/se(\hat{\delta})$	C, contrast matrix
Omnibus test	$W = \hat{\beta}_S^T \hat{V}_S^{-1} \hat{\beta}_S \sim \chi^2(df)$; $LRT = 2(\ell_{full} - \ell_{reduced})$	S, factor coefficients; df = |S|
Multiplicity	$p_{adj} = BH(p)$, per term and per contrast	Benjamini–Hochberg procedure

Supplementary Table 13 | Simulation parameters by family. Baseline means were drawn per feature from the range shown (log-uniformly for count families, uniformly otherwise; for LOGNO the range is on the log scale). *σ* and *ν* were drawn uniformly per feature from the range of the stress level: low (baseline and batch scenarios), high (noisy and outlier scenarios) and zero inflation. For real-valued families *σ* is expressed as a fraction of the baseline mean. Meaning of *ν*: shape for GG, structural-zero probability for ZAGA and ZAIG, zero-inflation probability for ZINBI, mass at zero for BEINF and degrees of freedom for TF; for ZIP, *σ* is the zero-inflation probability. Single-cell profiles replaced *σ* of NBI, PIG and ZINBI by the measured cell-level ranges given in Supplementary Table 5.

Family	Effect encoding	Baseline mean	σ low	σ high	σ zero infl.	ν low	ν high	ν zero infl.
NBI	Multiplicative	0.5–800	0.05–0.25	0.30–1.00	0.30–0.80	–	–	–
PIG	Multiplicative	0.5–800	0.10–0.40	0.50–1.50	0.40–1.00	–	–	–
PO	Multiplicative	0.5–800	–	–	–	–	–	–
ZIP	Multiplicative	0.5–800	0.05–0.15	0.15–0.40	0.40–0.65	–	–	–
ZINBI	Multiplicative	0.5–800	0.05–0.25	0.30–1.00	0.30–0.80	0.05–0.15	0.15–0.40	0.40–0.65
GA	Multiplicative	5–200	0.20–0.50	0.50–1.20	0.40–0.90	–	–	–
GG	Multiplicative	5–200	0.20–0.50	0.50–1.20	0.40–0.90	0.5–1.5	0.8–2.5	0.8–2.0
IG	Multiplicative	5–150	0.20–0.50	0.50–1.50	0.40–1.00	–	–	–
ZAGA	Multiplicative	5–150	0.20–0.50	0.50–1.20	0.40–1.00	0.05–0.15	0.15–0.35	0.30–0.55
ZAIG	Multiplicative	5–150	0.20–0.50	0.50–1.50	0.40–1.00	0.05–0.15	0.15–0.35	0.30–0.55
LOGNO	Log-additive	log 5–log 200	0.20–0.50	0.40–1.00	0.40–0.90	–	–	–
BE	Logit-additive	0.05–0.60	0.10–0.30	0.20–0.50	0.20–0.50	–	–	–
BEINF	Logit-additive	0.05–0.50	0.10–0.30	0.20–0.50	0.20–0.50	0.02–0.10	0.05–0.20	0.30–0.50
NO	Additive (s.d.)	20–200	0.15–0.35	0.40–0.80	0.30–0.60	–	–	–
TF	Additive (s.d.)	20–200	0.15–0.35	0.40–0.80	0.30–0.60	6–12	3–6	3–6

Family	Effect encoding	Baseline mean	σ low	σ high	σ zero infl.	ν low	ν high	ν zero infl.
GU	Additive (s.d.)	20–200	0.15–0.35	0.40–0.80	0.30–0.60	–	–	–

Supplementary Table 14 | Software versions. Versions recorded in the manifests of the archived synthetic runs. The real-data benchmark ran on the same server and R installation; its archived session confirms R, PERSEO, DESeq2, edgeR and limma.

Software	Version	Use
R	4.5.3	All benchmark analyses
PERSEO	1.0.0	Proposed method
gamlss.dist	6.1.1	GAMLSS families used by PERSEO
DESeq2	1.48.2	Reference method, size factors
edgeR	4.6.3	Reference method, TMM, filterByExpr
limma	3.66.0	limma-voom and limma
vsn	3.76.0	Variance-stabilizing normalization
ANCOMBC	2.13.2	ANCOM-BC2
ALDEx2	1.40.0	Reference method
lefser	1.20.2	LEfSe
MSstats	4.21.0	Reference method
glmmTMB	1.1.14	Cell-level mixed model (single-cell)