AWmeta empowers adaptively-weighted transcriptomic meta-analysis

Yanshi Hu [1,*], **Zixuan Wang**[1], **Yueming Hu** [1], **Cong Feng**[1,2], **Qiuyu Fang**[3], **and Ming Chen** [1,2,*]

[1]Department of Bioinformatics, College of Life Sciences, Zhejiang University, Hangzhou 310058, China
[2]Biomedical Big Data Center, The First Affiliated Hospital, Zhejiang University School of Medicine; Institute of Hematology, Zhejiang University, Hangzhou 310058, China
[3]Department of Thoracic Surgery, The Second Affiliated Hospital Zhejiang University School of Medicine, Zhejiang University, Hangzhou 310058, China
[*]Corresponding author e-mail(s): yanshihu@zju.edu.cn; mchen@zju.edu.cn

ABSTRACT

Transcriptomic meta-analysis, a statistical method to synthesize multiple independent studies on a common topic, significantly enhances statistical power and facilitates the identification of consistent molecular characteristics, thereby improving reliability and reproducibility of biological discoveries. While many meta-analysis methods and tools have been developed for transcriptomic data, each manifests certain limitations, typically relying on either P-value combination or effect size integration alone. To address these limitations, we introduce AWmeta, a novel method for transcriptomic meta-analysis, which, for the first time, integrates the advantages of transcriptomic meta-analysis P-value and effect size methods, resulting in performance enhancement. By leveraging 35 transcriptomic datasets from Parkinson's and Crohn's disease, encompassing diverse tissue types, AWmeta exerts superior performance over existing state-of-the-art methods, including REM (Random Effect Model), across a range of key metrics. Specifically, AWmeta provides highly reliable differentially expressed genes (DEG) via accurate DEG discrimination in real and simulated data. Moreover, AWmeta results in gene difference quantification converging closer to theoretical baselines at both gene- and study-level compared to independent studies and REM. Importantly, AWmeta requires only a small sample size of independent studies to achieve substantial gene difference quantification effects, aiding in reducing experimental costs. We also observed that AWmeta requires fewer DEGs from integrated independent studies to achieve significant gene difference quantification effects, easing experimental operation and reducing researcher workload. While maintaining reliable differential gene quantification, AWmeta demonstrates robust performance against external interference and internal defects. In addition, AWmeta generates gene difference quantification results that offer more biologically meaningful insights than those from original independent studies and REM. Systematic evaluation demonstrates that AWmeta outperforms existing methods by achieving superior convergence, stability, and robustness with smaller sample size and fewer DEGs. Overall, AWmeta serves as a powerful and effective tool for transcriptomic meta-analysis.

Main

The exponential growth of publicly available transcriptomic data presents unprecedented opportunities for meta-analysis, a statistical approach that integrates findings across multiple independent studies to reveal consistent molecular signatures that individual studies often fail to detect[1]. This integration not only overcomes sample size limitations but also addresses the inherent variability in experimental design, tissue heterogeneity, and technical platforms that frequently obscure subtle yet biologically meaningful expression patterns. As complex diseases continue to challenge traditional single-study approaches, meta-analysis has emerged as a critical methodology for establishing reproducible biomarkers and pathways with enhanced confidence and precision[2].

Meta-analysis methods have evolved substantially over recent decades and are generally classified into three fundamental approaches[3]: P-value-based, effect-size-based, and rank-based integration. P-value methods such as Fisher's approach[4], Stouffer's Z-score technique[5], and the Adaptively Weighted Fisher's method[6] effectively combine statistical significance across studies but typically overlook effect magnitude. In contrast, effect-size methods including the Fixed Effects Model (FEM)[7] and Random Effects Model (REM)[8] quantify expression differences but can be sensitive to outliers and between-study heterogeneity. Rank-based approaches like RankProd and RankSum[9] exhibit robustness against outliers, though often at the expense of statistical precision. Despite their respective strengths, these methodological categories face substantial limitations when applied independently to complex transcriptomic datasets.

This methodological dichotomy creates a significant barrier in transcriptomic meta-analysis, particularly for complex

disease research where both statistical significance and effect magnitude are essential for biological interpretation. *P*-value integration excels at identifying consistently differentially expressed genes (DEGs) but provides minimal information on expression magnitude. Conversely, effect-size methods accurately quantify expression differences but frequently struggle with heterogeneity across diverse experimental designs and tissue types. Furthermore, no existing approach successfully incorporates the complementary advantages of multiple methodological categories, representing a critical gap in the field.

Here, we introduce AWmeta, a novel transcriptomic meta-analysis method that integrates the complementary strengths of *P*-value and effect size approaches. AWmeta applies an adaptively weighted strategy to identify and emphasize the most informative studies while accounting for between-study heterogeneity. Using 35 transcriptomic datasets from Parkinson's and Crohn's disease spanning diverse tissue types, we demonstrate that AWmeta outperforms existing state-of-the-art methods, including REM, across multiple performance metrics. AWmeta achieves superior differential gene detection with higher accuracy and resolution, while requiring fewer independent studies and DEGs to attain significant quantification effects. Furthermore, AWmeta demonstrates exceptional stability against external interference and internal data gaps, ultimately providing gene expression differential quantification results with enhanced biological relevance[10,11]. This advancement offers researchers a more powerful and efficient tool for transcriptomic meta-analysis, with potential applications across diverse biomedical research domains.

Results

Overview of AWmeta

AWmeta (Fig. 1a) is a novel integrative framework for transcriptomic meta-analysis, combining statistical strengths from *P*-value and effect size integration methods. Following preprocessing of multiple transcriptomic datasets from independent studies of the same disease tissue, AWmeta implements two complementary modules for each gene: AW-Fisher and AW-REM. The AW-Fisher module calculates meta-analysis *P*-values by optimizing weights to minimize the combined probability, effectively filtering less informative studies while preserving statistical power. Subsequently, the AW-REM module utilizes these optimized weights within a restricted maximum likelihood estimation (REML) process to estimate between-study variance (τ^2) and produce weighted effect sizes ($\log_2$FC)[12].

In the following sections, we detail a comprehensive evaluation of AWmeta's performance. We compared AWmeta primarily against REM, which is widely used and represents a high-performing method within the effect size integration category[13,14]. This evaluation, leveraging 35 transcriptomic datasets from Parkinson's and Crohn's disease tissues (1), assessed key metrics including DEG detection capability, gene-wise and study-wise convergence, stability, robustness, and biological relevance. The results consistently demonstrate AWmeta's superior performance across these metrics compared to REM, establishing its utility as a powerful tool for integrative transcriptomic analysis and providing a solid foundation for downstream biological investigation.

AWmeta identifies more DEGs with higher accuracy

A primary goal of transcriptomic meta-analysis is to enhance statistical power for identifying DEGs reliably, typically defined by applying statistical significance and fold-change thresholds (Fig. 1b). To benchmark AWmeta, we compared its performance against REM using the 35 transcriptomic datasets previously described, spanning five distinct disease-tissue contexts. Across all five contexts and systematically evaluating performance across nine combinations of statistical significance (*P*-value: 0.01, 0.05, 0.1) and fold-change thresholds ($|\log_2 \mathrm{FC}|$: $\log_2 1.2$, $\log_2 1.5$, $\log_2 2.0$), AWmeta consistently identified significantly more DEGs than REM (2; $P < 10^{-4}$, Wilcoxon signed-rank test over 100 bootstrap iterations, Methods). For instance, using representative thresholds ($P < 0.01$ and $|\log_2 \mathrm{FC}| > \log_2 1.2$), AWmeta yielded increases in detected DEGs ranging from 20% to over 300% compared to REM across the tissues (Fig. 1c). This demonstrates AWmeta's superior statistical power in DEG identification compared to REM.

While increased sensitivity is desirable, it necessitates evaluation of discrimination accuracy to control for potential false positives. We therefore employed a semi-synthetic simulation framework inspired by previous work[15], involving benchmark dataset generation with known positive (shared DEGs) and negative (shared non-DEGs) genes (Fig. 1d, Methods) and the introduction of controlled noise by permuting sample labels in a varying number of studies within each tissue context (Fig. 1e, Methods). We tested three noise levels corresponding to permuting the minimum, median, and maximum possible number of studies while aiming to retain informative signal from the non-permuted studies (Fig. 1f).

AWmeta consistently demonstrated superior or comparable discrimination accuracy to REM across the different noise levels and DEG thresholds tested. Under low noise conditions, created by permuting the minimum number of studies (Fig. 1g), AWmeta showed significantly better discrimination across most tissues and all nine DEG threshold combinations (3). Accuracy predictably decreased for both methods as more studies were permuted to introduce higher noise corresponding to median and maximum permutation levels; however, AWmeta generally maintained significantly better discrimination than REM across most conditions (Fig. 1h,i; 4, 5). Notably, AWmeta's discrimination remained robust under high noise conditions involving

maximum study permutation, with mean AUROC values often exceeding 0.8, highlighting its resilience to noise introduced by potentially problematic studies. This comprehensive analysis demonstrates that AWmeta achieves a superior balance between heightened detection sensitivity and robust discrimination accuracy for DEG identification from diverse transcriptomic datasets.

AWmeta demonstrates superior gene-wise and study-wise convergence in differential gene expression analysis

We next investigated AWmeta's ability to converge toward theoretical true values when integrating transcriptomic datasets. We developed quantitative metrics to assess convergence at both gene and study levels. Gene-wise convergence evaluates how closely the meta-analysis estimate for a specific gene aligns with the range of estimates from the individual studies contributing data for that gene. This was quantified using the Mean Absolute Deviation (MAD) between the meta-analysis $\log_2$ fold change and the corresponding $\log_2$ fold changes from the contributing individual studies (Fig. 2a, Methods). Lower MAD values indicate better approximation of the theoretical true biological signal.

We first applied this metric across all analyzed genes in the five disease tissues. In all five tissues, AWmeta achieved significantly lower gene-wise convergence scores compared to both REM and the baseline convergence observed among individual studies (Fig. 2b-f; $P < 10^{-4}$, one-tailed Mann-Whitney U test). Notably, while some individual studies occasionally outperformed REM in specific contexts, AWmeta consistently achieved lower scores than any individual study across all tissues. This suggests AWmeta robustly identifies and integrates reliable signals while effectively down-weighting contributions from divergent studies, thereby providing a superior consensus representation of the gene expression landscape.

Since disease processes are primarily driven by DEGs, we further evaluated gene-wise convergence specifically for DEGs identified using nine different threshold combinations (varying significance levels and $\log_2$FC cutoffs). Across all threshold combinations and disease tissues, AWmeta maintained significantly lower convergence scores (6a-e; P typically $< 10^{-4}$). This demonstrates its effectiveness in deriving reliable effect size estimates for disease-associated genes from diverse transcriptomic datasets, irrespective of the specific criteria used to define DEGs.

We next assessed study-wise convergence to evaluate how well the meta-analysis results represent the collective evidence across contributing studies. We employed three complementary approaches: a similarity metric based on ranked gene lists emphasizing top genes, the arithmetic mean of Jaccard and Overlap coefficients (JC/OC) for DEG set concordance, and the Phi (ϕ) correlation coefficient[16] assessing classification association beyond chance agreement (Fig. 2g,h; Methods). Higher scores indicate better study-wise convergence for all metrics.

Across the five disease tissues, AWmeta consistently achieved significantly higher study-wise convergence scores than the baseline representing typical inter-study agreement (Fig. 2k,l,m). Furthermore, AWmeta significantly outperformed REM in the majority (10 out of 15) of comparisons across different metrics and tissues, while performing comparably to REM in the remaining cases. This superior performance demonstrates AWmeta's ability to effectively address underlying variability by optimally weighting study contributions. Notably, while overall convergence scores tended to be lower in Parkinson's disease tissues compared to Crohn's disease tissues, potentially reflecting greater inherent variability within these specific disease contexts, AWmeta consistently provided the most convergent results relative to both the baseline and REM. This improvement was particularly pronounced in tissues like Parkinson's SN and Crohn's peripheral blood, where AWmeta's convergence scores improved by 25-30% compared to REM (Fig 2k-m). This enhanced ability to extract reliable biological signals from complex transcriptomic data is crucial for identifying subtle yet meaningful expression changes.

To further validate the robustness of these findings for the set-based methods, we confirmed AWmeta's superior performance using the JC/OC and ϕ coefficient metrics across nine different DEG threshold combinations (7a-j). AWmeta consistently achieved significantly higher scores than the baseline and outperformed REM in most evaluations across these varied thresholds. This indicates that the differential expression results processed by AWmeta are closer to the overall true convergence value compared to those from individual studies or REM alone.

AWmeta achieves superior meta-analysis convergence with fewer samples and DEGs

There is general agreement that transcriptomic meta-analyses benefit from larger datasets. However, achieving reliable meta-analysis of differential gene expression with fewer samples remains a key challenge. To address this, we investigated the association between study-wise convergence and sample size, employing Spearman correlation coefficients to capture potential non-linear relationships. For enhanced statistical power and reduced sensitivity to specific thresholds, this correlation analysis utilized the results derived from the nine different DEG threshold combinations (varying significance levels and $\log_2$FC cutoffs, as described previously) across both the JC/OC and the ϕ coefficient convergence metrics.

Significant positive correlations between sample size and study-wise convergence were found in Parkinson's SN, Parkinson's blood, and Crohn's IM (Fig. 3a–c) for both AWmeta and REM, indicating that larger sample sizes generally yield results closer to the theoretical true biological signals. However, when integrating studies with equivalent sample sizes, AWmeta consistently demonstrated superior convergence performance compared to REM across all five disease tissues. This performance advantage widened as sample size increased, reflected by steeper correlation slopes for AWmeta. For example, using the ranked similarity

metric in Crohn's IM (Fig. 3a), AWmeta and REM displayed similar convergence scores of approximately 0.38 at a sample size of 62. Yet, when the sample size increased to 200, AWmeta's convergence score increased to 0.77 while REM's reached 0.50. This translates to a 103% increase in convergence for AWmeta relative to its baseline score, compared to a 32% increase for REM, demonstrating AWmeta's greater capacity to derive benefit from larger sample sizes. As a result, AWmeta enables achievement of a target level of convergence with considerably fewer samples per study compared to REM. For instance, to reach a convergence score of 0.2 in Parkinson's SN, AWmeta required 15 samples using the Jaccard/Overlap coefficient metric (Fig. 3b) compared to 27 for REM (a 44% reduction in required sample size), and 20 samples using the ϕ correlation coefficient metric (Fig. 3c) compared to 35 for REM (a 43% reduction). This efficiency offers an advantage for reducing experimental costs.

In addition to sample size, we used the same correlation-based approach, again leveraging results from the nine DEG threshold combinations for the set-based metrics, to investigate the association between the number of identified DEGs and study-wise convergence. We found robust positive correlations between DEG number and convergence across all five disease tissues (Fig. 3d–f). Under these conditions, when datasets with similar DEG counts were considered, AWmeta achieved greater convergence than REM across the three complementary convergence metrics, indicating greater sensitivity to underlying biological signals. Furthermore, AWmeta exhibited steeper correlation slopes than REM, meaning its performance advantage becomes more pronounced as the number of DEGs increases, underscoring its capacity to detect gene expression differences more effectively. For instance, using the ranked similarity metric, the average correlation slope across tissues for AWmeta was 0.426, notably steeper than the average slope of 0.224 for REM (Fig. 3d). Importantly, this means AWmeta can achieve comparable convergence levels with significantly fewer DEGs than required by REM—a meaningful advantage for experimental design. This implies AWmeta can achieve reliable outcomes with less stringent experimental conditions (e.g., milder treatments, fewer replicates), reducing complexity and cost, particularly for large-scale studies.

AWmeta demonstrates superior stability and robustness in gene differential quantification

We next assessed whether AWmeta's optimized weighting strategy enhances the stability and robustness of differential gene quantification compared to REM. We employed quantitative metrics evaluating consistency against random data splits and resilience to data perturbations across the five disease tissues.

First, we evaluated stability by quantifying the consistency of ranked differential gene lists derived from randomly halved sample sets within each study, repeated over 100 iterations (Fig. 4a, Methods). Across all five disease tissues, AWmeta demonstrated significantly higher stability scores than REM (Fig. 4b; $P < 10^{-4}$, Wilcoxon signed-rank test)AWmeta generally showed tighter interquartile ranges in its stability scores, indicating more consistent performance across the random splits. Mean stability scores were below 0.6 for both methods, likely reflecting the reduced statistical power and study-wise convergence associated with halved sample sizes.

We then examined robustness against two types of perturbations: external interference via inclusion of an unrelated external study (Fig. 4c) and internal consistency via systematic removal of one study at a time (Fig. 4e). Against external interference, AWmeta exhibited strong resistance, maintaining mean robustness scores above 0.8 and significantly outperforming REM across all tissues (Fig. 4d; $P < 10^{-4}$). This resilience reflects AWmeta's adaptive weighting minimizing the influence of divergent studies. For internal robustness assessed by single-study removal, AWmeta also achieved significantly higher scores than REM in four of the five tissues (Fig. 4f; $P < 10^{-4}$). The exception was Crohn's disease blood, where the low number of contributing studies, three in total, limited robustness scores to below 0.7 for both methods. In tissues with five or more studies, AWmeta achieved mean internal robustness scores near 0.9, indicating highly consistent results despite individual study exclusions.

These results demonstrate that AWmeta's adaptive weighting strategy significantly enhances both the stability and robustness of DEG quantification. This ensures more reliable findings from meta-analyses involving collections of transcriptomic studies with varying characteristics.

AWmeta enhances detection of biologically relevant genes

We employed biocorrelation metrics to evaluate whether transcriptomic meta-analysis methods yield biologically meaningful gene quantification. Following established protocols (Fig. 6a), we assessed the biological relevance of differential expression results against three distinct disease benchmark gene sets for both Parkinson's disease (PD) and Crohn's disease (CD). The approach ranks genes based on both statistical significance (P-value) and effect size magnitude ($|\log_2 \mathrm{FC}|$), providing an objective evaluation independent of arbitrary differential expression thresholds (see Methods for details). The benchmark gene sets were compiled from the DisGeNET database (GDA score > 0.2)[17,18], the MalaCards database, and our in-house curated genetic variant associations.

To ensure the reliability of these inputs, the benchmark gene sets compiled for each disease were validated for internal consistency through pairwise comparisons. We quantified the magnitude of overlap using the Odds Ratio (OR) and determined its statistical significance via Fisher's exact test (FET). The results revealed significant overlap among all three benchmark gene

sets for both PD (e.g., OR=138.8, $P = 5.1 \times 10^{-115}$ for DisGeNET vs MalaCards) and CD (e.g., OR=242.9, $P = 1.3 \times 10^{-27}$ for DisGeNET vs MalaCards) (Fig. 6b). This mutual consistency supports the reliability of our selected disease benchmark gene sets.

We next compared the biological relevance scores from AWmeta and REM against a baseline reflecting the individual studies contributing to the meta-analysis. Kruskal-Wallis tests assessed overall differences, with Nemenyi post-hoc tests for pairwise comparisons. For the genetic variant gene set (Fig. 6c), AWmeta yielded significantly higher relevance scores than both baseline and REM across all assessed tissues. With the DisGeNET gene set (Fig. 6d), AWmeta also held significant advantages over baseline and REM in Parkinson's SN, Crohn's peripheral blood, and Crohn's IM. Similarly, using the MalaCards gene set (Fig. 6e), AWmeta similarly produced superior scores in Parkinson's SN, Crohn's IM, and Crohn's CM. Notably, AWmeta's enhanced performance in Parkinson's SN and Crohn's IM was consistent across all three gene sets. Overall, across the 15 comparisons, AWmeta scored significantly higher than baseline in 11 cases and likewise surpassed REM in 11 cases. Conversely, REM failed to significantly improve upon baseline scores in 11 of 15 comparisons. This highlights AWmeta's distinct capacity to enhance biological signal detection, an improvement not consistently offered by conventional meta-analysis over analyzing studies individually.

These findings demonstrate that AWmeta yields gene quantification with significantly greater biological relevance compared to both individual studies and REM. By effectively prioritizing established disease-associated genes, AWmeta reflects the underlying pathology more accurately and serves as a potent tool for uncovering disease-relevant transcriptomic alterations within complex datasets.

Discussion

Transcriptomic meta-analysis has become essential for extracting meaningful biological signals from heterogeneous expression data, yet existing methods face inherent limitations by relying exclusively on either *P*-value combination or effect size integration. This methodological dichotomy forces researchers to prioritize statistical significance detection over effect magnitude quantification, or vice versa. The field has long needed an approach that harnesses the complementary strengths of both frameworks without their respective limitations.

AWmeta represents the first successful integration of *P*-value aggregation and effect size methodologies in transcriptomic meta-analysis. The core innovation—a cross-module information transfer where optimized weights from *P*-value calculations directly enhance effect size estimation—effectively addresses between-study heterogeneity while maximizing consistent biological signal extraction. Indeed, the substantial variability often observed between studies, visually apparent in metrics like gene-wise convergence (e.g., Fig. 2b-f showing skewed distributions), highlights the prevalence of such heterogeneity and strongly supports the use of random-effects frameworks like REM and AWmeta over simpler fixed-effects models[19]. In our comprehensive evaluation across 35 datasets from Parkinson's and Crohn's disease, AWmeta demonstrated superior differential gene discrimination (AUROC: 0.89 vs 0.83 for REM, $P < 0.01$) that remained robust under substantial experimental noise. This enhanced discrimination capacity enabled identification of subtle yet biologically meaningful expression changes that conventional methods frequently miss, substantially improving the reliability and reproducibility of transcriptomic discoveries.

Our convergence metrics revealed AWmeta's practical advantages in approximating theoretical true values at both gene and study levels. It's noteworthy that in our gene-wise convergence assessments, some individual studies occasionally outperformed standard REM, even without larger sample sizes (6). While not conclusive, this hints that inherent study quality or specific experimental contexts might significantly influence reliability, perhaps as much as sample size itself. We also observed a tendency for these well-performing studies to utilize RNA-seq technology. These observations underscore the complexity of integrating diverse datasets and highlight the benefit of AWmeta's adaptive approach which weights studies based on informational content rather than relying solely on metrics like sample size. AWmeta achieved equivalent convergence with significantly fewer samples and DEGs than conventional methods—a critical advantage in resource-constrained research environments that can substantially reduce experimental costs and researcher workload. This efficiency was particularly evident in tissues with high heterogeneity, such as substantia nigra in Parkinson's disease, where convergence improved by 25-30% compared to REM. Furthermore, AWmeta maintained exceptional robustness against both external interference and internal data deletions, with stability values consistently exceeding 0.8 across all disease tissues, significantly outperforming existing approaches.

The biological relevance assessment using benchmark gene sets demonstrated AWmeta's capacity to generate results with enhanced biological meaningfulness compared to both REM and individual studies. Using disease-specific gene sets from DisGeNET, MalaCards, and genetic variant databases, AWmeta demonstrated significantly higher biological relevance scores in 11 of 15 tissue-benchmark combinations, providing researchers with more accurate representations of underlying disease mechanisms and potentially facilitating new insights into disease pathophysiology.

While AWmeta substantially advances transcriptomic meta-analysis, future development could extend its capabilities through integration with tissue-specific reference networks or multi-omics data types. Additionally, the methodology could be

expanded to facilitate drug target identification through integration with pharmacological databases. Furthermore, AWmeta presents a valuable opportunity for researchers to integrate their own, potentially limited-scale, datasets with larger public repositories, thereby leveraging collective evidence to enhance the statistical power and contextual relevance of their specific findings. Despite these future directions, AWmeta already represents a significant methodological advancement with immediate practical applications in biomedical research and biotechnology development, providing researchers with a powerful tool for extracting reliable biological signals from complex and often contradictory transcriptomic data—ultimately enhancing the transition from experimental biology to practical applications in clinical and industrial settings.

Methods

Overview of AWmeta Framework

The AWmeta framework performs gene-by-gene meta-analysis of transcriptomic data by integrating results from independent studies. It comprises two sequential modules: the AW-Fisher module for adaptive P-value integration and the AW-REM module for adaptive effect size integration. The computational workflow is illustrated in Fig. 1a. Input data for each gene typically consists of a P-value (P_i), a $\log_2$ fold change (FC_i), and its corresponding within-study variance (Var_i) derived from individual study analyses.

AW-Fisher Module (Adaptive P-value Integration)

This module computes a meta-analysis P-value for each gene using the Adaptive Weighted Fisher's method[?]. The analysis for a specific gene uses data only from the set S_{gene}, which includes studies providing data for that gene (studies lacking data, e.g., study 2 in Fig. 1 where $2 \notin S_{gene}$, are excluded). Within this set S_{gene}, the method identifies an optimal subset of studies by minimizing the combined P-value derived from the weighted Fisher's statistic.

Let $N' = |S_{gene}|$ be the number of studies providing data for the gene. The vector of available P-values from these studies is $\overrightarrow{\mathbf{P}} = (P_1, \ldots, P_i)_{i \in S_{gene}} \in (0,1)^{N'}$. The corresponding binary weight vector, indicating inclusion ($w_i = 1$) or exclusion ($w_i = 0$) of study $i \in S_{gene}$ in the evaluated subset, is $\overrightarrow{\mathbf{w}} = (\hat{w}_1, \ldots, \hat{w}_i)_{i \in S_{gene}} \in \{0,1\}^{N'}$.

The AW-Fisher statistic is defined as:

$$T(\overrightarrow{\mathbf{P}}; \overrightarrow{\mathbf{w}}) = -2 \sum_{i \in S_{gene}} w_i \ln P_i \tag{1}$$

The significance level of $T(\overrightarrow{\mathbf{P}}; \overrightarrow{\mathbf{w}})$ under the null hypothesis is calculated using the chi-squared distribution:

$$L(T(\overrightarrow{\mathbf{P}}; \overrightarrow{\mathbf{w}})) = 1 - F_{\chi^2_{d(\overrightarrow{\mathbf{w}})}}(T(\overrightarrow{\mathbf{P}}; \overrightarrow{\mathbf{w}})) \tag{2}$$

where the degrees of freedom are $d(\overrightarrow{\mathbf{w}}) = 2\sum_{i \in S_{gene}} w_i$, and $F_{\chi^2_d}(\cdot)$ is the cumulative distribution function (CDF) of the chi-squared distribution with d degrees of freedom.

The meta-analysis P-value, $s(\overrightarrow{\mathbf{P}})$, is the minimum significance level obtained by optimizing the weight vector over the studies in S_{gene}[?]:

$$s(\overrightarrow{\mathbf{P}}) = \min_{\overrightarrow{\mathbf{w}}} L(T(\overrightarrow{\mathbf{P}}; \overrightarrow{\mathbf{w}})) \tag{3}$$

The optimal weight vector $\hat{\boldsymbol{w}}$ that achieves this minimum is determined by:

$$\hat{\boldsymbol{w}} = w(\overrightarrow{\boldsymbol{P}}) = \underset{\overrightarrow{\mathbf{w}}}{\operatorname{argmin}} L(T(\overrightarrow{\mathbf{P}}; \overrightarrow{\mathbf{w}})) = (\hat{w}_1, \ldots, \hat{w}_i)_{i \in S_{gene}} \tag{4}$$

This optimal weight vector $\hat{\boldsymbol{w}}$, containing binary weights for each study in S_{gene}, is passed to the AW-REM module.

AW-REM Module (Adaptive Effect Size Integration)

This module calculates the meta-analysis effect size ($\log_2$FC) using an adaptively weighted REM. It leverages the $\log_2$FC (FC_i) and variance (Var_i) data from studies in S_{gene}, modulated by the adaptive weights $\hat{\boldsymbol{w}}$ derived from the AW-Fisher module for those same studies. The contribution weight for study $i \in S_{gene}$ in the AW-REM is defined as:

$$W_i = \frac{\hat{w}_i}{Var_i + T^2} \tag{5}$$

where $\hat{w}_i$ is the binary weight (0 or 1) for study $i \in S_{gene}$ obtained from Equation 4. If the AW-Fisher module assigned $\hat{w}_i = 0$ to study i, then $W_i = 0$, effectively removing that study's influence from the effect size calculation despite its data being available in S_{gene}. Var_i is the within-study variance for the gene in study i, and T^2 represents the between-study variance. T^2 is estimated using the Restricted Maximum Likelihood (REML) method[12], incorporating data only from studies $i \in S_{gene}$ for which $\hat{w}_i = 1$.

The final meta-analysis $\log_2$FC, denoted as $\mathbf{M}$, is the weighted arithmetic mean calculated using the effect sizes and AW-REM weights for all studies in S_{gene}:

$$\mathbf{M} = \frac{\sum_{i \in S_{gene}} W_i FC_i}{\sum_{i \in S_{gene}} W_i} \tag{6}$$

The summation implicitly includes only studies from S_{gene} where $\hat{w}_i = 1$, as W_i is zero otherwise. Studies initially excluded because they were not in S_{gene} (like study 2 in the example) are entirely absent from this calculation.

Transcriptomic datasets

To proof-of-concept the AWmeta framework, we compiled 35 publicly available human transcriptomic datasets for Parkinson's and Crohn's disease from the Gene Expression Omnibus (GEO)[20], Sequence Read Archive (SRA)[21], and ArrayExpress[22] databases. These datasets, encompassing both microarray and RNA-sequencing (RNA-seq) platforms, included samples derived from Parkinson's substantia nigra (SN), Parkinson's peripheral blood, Crohn's ileal mucosa (IM), Crohn's colonic mucosa (CM)[23,24], and Crohn's peripheral blood. A complete list of the datasets, detailing accession numbers, platform information, and sample sizes, is provided in Extended Data Fig. 1.

Transcriptome data preprocessing

Due to the inclusion of datasets generated on different platforms, specific preprocessing pipelines were applied separately to microarray and RNA-seq data.

Microarray data processing

To ensure accurate and up-to-date probe annotations, microarray probe identifiers were mapped to Entrez Gene IDs using information retrieved from GEO SOFT files, platform-specific Bioconductor annotation packages, and the AnnoProbe R package. For genes represented by multiple probes, the probe exhibiting the largest Inter-Quartile Range (IQR) of intensity values across samples within a study was selected to represent the gene's expression level, a strategy demonstrated to optimize biological relevance[13,25]. Subsequently, the limma R package[26] was utilized for microarray data preprocessing, normalization, and differential gene identification.

RNA-seq data processing

An automated workflow for RNA-seq data analysis was implemented using snakemake[27]. Raw sequencing reads were processed with Trimmomatic[28] to remove adapter sequences and low-quality bases. Following best practice recommendations[29], the cleaned reads were aligned to the human reference genome (GRCh38 assembly) using HISAT2[30]. Gene-level read counts were quantified from the aligned reads using featureCounts[31]. Finally, the DESeq2 R package[32] was used for gene difference quantification.

Evaluation Metrics for Meta-Analysis Methods for Transcriptome Data

To comprehensively assess the performance of AWmeta and compare it with standard approaches, we applied several quantitative metrics. These metrics were designed to measure key aspects of meta-analysis efficacy using the compiled transcriptomic datasets. The REM[3,13,14] was used as a benchmark method for comparison in these evaluations.

Ability to detect differential genes

DEG Detection Capability was defined as the number of DEGs identified by each meta-analysis method. The identification of DEGs required genes to satisfy pre-defined thresholds for both statistical significance (*P*-value) and fold change magnitude ($\log_2$FC) (Fig.1b). To assess the stability and reliability of this capability, we implemented a bootstrap resampling strategy with 100 iterations. In each iteration, we created bootstrapped datasets by randomly sampling with replacement from the original case and control groups while maintaining the original sample sizes. The meta-analysis method was then applied to these bootstrapped datasets, and the distribution of identified DEG counts was analyzed to evaluate the robustness of detection.

Evaluation of DEG Discrimination Capacity using Semi-synthetic Data

To evaluate the ability to discriminate between DEGs and non-DEGs, particularly considering potential false positives arising from higher detection sensitivity, we utilized an evaluation metric based on semi-synthetic simulated data, inspired by previous work[15]. This approach consisted of benchmark dataset generation and evaluation using datasets with simulated noise (Fig.1c).

For each tissue context, we generated an unbiased benchmark dataset as follows: (1) DEGs and non-DEGs were determined for both AWmeta and REM using predefined screening thresholds; (2) for fairness, an unbiased positive sample set was created by randomly sampling half of the DEGs identified by both methods; and (3) an unbiased negative sample set, equal in size to the positive set, was created by randomly sampling from the non-DEGs identified by both methods.

We then evaluated discrimination capacity using semi-synthetic datasets. Sample labels (case/control) were randomly permuted within selected independent studies for a given tissue context (e.g., studies 1 and 3 as illustrated in Fig.1c). Theoretically, these label-permuted studies would no longer contain true DEGs. AWmeta and REM were applied to these semi-synthetic datasets (containing both original and label-permuted studies). The Area Under the Receiver Operating Characteristic curve (AUROC) and the Area Under the Precision-Recall curve (AUPRC) metrics were calculated based on the meta-analysis results for the benchmark genes, quantifying the DEG discrimination capacity. Higher AUROC and AUPRC values indicate stronger discrimination capacity. This entire procedure was repeated 100 times to obtain distributions of these metrics, ensuring the stability and reliability of our assessment.

Gene-wise convergence assessment for differential gene expression

To evaluate the convergence of meta-analysis effect sizes at the individual gene level, we calculated the Mean Absolute Deviation (MAD) between the meta-analysis $\log_2$ fold change (FC_{meta}) and the $\log_2$FC values from individual studies contributing data for that gene. For a given gene, let FC_{meta} be the meta-analysis $\log_2$FC (e.g., calculated using Equation 6) and FC_i be the $\log_2$FC from study i. Let S_{gene} denote the set of studies providing data for this specific gene (as defined in the AWmeta Framework section). The Gene-wise Convergence score (C_{gene}) for that gene relative to the meta-analysis result was defined using only studies within S_{gene}:

$$C_{gene} = \frac{1}{|S_{gene}|} \sum_{i \in S_{gene}} |FC_{meta} - FC_i| \tag{7}$$

where $|S_{gene}|$ is the number of studies in the set S_{gene}. A lower C_{gene} value indicates better convergence, indicating closer agreement between the meta-analysis estimate and individual study values within S_{gene}.

For baseline comparison, we also calculated gene-wise convergence relative to each individual study j within S_{gene} acting as a reference point:

$$C^j_{gene} = \frac{1}{|S_{gene}|} \sum_{i \in S_{gene}} |FC_j - FC_i| \quad \text{for } j \in S_{gene} \tag{8}$$

This approach allowed systematic comparison of AWmeta's convergence against other methods and individual studies, considering only the relevant set of studies (S_{gene}) for each gene.

Study-wise convergence assessment for differential gene expression

To assess the overall consistency between the complete gene list generated from a meta-analysis method and the individual contributing studies, three complementary approaches were employed. For all approaches, higher scores indicate better study-level convergence.

Ranked gene list similarity metric

First, we developed a similarity metric based on ranked gene lists (Fig. 4-10, 4-11A). For both the meta-analysis and each individual study, genes were assigned a composite rank derived from the product of their P-value rank (ascending) and absolute $\log_2$FC rank (descending). Genes were then ordered according to this composite rank (ascending) to create lists (e.g., G_{meta} for meta-analysis, G_i for study i).

The weighted similarity $S(G_{meta}, G_i)$ between the meta-analysis and each individual study gene lists (containing N genes) was computed using a non-linear weighting scheme[33], which assigns greater emphasis to concordance among top-ranked genes:

$$S(G_{meta}, G_i) = \sum_{n=1}^{N} e^{-\alpha n} O_n(G_{meta}, G_i) \tag{9}$$

where $O_n(G_{meta}, G_i)$ is the number of identical genes in the top n positions and α is a weighting exponent (0.001). This score was then normalized[14] to the interval [-1, 1] yielding the adjusted similarity $S_{adj}(G_{meta}, G_i)$:

$$S_{adj}(G_{meta}, G_i) = \frac{S(G_{meta}, G_i) - E_{\text{null}}(S(G_{meta}, G_i))}{\max(S(G_{meta}, G_i)) - E_{\text{null}}(S(G_{meta}, G_i))} \tag{10}$$

where $E_{\text{null}}(S(G_{meta}, G_i)) = \sum_{n=1}^{N} \frac{n^2}{N} e^{-\alpha n}$ and $\max(S(G_{meta}, G_i)) = \sum_{n=1}^{N} n e^{-\alpha n}$ are the expected and maximum scores under a null hypothesis of random gene lists.

As a baseline comparison, we sequentially designated each individual study list G_j as the reference point and calculated $S_{adj}(G_j, G_i)$ between it and all other study lists G_i (where $j \neq i$).

Set-based overlap metrics

Second, we assessed similarity using set-based overlap metrics that ignore rank order (Fig. 4-11B, 4-11C). Differentially Expressed Gene (DEG) sets were determined for the meta-analysis (Set_{meta}) and each individual study (Set_i) using defined significance thresholds.

We calculated two complementary metrics: the Jaccard Coefficient ($JC_i = |Set_{meta} \cap Set_i| / |Set_{meta} \cup Set_i|$) and the Overlap Coefficient ($OC_i = |Set_{meta} \cap Set_i| / \min(|Set_{meta}|, |Set_i|)$). The convergence metric for study i relative to the meta-analysis was the arithmetic mean $(JC_i + OC_i)/2$.

As a baseline comparison, we sequentially designated each individual study's DEG set Set_j as the reference point and calculated the same metrics between it and all other study DEG sets Set_i (where $j \neq i$).

This approach addresses the limitations of the first method by being insensitive to gene ordering, focusing instead on whether genes are classified as DEGs or not, regardless of their rankings.

Phi correlation coefficient

Third, we measured the association between DEG classifications using the ϕ correlation coefficient (Phi Coefficient, PC)[16] (Fig. 4-11D). This approach considers the extreme case where shared DEGs or non-DEGs between two gene sets might be randomly generated.

For each comparison between the meta-analysis and an individual study i, we constructed a 2×2 contingency table categorizing all genes as DEG or non-DEG in both datasets. The ϕ coefficient (ϕ_i) was then calculated as:

$$\phi_i = \frac{n_{11}n_{22} - n_{12}n_{21}}{\sqrt{n_{1\cdot}n_{2\cdot}n_{\cdot 1}n_{\cdot 2}}} \tag{11}$$

where n_{11} represents genes that are DEGs in both datasets, n_{22} represents genes that are non-DEGs in both datasets, and n_{12} and n_{21} represent genes classified differently between datasets. The row and column sums are denoted by $n_{1\cdot}$, $n_{2\cdot}$, $n_{\cdot 1}$, and $n_{\cdot 2}$.

As a baseline comparison, we sequentially designated each individual study's classification as the reference point and calculated the ϕ coefficient between it and all other studies.

While the first method can process all genes together, the second and third methods require separate calculations for DEGs and non-DEGs. However, since DEGs and non-DEGs are mutually exclusive events, results from either category will be complementary, and typically only one needs to be reported. Across all three methods, higher values indicate better study-level convergence and closer approximation to the overall converged true values.

Stability Assessment of Differential Gene Quantification

The stability of meta-analysis results was assessed through consistency evaluation under random within-study subsampling. For each independent study in the original cohort, case and control samples were randomly partitioned into two approximately equal halves. This process generated two corresponding sets of sub-studies. DEG analysis was performed individually for each sub-study. Subsequently, the transcriptomic meta-analysis procedure was applied independently to each of the two sets of sub-studies. The similarity between the resulting ordered lists of DEGs from the two meta-analyses was quantified using the adjusted differential gene list similarity measure (defined in Equation 10). This entire procedure, from sample splitting to similarity calculation, was repeated 100 times to assess a distribution of stability scores.

External and Internal Robustness Assessment of Differential Gene Quantification

We assessed meta-analysis robustness against two perturbation types. First, external robustness was assessed by adding a single, independent external study to the original dataset. Meta-analysis was performed on both the original and expanded datasets, and the similarity between the resulting ranked lists of DEGs was calculated using an adjusted similarity measure (Equation 10). This procedure quantifies the method's resilience to the inclusion of potentially disparate data. Second, internal robustness was assessed by removing one study at a time from the original dataset. The DEG list from the meta-analysis of the reduced dataset was compared to the original DEG list using the same similarity measure (Equation 10), evaluating the influence of each individual study. Repeating these processes—adding different external studies or removing each internal study in turn—yielded distributions of external and internal robustness scores.

Biological Relevance Assessment of Differential Gene Quantification

We evaluated the biological relevance of meta-analysis results using benchmark gene sets for Parkinson's (PD) and Crohn's disease (CD). Benchmark genes were sourced from: 1) DisGeNET[34] with Gene-Disease Association (GDA) score > 0.2[17,18]; 2) the MalaCards[35]; and 3) our in-house curated disease-linked genetic variations .

For each analysis result (meta-analysis or individual study), we first ranked all analyzed genes based on ascending P-value and, separately, on descending absolute $\log_2$ fold change ($|\log_2 \mathrm{FC}|$). We then identified the rank of each benchmark gene within these complete rankings to assign its $\mathrm{Rank}_{P_\mathrm{value}}$ and $\mathrm{Rank}_{|\log\mathrm{FC}|}$. An integrated rank for each benchmark gene was calculated as:

$$\text{Integrated Rank} = \mathrm{Rank}_{|\log\mathrm{FC}|} \times \mathrm{Rank}_{P_\mathrm{value}} \tag{12}$$

Benchmark genes were then re-ranked according to this Integrated Rank (ascending) to obtain $\mathrm{Rank}_{\text{Integrated Rank}}$. The Biological Relevance score was calculated for each benchmark gene as:

$$\text{Biological Relevance} = 1 - \frac{\mathrm{Rank}_{\text{Integrated Rank}}}{N_{benchmark}} \tag{13}$$

where $N_{benchmark}$ is the size of the benchmark set. Higher scores reflect greater biological relevance, indicating that benchmark genes ranked better overall by combining statistical significance and effect size. This rank-based score avoids arbitrary DEG thresholds. We compared score distributions from AWmeta, REM, and individual studies using Kruskal-Wallis and Nemenyi tests.

Data availability

All transcriptomic datasets in this study are publicly available, accessible on GEO, SRA and ArrayExpress, spanning substantial nigra (GSE114517, GSE8397, GSE20163, GSE20164, GSE20292, GSE7621, GSE49036, GSE42966, GSE43490, GSE26927, GSE54282) and periphery blood (GSE57475, GSE54536, GSE34287, GSE99039, GSE72267, GSE6613, GSE18838, GSE165082) from Parkinson's disease, and periphery blood (GSE119600, GSE112057, GSE94648), ileal (GSE102133, GSE75214, GSE16879, GSE68570, GSE101794, GSE57945) and colonic mucosa (GSE75214, GSE16879, GSE36807, GSE4183, GSE9686, GSE66207, E-MTAB-184) from Crohn's disease. Benchmark gene sets for Parkinson's and Crohn's disease were downloaded from DisGeNET (https://disgenet.com/) and MalaCards (https://www.malacards.org/).

Code availability

AWmeta is available on GitHub at https://github.com/YanshiHu/AWmeta.

References

1. Shorten, A. & Shorten, B. What is meta-analysis? *Evid. Based Nurs.* **16**, 3–4 (2013).
2. Zhong, G., Bolitho, S., Grunstein, R., Naismith, S. L. & Lewis, S. J. G. The relationship between thermoregulation and rem sleep behaviour disorder in parkinson's disease. *PLoS One* **8**, e72661 (2013).
3. Makinde, F. L. *et al.* Reviewing and assessing existing meta-analysis models and tools. *Briefings Bioinforma.* **22**, bbab324 (2021).
4. Fisher, R. A. Statistical methods for research workers. In *Breakthroughs in statistics: Methodology and distribution*, 66–70 (Springer, 1970).
5. Stouffer, S. A. A study of attitudes. *Sci. Am.* **180**, 11–15 (1949).
6. Li, J. & Tseng, G. C. An adaptively weighted statistic for detecting differential gene expression when combining multiple transcriptomic studies. *Annals Appl. Stat.* **5**, 994–1019 (2011).
7. Cochran, W. G. The combination of estimates from different experiments. *Biometrics* **10**, 101 (1954).
8. Wang, X. *et al.* Comparing methods for performing trans-ethnic meta-analysis of genome-wide association studies. *Hum. molecular genetics* **22**, 2303–2311 (2013).
9. Breitling, R., Armengaud, P., Amtmann, A. & Herzyk, P. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Lett.* **573**, 83–92 (2004).
10. Piras, I. S. *et al.* Peripheral biomarkers in schizophrenia: a meta-analysis of microarray gene expression datasets. *Int. J. Neuropsychopharmacol.* **22**, 186–193 (2019).
11. Cho, H. *et al.* Meta-analysis method for discovering reliable biomarkers by integrating statistical and biological approaches: An application to liver toxicity. *Biochem. Biophys. Res. Commun.* **471**, 274–281 (2016).
12. Kenward, M. G. & Roger, J. H. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics* **53**, 983–997 (1997).
13. Ma, T. *et al.* Metaomics: analysis pipeline and browser-based software suite for transcriptomic meta-analysis. *Bioinformatics* **35**, 1597–1599 (2019).
14. Chang, L.-C., Lin, H.-M., Sibille, E. & Tseng, G. C. Meta-analysis methods for combining multiple expression profiles: comparisons, statistical characterization and an application guideline. *BMC Bioinforma.* **14**, 368 (2013).
15. 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).
16. Akoglu, H. User's guide to correlation coefficients. *Turkish J. Emerg. Medicine* **18**, 91–93 (2018).
17. Yuan, X. *et al.* The common genes involved in the pathogenesis of alzheimer's disease and type 2 diabetes and their implication for drug repositioning. *Neuropharmacology* **223**, 109327 (2023).

18. Ferreira, J. C., Alshamali, F., Pereira, L. & Fernandes, V. Characterization of arabian peninsula whole exomes: contributing to the catalogue of human diversity. *iScience* **25**, 105336 (2022).

19. Borenstein, M., Hedges, L. V., Higgins, J. P. & Rothstein, H. R. A basic introduction to fixed-effect and random-effects models for meta-analysis. *Res. synthesis methods* **1**, 97–111 (2010).

20. Barrett, T. *et al.* Ncbi geo: archive for functional genomics data sets—update. *Nucleic Acids Res.* **41**, D991–D995 (2012).

21. Sayers, E. W. *et al.* Database resources of the national center for biotechnology information. *Nucleic Acids Res.* **47**, D23–D28 (2019).

22. Kolesnikov, N. *et al.* Arrayexpress update—simplifying data submissions. *Nucleic Acids Res.* **43**, D1113–D1116 (2015).

23. Atreya, R. & Siegmund, B. Location is important: differentiation between ileal and colonic crohn's disease. *Nat. Rev. Gastroenterol. Hepatol.* **18**, 544–558 (2021).

24. Trist, B. G., Hare, D. J. & Double, K. L. Oxidative stress in the aging substantia nigra and the etiology of parkinson's disease. *Aging Cell* **18**, e13031 (2019).

25. Gentleman, R., Carey, V., Huber, W., Irizarry, R. & Dudoit, S. *Bioinformatics and computational biology solutions using R and Bioconductor* (Springer Science & Business Media, 2005).

26. Ritchie, M. E. *et al.* limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47–e47 (2015).

27. Köster, J. & Rahmann, S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* **28**, 2520–2522 (2012).

28. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

29. Sahraeian, S. M. E. *et al.* Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum rna-seq analysis. *Nat. Commun.* **8**, 59 (2017).

30. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with hisat2 and hisat-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).

31. Liao, Y., Smyth, G. K. & Shi, W. featurecounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).

32. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biol.* **15** (2014).

33. Yang, X., Bentink, S., Scheid, S. & Spang, R. Similarities of ordered gene lists. *J. Bioinforma. Comput. Biol.* **4**, 693–708 (2011).

34. Piñero, J. *et al.* The disgenet knowledge platform for disease genomics: 2019 update. *Nucleic acids research* **48**, D845–D855 (2020).

35. Rappaport, N. *et al.* Malacards: an amalgamated human disease compendium with diverse clinical and genetic annotation and structured search. *Nucleic acids research* **45**, D877–D887 (2017).

36. Lewis, P. *et al.* Evidence for immune response, axonal dysfunction and reduced endocytosis in the substantia nigra in early stage parkinson's disease. *PLoS One* **10**, e0128651 (2015).

37. Cookson, M. R. *et al.* Systems-based analyses of brain regions functionally impacted in parkinson's disease reveals underlying causal mechanisms. *PLoS ONE* **9**, e102909 (2014).

38. Simchovitz, A. *et al.* A lncrna survey finds increases in neuroprotective linc-pint in parkinson's disease substantia nigra. *Aging Cell* **19**, e13115 (2020).

39. Duke, D. C., Moran, L. B., Pearce, R. K. B. & Graeber, M. B. The medial and lateral substantia nigra in parkinson's disease: mrna profiles associated with higher brain tissue vulnerability. *Neurogenetics* **8**, 83–94 (2007).

40. Leal, S. M. *et al.* A genomic pathway approach to a complex disease: axon guidance and parkinson disease. *PLoS Genet.* **3**, e98 (2007).

41. Corradini, B. R. *et al.* Complex network-driven view of genomic mechanisms underlying parkinson's disease: analyses in dorsal motor vagal nucleus, locus coeruleus, and substantia nigra. *BioMed Res. Int.* **2014**, 1–16 (2014).

42. Zheng, B. *et al.* Pgc-1alpha, a potential therapeutic target for early intervention in parkinson's disease. *Sci Transl Med* **2**, 52ra73 (2010).

43. Durrenberger, P. F. *et al.* Selection of novel reference genes for use in the human central nervous system: a brainnet europe study. *Acta Neuropathol.* **124**, 893–903 (2012).

44. Cookson, M. R. *et al.* Srrm2, a potential blood biomarker revealing high alternative splicing in parkinson's disease. *PLoS One* **5**, e9104 (2010).

45. Scherzer, C. R. *et al.* Gata transcription factors directly regulate the parkinson's disease-linked gene -synuclein. *Proc. Natl. Acad. Sci.* **105**, 10907–10912 (2008).

46. Calligaris, R. *et al.* Blood transcriptomics of drug-naïve sporadic parkinson's disease patients. *BMC Genomics* **16**, 876 (2015).

47. Mosley, R. L. *et al.* Biosignatures for parkinson's disease and atypical parkinsonian disorders patients. *PLoS One* **7**, e43595 (2012).

48. Locascio, J. J. *et al.* Association between -synuclein blood transcripts and early, neuroimaging-supported parkinson's disease. *Brain* **138**, 2659–2671 (2015).

49. Alieva, A. K. *et al.* Involvement of endocytosis and alternative splicing in the formation of the pathological process in the early stages of parkinson's disease. *BioMed Res. Int.* **2014**, 1–6 (2014).

50. Shamir, R. *et al.* Analysis of blood-based gene expression in idiopathic parkinson disease. *Neurology* **89**, 1676–1683 (2017).

51. Henderson, A. R. *et al.* Dna methylation and expression profiles of whole blood in parkinson's disease. *Front. Genet.* **12**, 640266 (2021).

52. Haberman, Y. *et al.* Pediatric crohn disease patients exhibit specific ileal transcriptome and microbiome signature. *J. Clin. Investig.* **124**, 3617–3633 (2014).

53. Haberman, Y. *et al.* Age-of-diagnosis dependent ileal immune intensification and reduced alpha-defensin in older versus younger pediatric crohn disease patients despite already established dysbiosis. *Mucosal Immunol.* **12**, 491–502 (2019).

54. Pérez-Torras, S. *et al.* Transportome profiling identifies profound alterations in crohn's disease partially restored by commensal bacteria. *J. Crohn's Colitis* **10**, 850–859 (2016).

55. Verstockt, S. *et al.* Gene and mirna regulatory networks during different stages of crohn's disease. *J. Crohn's Colitis* **13**, 916–930 (2019).

56. Timmer, A. *et al.* Mucosal gene expression of antimicrobial peptides in inflammatory bowel disease before and after first infliximab treatment. *PLoS One* **4**, e7984 (2009).

57. Vancamelbeke, M. *et al.* Genetic and transcriptomic bases of intestinal epithelial barrier dysfunction in inflammatory bowel disease. *Inflamm. Bowel Dis.* **23**, 1718–1729 (2017).

58. van Beelen Granlund, A. *et al.* Reg gene expression in inflamed and healthy colon mucosa explored by in situ hybridisation. *Cell Tissue Res.* **352**, 639–646 (2013).

59. Carey, R. *et al.* Activation of an il-6:stat3-dependent transcriptome in pediatric-onset inflammatory bowel disease. *Inflamm. Bowel Dis.* **14**, 446–457 (2008).

60. Galamb, O. *et al.* Inflammation, adenoma and cancer: objective classification of colon biopsy specimens with gene expression signature. *Dis. Markers* **25**, 1–16 (2008).

61. Calogero, R. A. *et al.* Identification of novel predictor classifiers for inflammatory bowel disease by gene expression profiling. *PLoS One* **8**, e76235 (2013).

62. Peck, B. C. E. *et al.* Micrornas classify different disease behavior phenotypes of crohns disease and may have prognostic utility. *Inflamm. Bowel Dis.* **21**, 2178–2187 (2015).

63. Planell, N. *et al.* Usefulness of transcriptional blood biomarkers as a non-invasive surrogate marker of mucosal healing and endoscopic response in ulcerative colitis. *J. Crohn's Colitis* **11**, 1335–1346 (2017).

64. Mo, A. *et al.* Disease-specific regulation of gene expression in a comparative analysis of juvenile idiopathic arthritis and inflammatory bowel disease. *Genome Medicine* **10**, 48 (2018).

65. Ostrowski, J. *et al.* Common functional alterations identified in blood transcriptome of autoimmune cholestatic liver and inflammatory bowel diseases. *Sci. Reports* **9**, 7190 (2019).

Acknowledgements

The authors thank all members from Ming Chen's Group of Bioinformatics for insightful discussions. This work was supported by the National Key Research and Development Program of China (2023YFE0112300); National Natural Sciences Foundation of China (32261133526; 32270709; 32070677); the 151 talent project of Zhejiang Province (first level), the Science and Technology Innovation Leading Scientist (2022R52035).

Author contributions

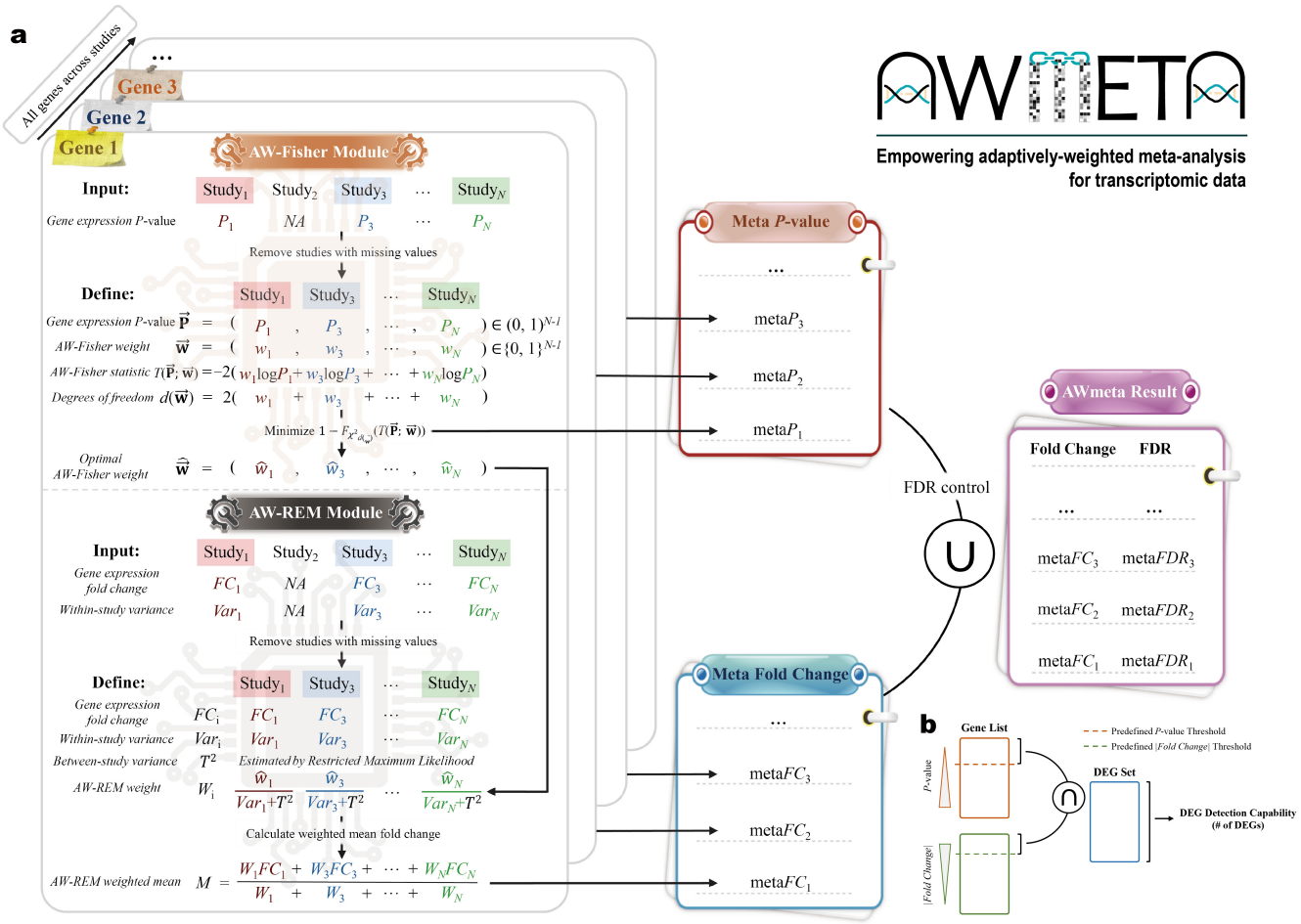
Y.S.H. conceptualized the study. M.C. supervised the study. Y.S.H., Z.W., and Y.H. performed data analysis. Y.S.H., Z.W., C.F., and Q.F. contributed to writing the manuscript. All authors reviewed and approved the final manuscript.

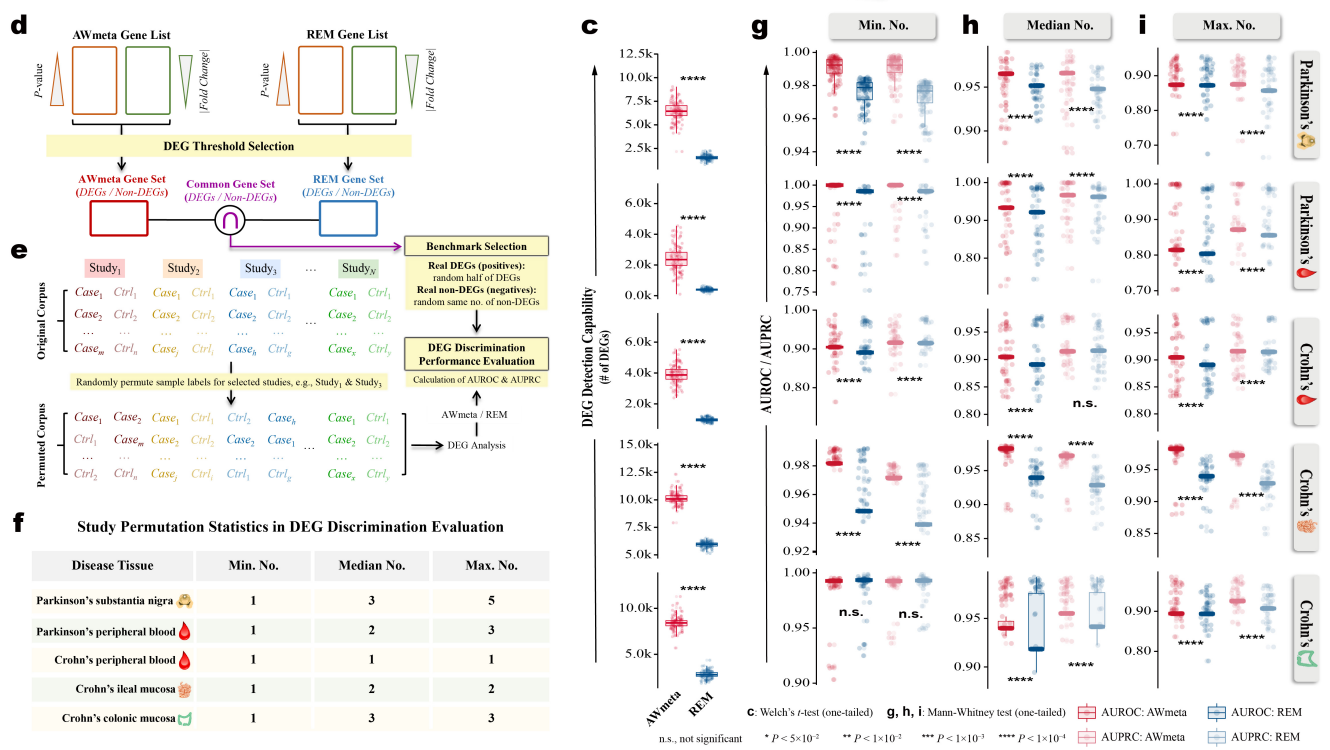
Competing interests

The authors declare no competing interests.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.





f Study Permutation Statistics in DEG Discrimination Evaluation

Disease Tissue	Min. No.	Median No.	Max. No.
Parkinson's substantia nigra	1	3	5
Parkinson's peripheral blood	1	2	3
Crohn's peripheral blood	1	1	1
Crohn's ileal mucosa	1	2	2
Crohn's colonic mucosa	1	3	3

Fig. 1 | Overview of AWmeta and DEG identification evaluation. a, Schematic of the AWmeta framework, comprising the AW-Fisher module for *P*-value calculation and the AW-REM module for effect size ($\log_2$ Fold Change) estimation. DEG identification evaluation includes DEG detection capability (**b, c**) and discrimination (**d-i**). **b,** Schematic defining DEG detection capability as the number of DEGs identified based on predefined statistical significance (*P*-value) and gene expression fold change ($\log_2$ based) thresholds. **c,** DEG detection capability comparison between AWmeta and REM ($P < 0.01$, $|\log_2 FC| > \log_2 1.2$). **d,** Strategy for generating the semi-synthetic benchmark dataset, sampling DEGs and non-DEGs from common genes identified by both AWmeta and REM. **e,** Workflow for evaluating DEG discrimination performance using label permutation within the semi-synthetic benchmark dataset, followed by AWmeta/REM analysis and calculation of AUROC and AUPRC. **f,** Study permutation statistics (number of studies permuted) in the DEG discrimination evaluation across the five disease-tissue datasets for simulations varying permutation count. **g,** DEG discrimination performance comparison between AWmeta and REM using minimum-permuted semi-synthetic simulation strategy ($P < 0.01$, $|\log_2 FC| > \log_2 1.2$). **h,** DEG discrimination performance comparison between AWmeta and REM using median-permuted semi-synthetic simulation strategy ($P < 0.01$, $|\log_2 FC| > \log_2 1.2$). **i,** DEG discrimination performance comparison between AWmeta and REM using maximum-permuted semi-synthetic simulation strategy ($P < 0.01$, $|\log_2 FC| > \log_2 1.2$). Icons represent tissue sources: 🧠: substantia nigra (SN); 🩸: peripheral blood; 🍿: ileal mucosa (IM); 🧫: colonic mucosa (CM).

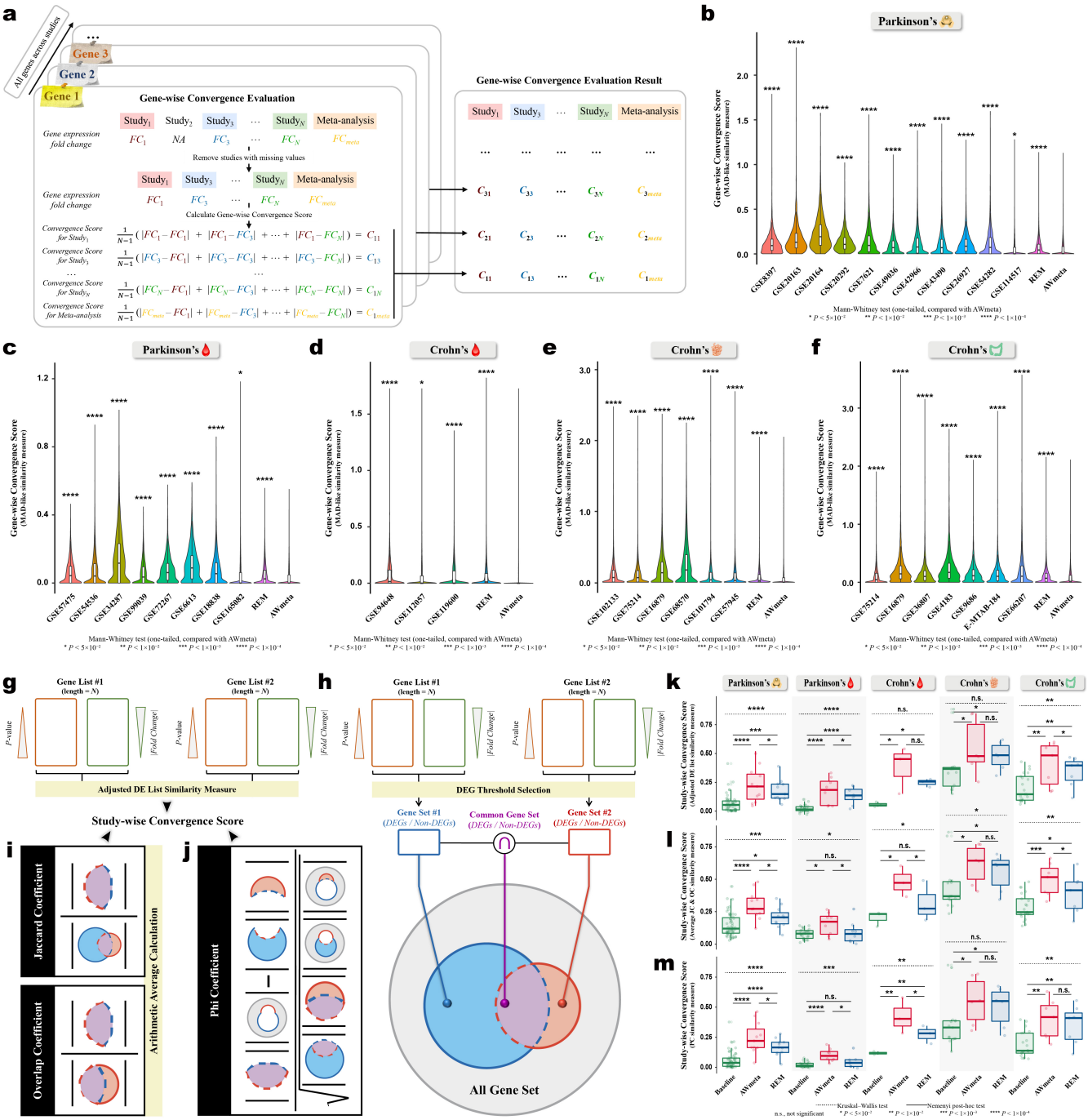


Fig. 2 | Calculation workflow and evaluation of gene-level and study-level meta-analysis convergence. **a,** Schematic illustrating the workflow for calculating gene-level convergence based on the MAD of quantitative gene differences across studies. **b-f,** Evaluation of gene-level convergence for all genes using the MAD-based metric across five disease tissue datasets: **b,** Parkinson's SN; **c,** Parkinson's blood; **d,** Crohn's peripheral blood; **e,** Crohn's IM; **f,** Crohn's CM. Statistical significance compared to AWmeta was assessed using one-tailed Mann-Whitney U tests. **g, i, j,** Principles for calculating study-level convergence metrics based on DEGs. **g,** Metric derived from the similarity of ranked DEG lists. **i,** Metric based on the arithmetic mean of Jaccard and Overlap coefficients assessing DEG set concordance. **j,** Metric using the Phi (ϕ) correlation coefficient measuring association between DEG status assignments. (**h** illustrates DEG set overlap used in set-based metrics like **i** and **j**). **k-m,** Evaluation of study-level convergence using DEGs defined by a common threshold ($|\log_2 FC| > \log_2 1.2$ and $P < 0.05$) across the five tissue datasets. Convergence assessed using: **k,** Ranked list similarity; **l,** Jaccard/Overlap mean; **m,** Phi (ϕ) coefficient. Overall differences among baseline, AWmeta, and REM were tested using the Kruskal-Wallis test, followed by Nemenyi post-hoc tests for pairwise comparisons. One-tailed Mann-Whitney U tests were used for specific comparisons against AWmeta. Icons represent tissue sources: : substantia nigra (SN); : peripheral blood (PB); : ileal mucosa (IM); : colonic mucosa (CM).

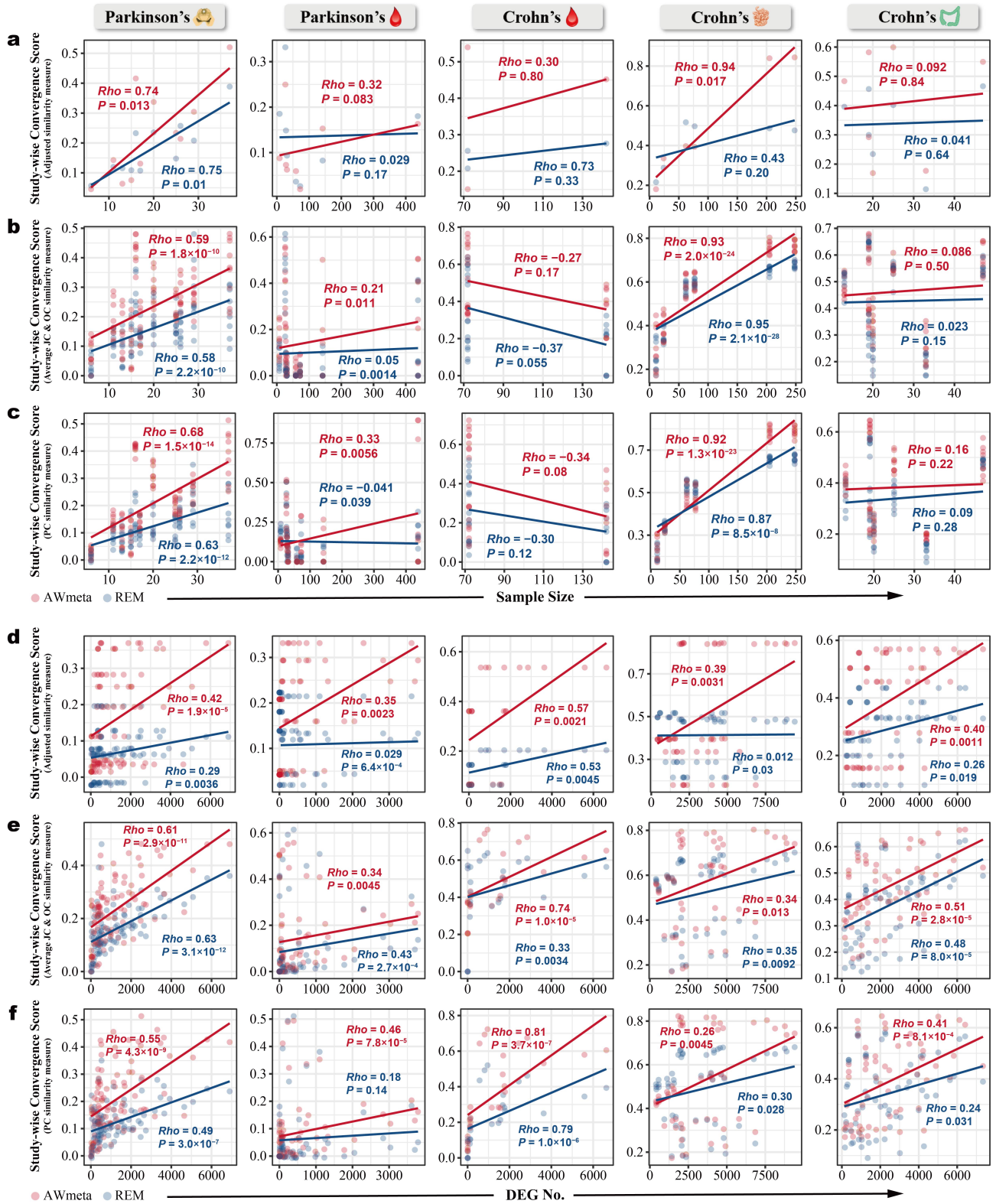


Fig. 3 | Comparison of meta-analysis convergence between AWmeta and REM across varying sample sizes and DEG numbers. a, Correlation between the similarity of ranked differential gene lists metric and sample size. **b,** Correlation between the arithmetic mean of Jaccard (JC) and overlap coefficient (OC) and sample size. **c,** Correlation between the Phi (ϕ) correlation coefficient metric and sample size. **d,** Correlation between the similarity of ranked differential gene lists metric and DEG number identified per study. **e,** Correlation between the arithmetic mean of JC and OC and DEG number identified per study. **f,** Correlation between the Phi (ϕ) correlation coefficient metric and DEG number identified per study. Icons represent tissue sources: 🧠: substantia nigra (SN); 🩸: peripheral blood; 🧫: ileal mucosa (IM); 🧬: colonic mucosa (CM).

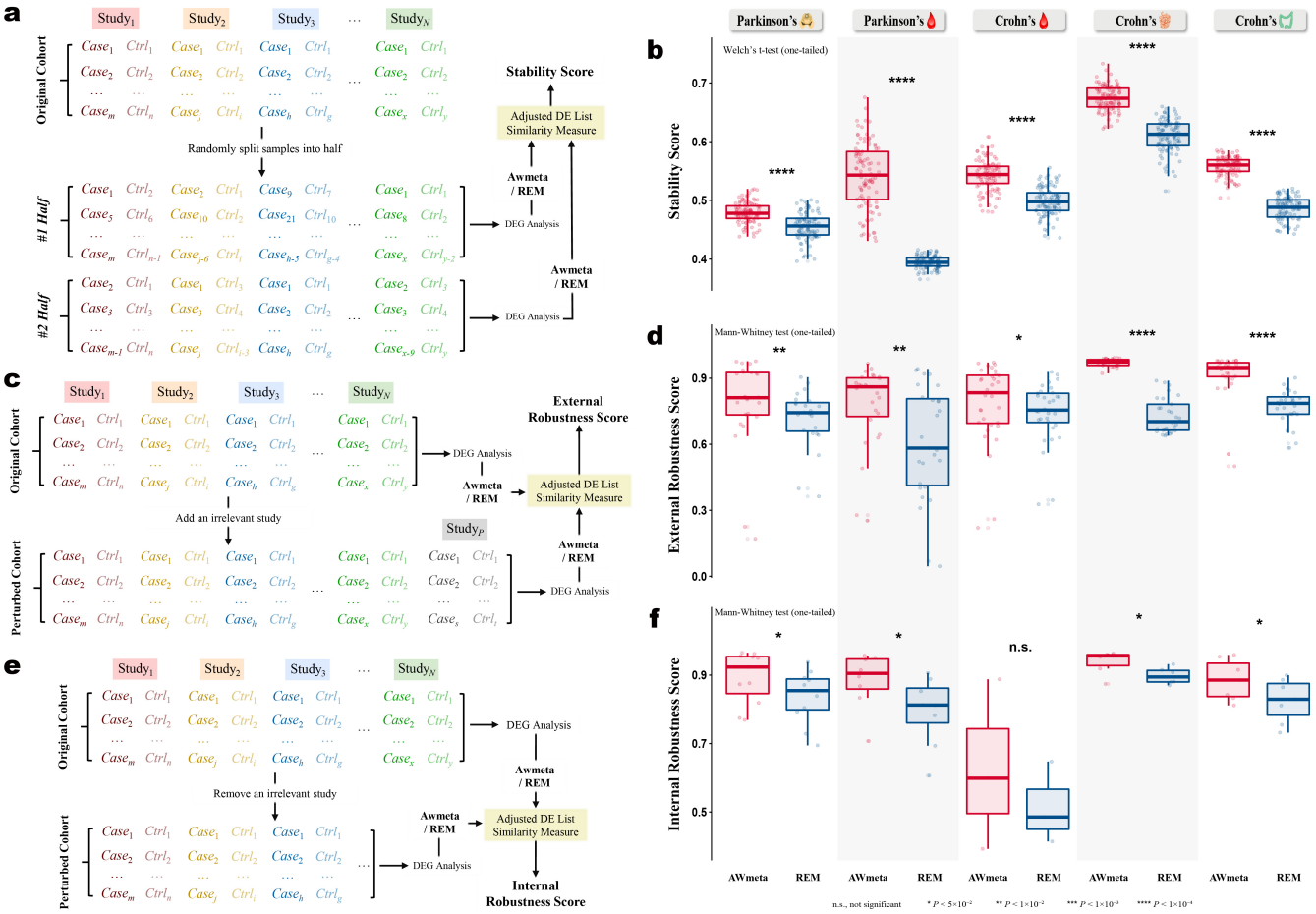


Fig. 4 | Evaluation of stability and robustness for quantitative gene difference estimates. a, Basic workflow for evaluating the stability of quantitative gene difference estimates. **b,** Evaluation of the stability of quantitative gene difference estimates. Statistical differences between AWmeta and REM were assessed using a one-tailed Welch's t-test. **c,** Workflow for calculating robustness to external outliers in quantitative gene difference estimates. **d,** Evaluation of robustness to external outliers. **e,** Workflow for calculating robustness to internal outliers in quantitative gene difference estimates. **f,** Evaluation of robustness to internal outliers. Icons represent tissue sources: : substantia nigra; : peripheral blood; : ileal mucosa; : colonic mucosa.

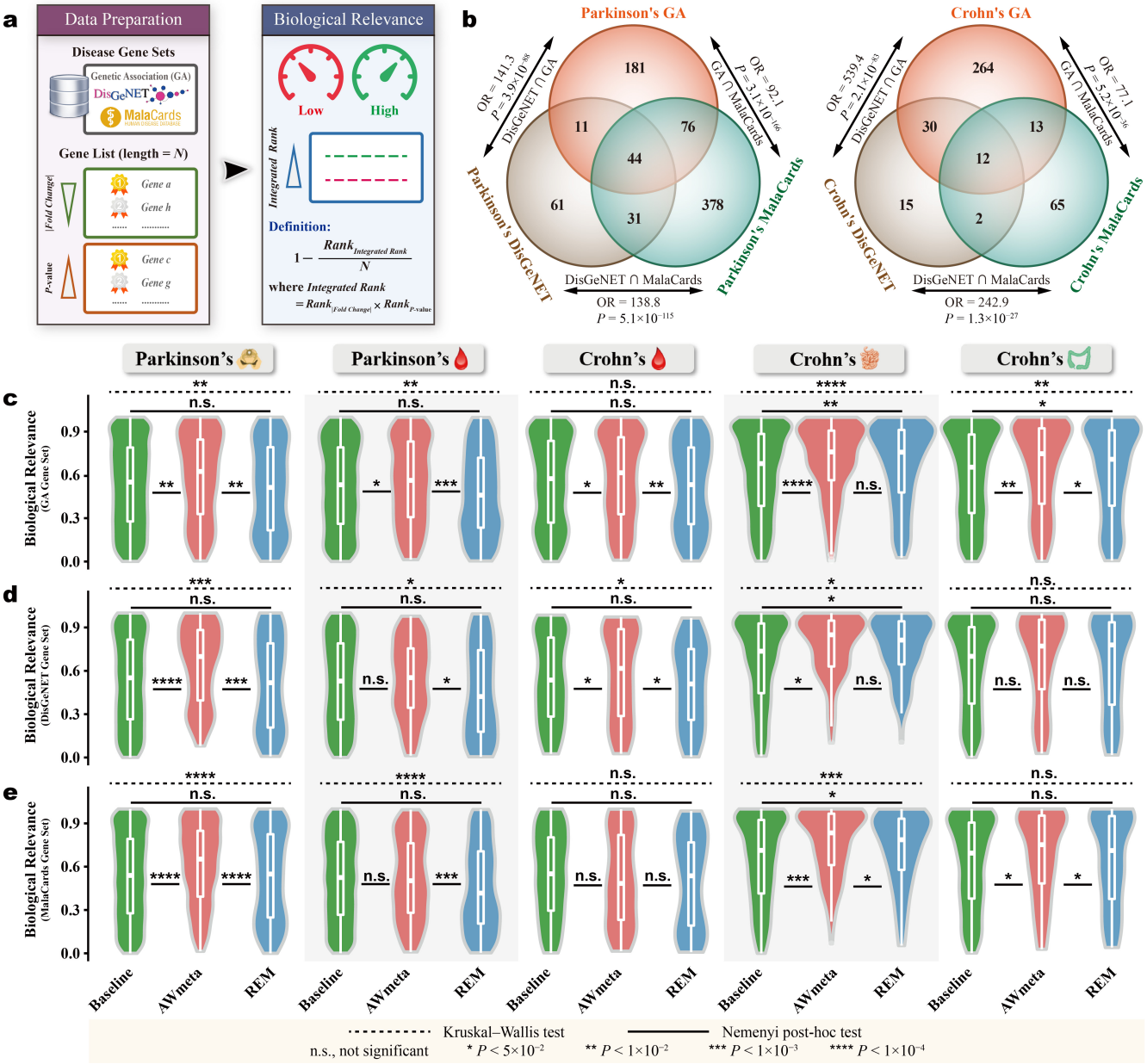


Fig. 5 | Calculation workflow and evaluation of biological relevance scores for differential gene expression. a, Workflow for calculating the quantitative biological relevance score for gene differences. **b,** Pairwise association analysis of the three benchmark gene sets for Parkinson's and Crohn's disease. The degree of overlap between gene sets was quantified using Odds Ratios (OR), and the statistical significance of this overlap was determined using Fisher's exact test. **c,** Evaluation results using genes associated with known genetic variants as the benchmark. Scores from Benchmark, AWmeta, and REM are compared. **d,** Evaluation results using DisGeNET disease-associated genes as the benchmark. Scores from Benchmark, AWmeta, and REM are compared. **e,** Evaluation results using MalaCards disease-associated genes as the benchmark. Scores from Benchmark, AWmeta, and REM are compared. Overall differences in biological relevance scores among the Benchmark, AWmeta, and REM methods were tested using the Kruskal-Wallis test (KWT). Nemenyi post hoc tests were used for pairwise comparisons between methods. Icons in the figure represent tissue sources: : substantia nigra; : peripheral blood; : ileal mucosa; : colonic mucosa.

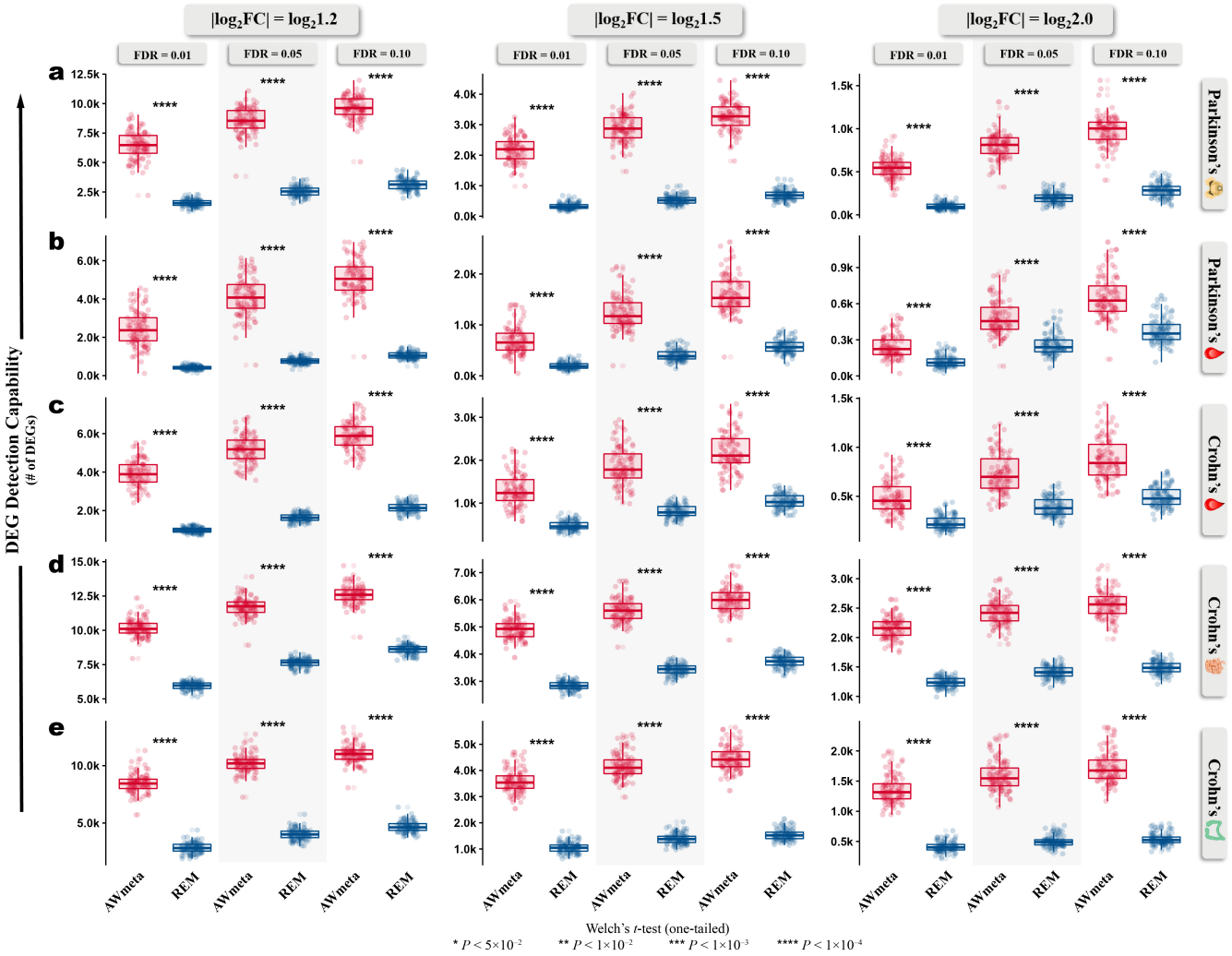
Dataset	Sequencing platform	Dataset source	Tissue source	Samples
Data IDs	Microarray or next-generation sequencing platforms	GEO, SRA and ArrayExpress	Blood, SN, IM and CM	No. of patients versus normals
PD datasets				
GSE57475	Illumina HumanHT-12 V3.0 expression beadchip	GEO	Blood	93 versus 49
GSE54536	Illumina HumanHT-12 V4.0 expression beadchip	GEO	Blood	4 versus 4
GSE34287	ExonHit Human Genome Wide SpliceArray 1.0	GEO	Blood	19 versus 12
GSE99039	Affymetrix Human Genome U133 Plus 2.0 Array	GEO	Blood	205 versus 233
GSE72267	Affymetrix Human Genome U133A 2.0 Array	GEO	Blood	40 versus 19
GSE6613	Affymetrix Human Genome U133A Array	GEO	Blood	50 versus 22
GSE18838	Affymetrix Human Exon 1.0 ST Array	GEO	Blood	17 versus 11
GSE165082	Illumina HiSeq 2000	GEO	Blood	12 versus 14
GSE114517	Illumina NextSeq 500	GEO	SN	17 versus 12
GSE8397	Affymetrix Human Genome U133A Array	GEO	SN	24 versus 13
GSE20163	Affymetrix Human Genome U133A Array	GEO	SN	8 versus 9
GSE20164	Affymetrix Human Genome U133A Array	GEO	SN	6 versus 5
GSE20292	Affymetrix Human Genome U133A Array	GEO	SN	11 versus 15
GSE7621	Affymetrix Human Genome U133 Plus 2.0 Array	GEO	SN	16 versus 9
GSE49036	Affymetrix Human Genome U133 Plus 2.0 Array	GEO	SN	8 versus 8
GSE42966	Agilent-014850 Whole Human Genome Microarray 4x44K G4112F	GEO	SN	9 versus 6
GSE43490	Agilent-014850 Whole Human Genome Microarray 4x44K G4112F	GEO	SN	8 versus 5
GSE26927	Illumina humanRef-8 v2.0 expression beadchip	GEO	SN	12 versus 8
GSE54282	Affymetrix Human Gene 1.0 ST Array	GEO	SN	3 versus 3
CD datasets				
GSE119600	Illumina HumanHT-12 V4.0 expression beadchip	GEO	Blood	95 versus 47
GSE112057	Illumina HiSeq 2000	GEO	Blood	60 versus 12
GSE94648	Affymetrix Human Genome U133 Plus 2.0 Array	GEO	Blood	50 versus 22
GSE102133	Affymetrix Human Gene 1.0 ST Array	GEO	IM	65 versus 12
GSE75214	Affymetrix Human Gene 1.0 ST Array	GEO	IM	51 versus 11
GSE16879	Affymetrix Human Genome U133 Plus 2.0 Array	GEO	IM	18 versus 6
GSE68570	Illumina HumanHT-12 V4.0 expression beadchip	GEO	IM	6 versus 5
GSE101794	Illumina HiSeq 2000	GEO	IM	198 versus 50
GSE57945	Illumina HiSeq 2000	GEO	IM	163 versus 42
GSE75214	Affymetrix Human Gene 1.0 ST Array	GEO	CM	8 versus 11
GSE16879	Affymetrix Human Genome U133 Plus 2.0 Array	GEO	CM	19 versus 6
GSE36807	Affymetrix Human Genome U133 Plus 2.0 Array	GEO	CM	13 versus 7
GSE4183	Affymetrix Human Genome U133 Plus 2.0 Array	GEO	CM	5 versus 8
GSE9686	Affymetrix Human Genome U133 Plus 2.0 Array	GEO	CM	11 versus 8
E-MTAB-184	Illumina HumanHT-12 v3.0 expression beadChip	SRA	CM	15 versus 32
GSE66207	Illumina HiSeq 2500	GEO	CM	20 versus 13

No. of Parkinson's (PD) / Crohn's disease (CD) patients: 4 8 16 32 64 128 256

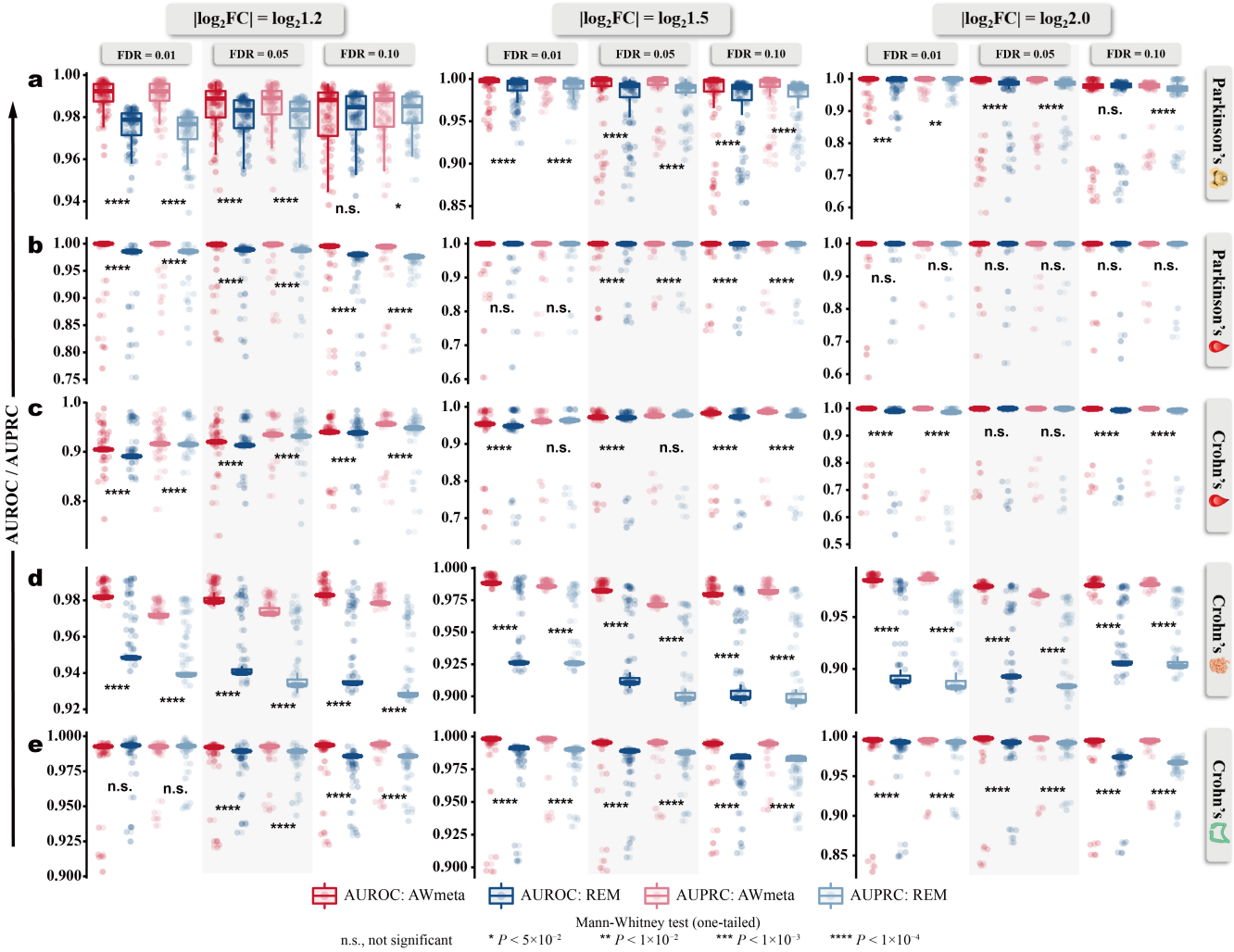
No. of normals:

GEO | SRA | ArrayExpress | Peripheral blood | Substantia nigra (SN) | Ileal mucosa (IM) | Colonic mucosa (CM) | No. of PD / CD patients | No. of normals

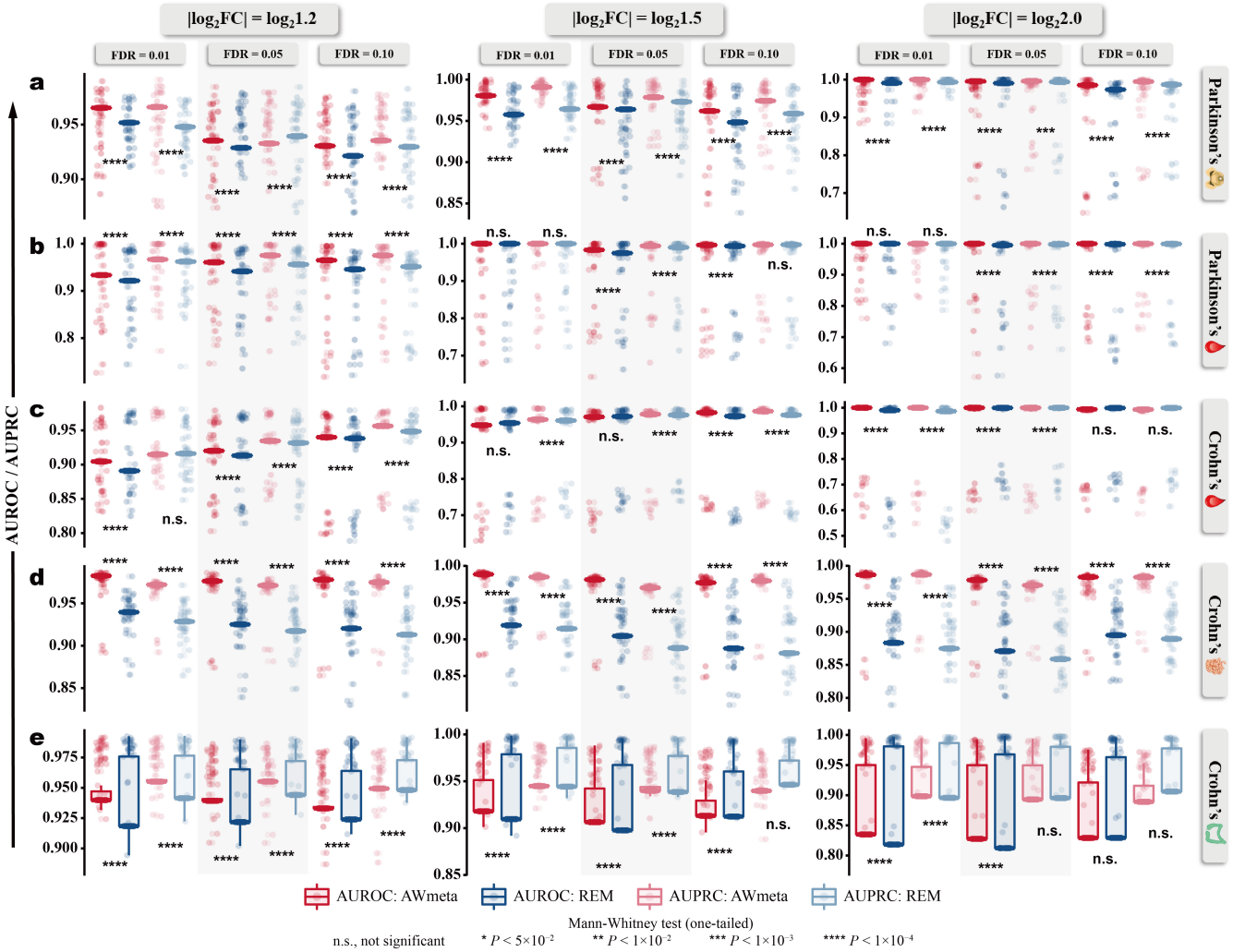
Extended Data Fig. 1 | Overview of transcriptome datasets for Parkinson's and Crohn's disease. Datasets for Parkinson's disease included substantia nigra[36–43] and peripheral blood[39,44–51]. Datasets for Crohn's disease comprised ileal mucosa[52–57], colonic mucosa[56–62], and peripheral blood[63–65]. Icons represent tissue sources: : substantia nigra (SN); : peripheral blood; : ileal mucosa (IM); : colonic mucosa (CM).



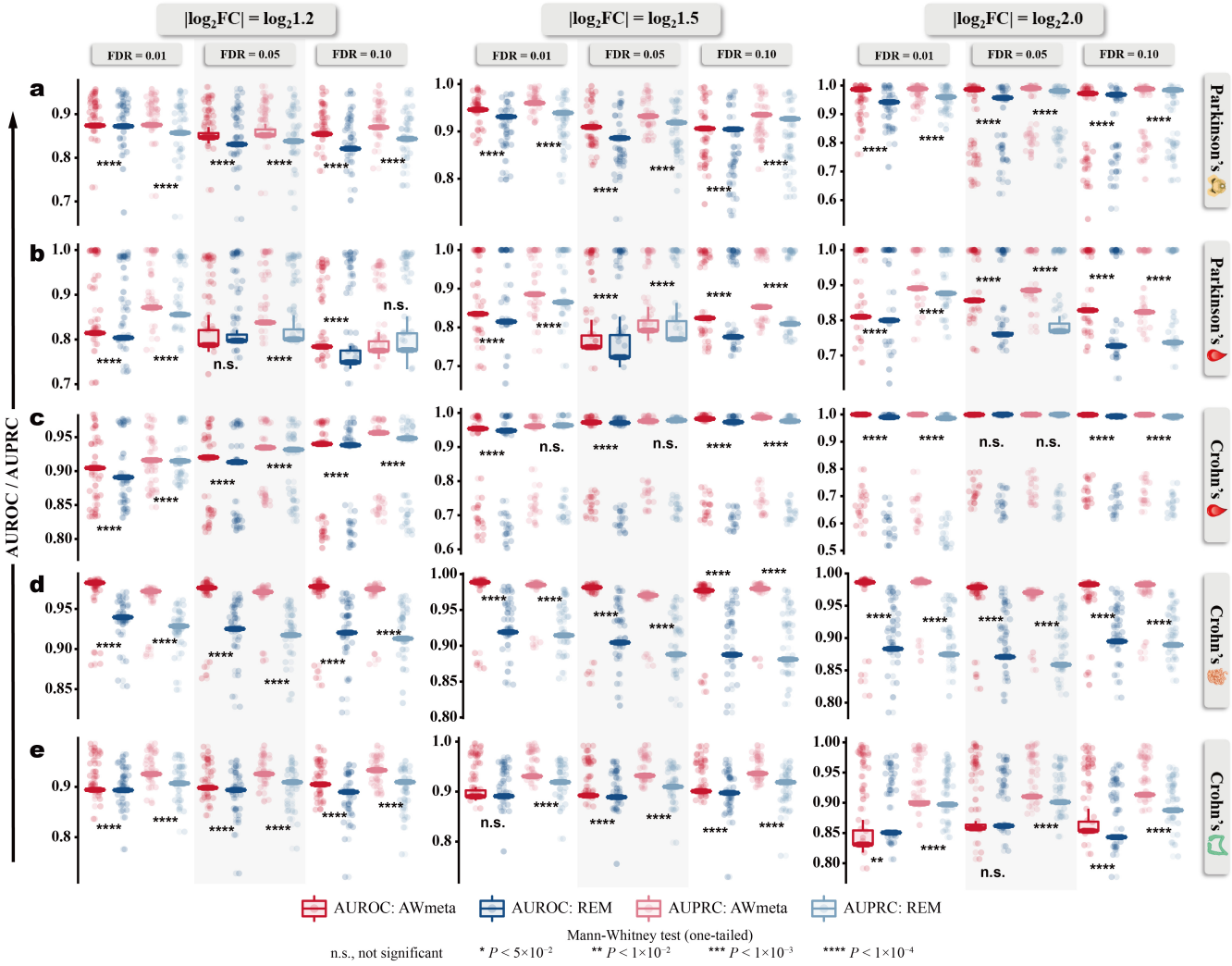
Extended Data Fig. 2 | Performance evaluation in identifying DEGs using semi-synthetic data based on the number of permuted independent studies. The ability to resolve DEGs was assessed using nine distinct thresholds, combining three statistical significance *P*-values (0.01, 0.05, 0.1) and three absolute log2 fold change cutoffs ($|\log_2 FC|$: $\log_2(1.2)$, $\log_2(1.5)$, $\log_2(2.0)$). Evaluations were performed using characteristics derived from five disease-tissue specific transcriptome datasets: **a,** Parkinson's SN, **b,** Parkinson's peripheral blood, **c,** Crohn's peripheral blood, **d,** Crohn's IM, and **e,** Crohn's CM. Icons represent tissue sources: 🧠: substantia nigra; 🩸: peripheral blood; 🧫: ileal mucosa; 🧫: colonic mucosa. Statistical significance for comparisons was determined using a one-tailed Mann-Whitney U test.



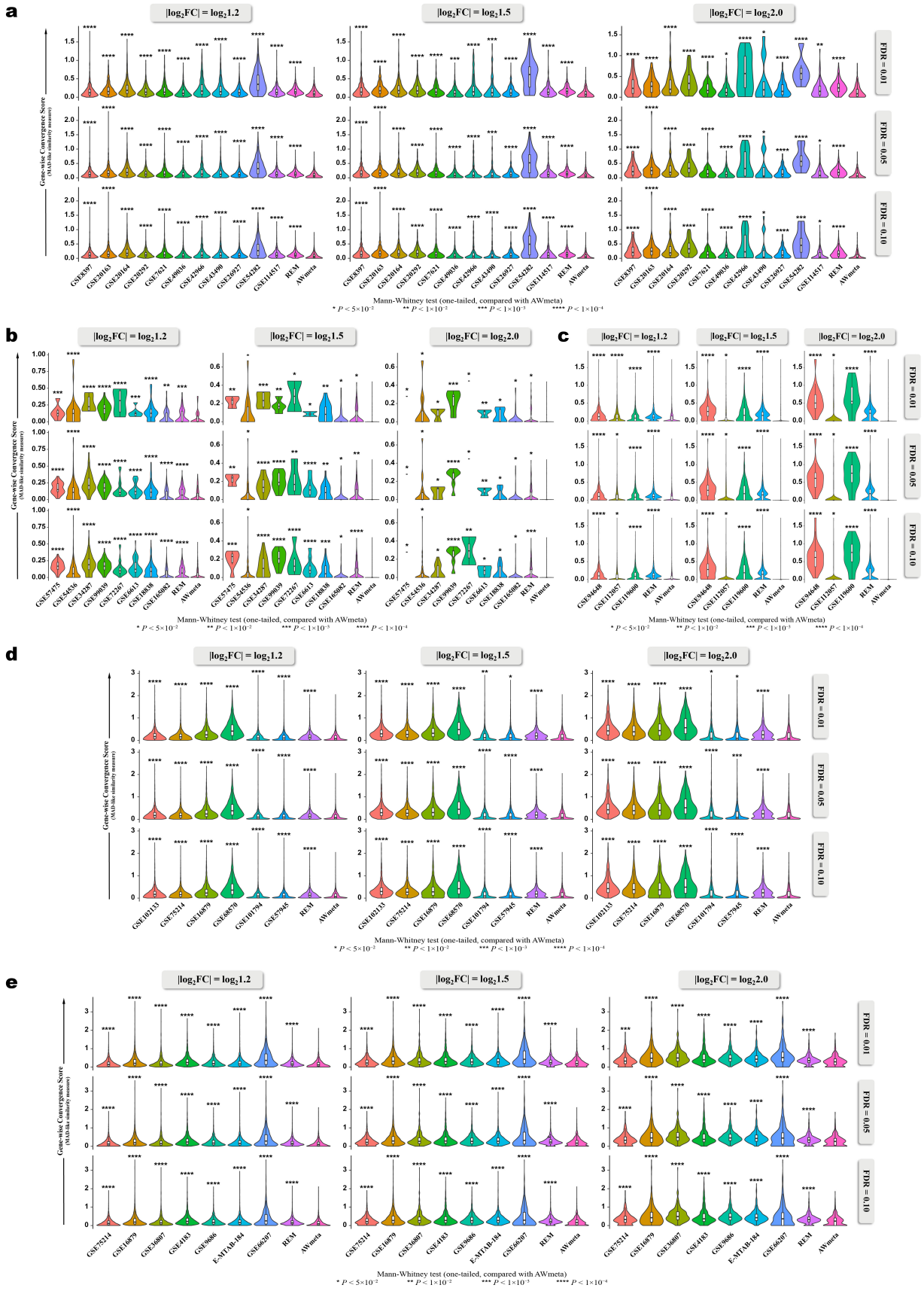
Extended Data Fig. 3 | Performance evaluation in identifying DEGs using semi-synthetic simulations based on the minimum number of included studies. The ability to resolve DEGs was assessed using nine distinct thresholds (combinations of *P*-value < 0.01, 0.05, or 0.1; and absolute $\log_2$ Fold Change $|\log_2 FC| > \log_2(1.2)$, $\log_2(1.5)$, or $\log_2(2.0)$). Evaluations were performed using characteristics derived from five disease-tissue specific transcriptome datasets: **a,** Parkinson's SN, **b,** Parkinson's peripheral blood, **c,** Crohn's PWB, **d,** Crohn's IM, and **e,** Crohn's CM. Icons represent tissue sources: 🧠: substantia nigra (SN); 🩸: peripheral blood; 🧫: ileal mucosa (IM); 🧪: colonic mucosa (CM). Statistical significance for comparisons was determined using a one-tailed Mann-Whitney U test.



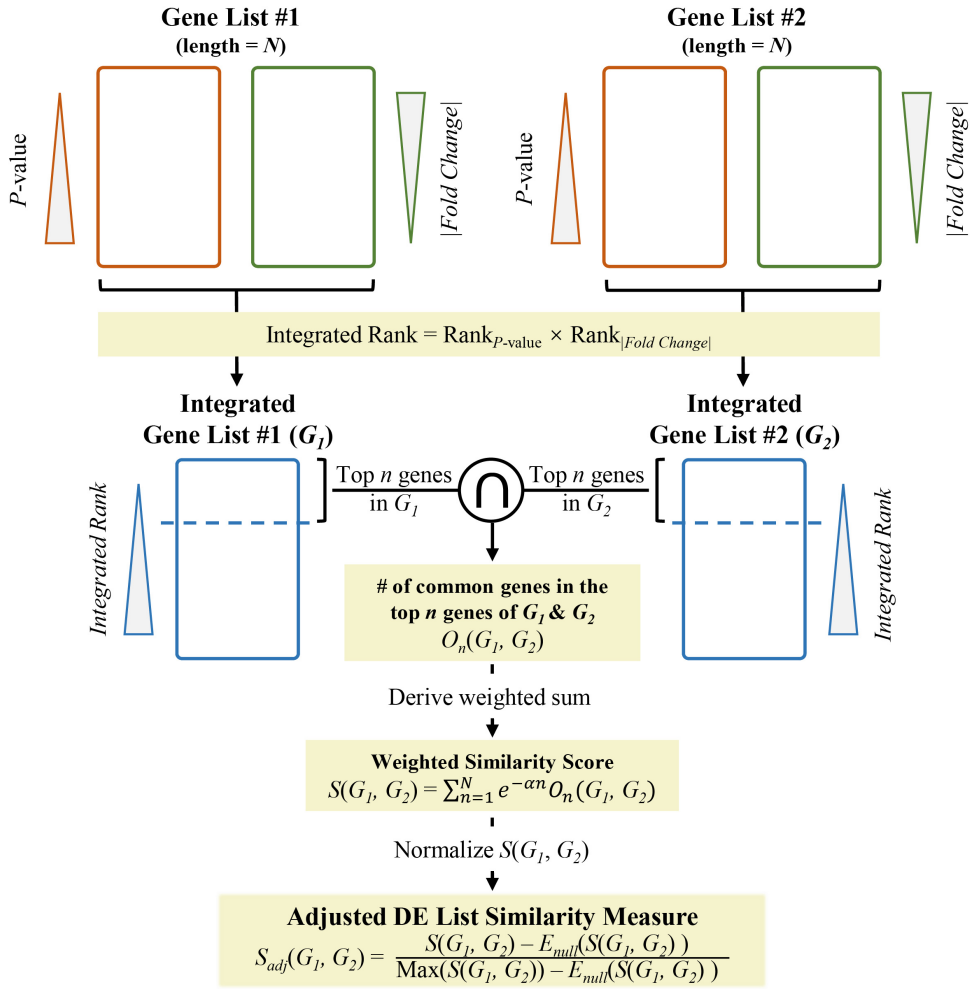
Extended Data Fig. 4 | Performance evaluation in identifying differentially expressed genes (DEGs) using semi-synthetic simulations based on the median number of included studies. The ability to resolve DEGs was assessed using nine distinct thresholds (combinations of *P*-value < 0.01, 0.05, or 0.1; and absolute $\log_2$ Fold Change $|\log_2 FC| > \log_2(1.2)$, $\log_2(1.5)$, or $\log_2(2.0)$). Evaluations were performed using characteristics derived from five disease-tissue specific transcriptome datasets: **a,** Parkinson's SN, **b,** Parkinson's peripheral blood, **c,** Crohn's peripheral blood, **d,** Crohn's IM, and **e,** Crohn's CM. Icons represent tissue sources: 🧠: substantia nigra (SN); 🩸: peripheral blood; 🧫: ileal mucosa (IM); 🧫: colonic mucosa (CM). Statistical significance for comparisons was determined using a one-tailed Mann-Whitney U test.



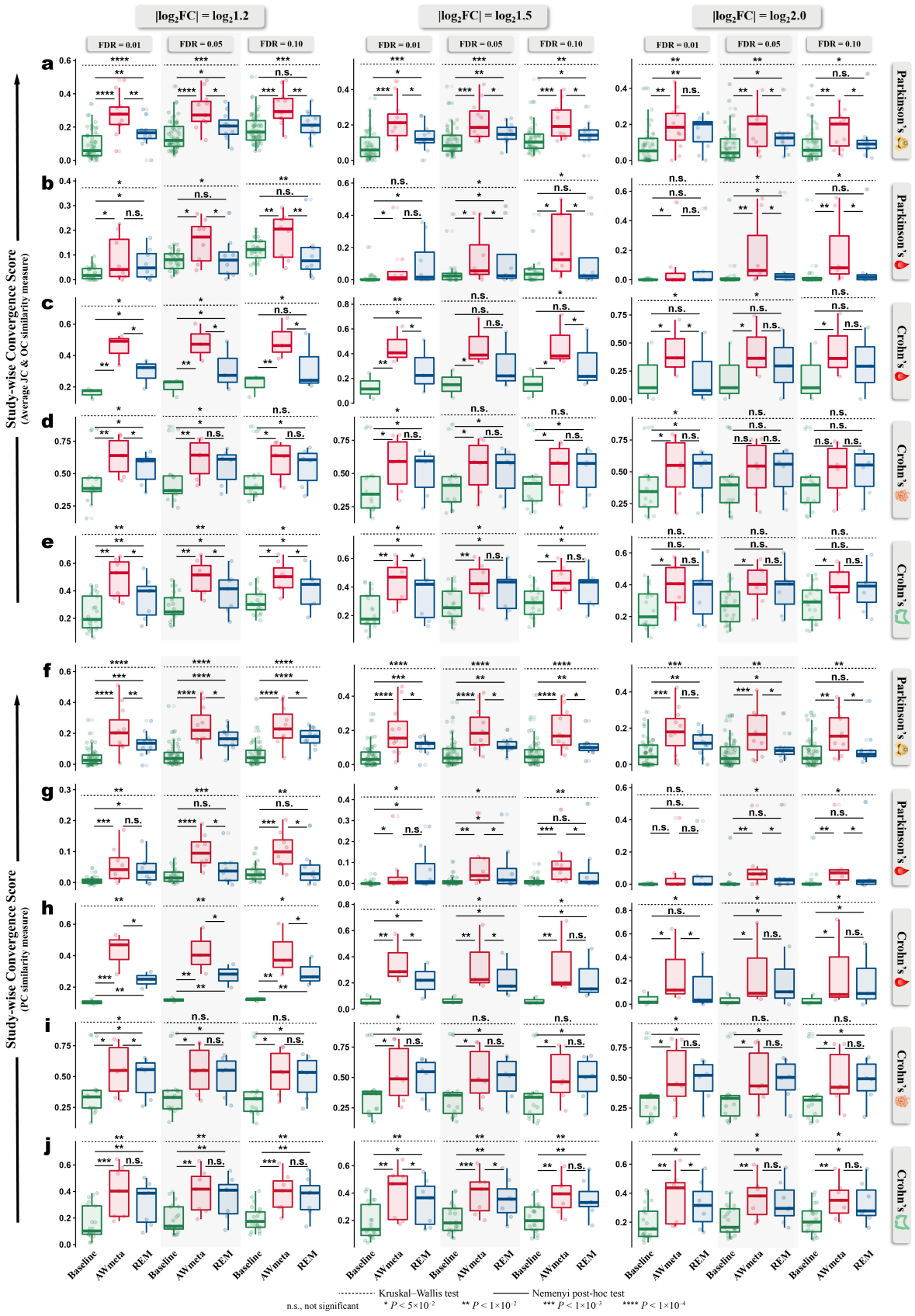
Extended Data Fig. 5 | Performance evaluation in identifying DEGs using semi-synthetic simulations based on the maximum number of included studies. The ability to resolve DEGs was assessed using nine distinct thresholds (combinations of P-value < 0.01, 0.05, or 0.1; and absolute $\log_2$ Fold Change $|\log_2 FC| > \log_2(1.2)$, $\log_2(1.5)$, or $\log_2(2.0)$). Evaluations were performed using characteristics derived from five disease-tissue specific transcriptome datasets: **a,** Parkinson's SN, **b,** Parkinson's peripheral blood, **c,** Crohn's peripheral blood, **d,** Crohn's IM, and **e,** Crohn's CM. Icons represent tissue sources: 🧠: substantia nigra (SN); 🩸: peripheral blood; 🧫: ileal mucosa (IM); 🧫: colonic mucosa (CM). Statistical significance for comparisons was determined using a one-tailed Mann-Whitney U test.


a
|log2FC| = log2 1.2
|log2FC| = log2 1.5
|log2FC| = log2 2.0
FDR = 0.01
FDR = 0.05
FDR = 0.10
Gene-wise Convergence Score
(MAD-like similarity measure)
GSE8397
GSE20163
GSE20164
GSE20292
GSE7621
GSE49036
GSE42066
GSE43490
GSE26927
GSE54282
GSE114517
REM
AWmeta
Mann-Whitney test (one-tailed, compared with AWmeta)
* P < 5×10−2 ** P < 1×10−2 *** P < 1×10−3 **** P < 1×10−4
b
GSE57475
GSE54536
GSE34287
GSE99039
GSE72267
GSE6613
GSE18838
GSE165082
c
GSE84648
GSE112057
GSE119600
d
GSE102133
GSE75214
GSE16879
GSE68570
GSE101794
GSE57945
e
GSE36807
GSE4183
GSE9686
E-MTAB-184
GSE66207

Extended Data Fig. 6 | Gene difference quantitative biocorrelation calculation process and evaluation results. **a,** Gene difference quantitative biocorrelation calculation process. **b,** Association analysis of three disease benchmark gene sets for each of Parkinson's and Crohn's disease. The degree of overlap between the two disease gene sets was calculated using Odds Ratio (OR), and the statistical significance of overlap between the two disease gene sets was calculated using Fisher's exact test. **c,** Results of quantitative biological correlation evaluation of gene differences for genetically variant genes. **d,** Results of the quantitative biological relevance evaluation of gene differences for the DisGeNET gene. **e,** Results of quantitative biological relevance evaluation of gene differences for MalaCards genes. The Kruskal-Wallis test (KWT) was used to test whether there were differences between the benchmark, AWmeta, and REM as a whole, and the Nemenyi post hoc test was used as a two-by-two statistical test between the benchmark, AWmeta, and REM. 🧠: substantia nigra; 🩸: peripheral blood; 🖐: ileal mucosa; 🧫: colonic mucosa.



Extended Data Fig. 7 | Principle of the similarity metric for ranked lists of differentially expressed genes.


|log2FC| = log21.2
|log2FC| = log21.5
|log2FC| = log22.0
FDR = 0.01
FDR = 0.05
FDR = 0.10
Study-wise Convergence Score
(Average JC & OC similarity measure)
Study-wise Convergence Score
(PC similarity measure)
Parkinson's
Crohn's
Baseline
AWmeta
REM
Kruskal–Wallis test
Nemenyi post-hoc test
n.s., not significant
* P < 5×10-2
** P < 1×10-2
*** P < 1×10-3
**** P < 1×10-4

Extended Data Fig. 8 | Gene difference quantitative biocorrelation calculation process and evaluation results. a, Gene difference quantitative biocorrelation calculation process. **b,** Association analysis of three disease benchmark gene sets for each of Parkinson's and Crohn's disease. The degree of overlap between the two disease gene sets was calculated using Odds Ratio (OR), and the statistical significance of overlap between the two disease gene sets was calculated using Fisher's exact test. **c,** Results of quantitative biological correlation evaluation of gene differences for genetically variant genes. **d,** Results of the quantitative biological relevance evaluation of gene differences for the DisGeNET gene. **e,** Results of quantitative biological relevance evaluation of gene differences for MalaCards genes. The Kruskal-Wallis test (KWT) was used to test whether there were differences between the benchmark, AWmeta, and REM as a whole, and the Nemenyi post hoc test was used as a two-by-two statistical test between the benchmark, AWmeta, and REM. 🧠: substantia nigra; 🩸: Peripheral blood; 🤚: ileal mucosa; 🧫: colonic mucosa.