nature communications

Article

https://doi.org/10.1038/s41467-025-60154-0

Worm Perturb-Seq: massively parallel whole-animal RNAi and RNA-seq

Received: 14 May 2024

Accepted: 15 May 2025

Published online: 23 May 2025

Check for updates

Hefei Zhang[1,8], Xuhang Li[1,8], Dongyuan Song[2,6], Onur Yukselen[3], Shivani Nanda[1,7], Alper Kucukural[3,4], Jingyi Jessica Li[2,5], Manuel Garber[4] & Albertha J. M. Walhout[1]

Transcriptomes provide highly informative molecular phenotypes that, combined with gene perturbation, can connect genotype to phenotype. An ultimate goal is to perturb every gene and measure transcriptome changes, however, this is challenging, especially in whole animals. Here, we present 'Worm Perturb-Seq (WPS)', a method that provides high-resolution RNA-sequencing profiles for hundreds of replicate perturbations at a time in living animals. WPS introduces multiple experimental advances combining strengths of Caenorhabditis elegans genetics and multiplexed RNA-sequencing with a novel analytical framework, EmpirDE. EmpirDE leverages the unique power of large transcriptomic datasets and improves statistical rigor by using gene-specific empirical null distributions to identify DEGs. We apply WPS to 103 nuclear hormone receptors (NHRs) and find a striking 'pairwise modularity' in which pairs of NHRs regulate shared target genes. We envision the advances of WPS to be useful not only for C. elegans, but broadly for other models, including human cells.

Since the dawn of functional genomics, the transcriptome has proven to be one of the most powerful molecular phenotypes to connect genotype to phenotype[1–3]. While early work in yeast provided insights into the transcriptional responses to gene deletions[4,5], similar large-scale and systematic studies in multicellular organisms have been lacking. Moreover, statistical analyses of large-scale, high-throughput genomics data suffer from technical biases and high false discovery rates (FDRs)[6], e.g., many false positives in the identification of differentially expressed genes (DEGs). More recently, a method commonly referred to as Perturb-seq has been developed that uses pooled CRISPR-based gene perturbation screens with single-cell RNA-seq. This method has proven powerful in cell-based functional screens to annotate gene function, identify genetic interactions, and to infer disease-related pathways[7–13]. Empowered by single-cell RNA-seq and pooled screening, this type of approach provides unparalleled multiplexity, enabling genome-wide perturbation and sequencing in a single or just a few experiments. However, this unique advantage also comes with trade-offs, including low sensitivity for gene detection, lack of biological replicate experiments (due to high costs), and challenges to perturb many genes in vivo[3,14–16].

Here, we present 'Worm Perturb-Seq' (WPS) in which individual genes are knocked down in the nematode *C. elegans* by feeding bacteria expressing double-stranded RNA, followed by RNA-seq using a strategy that adopts the high multiplexity of single-cell sequencing but uses bulk samples to produce high-resolution RNA-seq profiles. WPS is labor- and cost-efficient and enables replicate experiments. Using

[1]Department of Systems Biology, University of Massachusetts Chan Medical School, Worcester, MA, USA. [2]Bioinformatics Interdepartmental Ph.D. Program, University of California, Los Angeles, CA, USA. [3]Via Scientific Inc., Cambridge, MA, USA. [4]Department of Genomics and Computational Biology, University of Massachusetts Chan Medical School, Worcester, MA, USA. [5]Department of Statistics and Data Science, Department of Biostatistics, Department of Computational Medicine, and Department of Human Genetics, University of California, Los Angeles, CA, USA. [6]Present address: Department of Genetics and Genome Sciences, University of Connecticut Health Center, Farmington, CT, USA. [7]Present address: Pathology, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA. [8]These authors contributed equally: Hefei Zhang, Xuhang Li. ✉e-mail: manuel.garber@umassmed.edu; marian.walhout@umassmed.edu

large-scale WPS data, we find subtle yet systematic fluctuations in gene expression caused pervasive false positive DEGs when analyzed by standard differential expression (DE) analysis methods that compare experiment to control conditions. To circumvent this issue, we develop a two-pronged data analysis framework, EmpirDE ('Empirical Differential Expression') that leverages the large WPS dataset to identify DEGs. EmpirDE systematically mitigates technical confounders by using gene-specific models with empirical null distributions to correct for anti-conservative *P* values (i.e., where the significance is overestimated) obtained by standard DE analysis. We demonstrate the rigorous control of FDR by EmpirDE using both simulations and experimental benchmarking. We apply WPS to the knockdown of 103 nuclear hormone receptors (NHRs) and discover that NHR pairs frequently share overlapping target genes, which cannot be explained by protein similarity, but is more related to NHR coexpression. WPS will enable examining different perturbations in addition to RNAi, including mutants, bacterial diets, and exposure to drugs or toxins. Importantly, EmpirDE will also broadly enable statistically rigorous analyses of large-scale transcriptomic data (i.e., >100 conditions) in other systems.

Results

In *C. elegans*, gene expression can be knocked down by feeding the animals bacteria expressing double-stranded RNA for a target gene of interest[17,18]. This whole-animal RNA interference (RNAi) is easy to perform for large sets of genes in parallel, and multiple RNAi libraries have been developed[19–23]. We developed WPS, which is composed of two major components: (1) an experimental approach to perform high-throughput whole-animal RNAi experiments and generate multiplexed RNA-seq libraries, and (2) a computational pipeline for quality control and rigorous statistical analysis of DEGs (Fig. 1a).

An overview of Worm Perturb-Seq (WPS)

WPS consists of several steps, many of which were optimized to enable high-throughput, cost-effective experiments (Fig. 1b). Briefly, RNAi is started with animals at the first larval stage (L1) and, when grown to the desired stage, animals are harvested, and total RNA is extracted in a 96-well extraction plate. This streamlined workflow allows efficient triplicate experiments for hundreds of knockdowns (Fig. 1a). Multiplex RNA-seq libraries are constructed using an early barcoding step during reverse transcription in the 96-well plates, with each barcode linked to a single perturbation, followed by pooling of ~50 samples and sequencing library construction. After sequencing, several quality control steps are performed (see below) and DEGs are identified with EmpirDE, which uses gene-specific models with empirical null distributions.

An experimental WPS platform

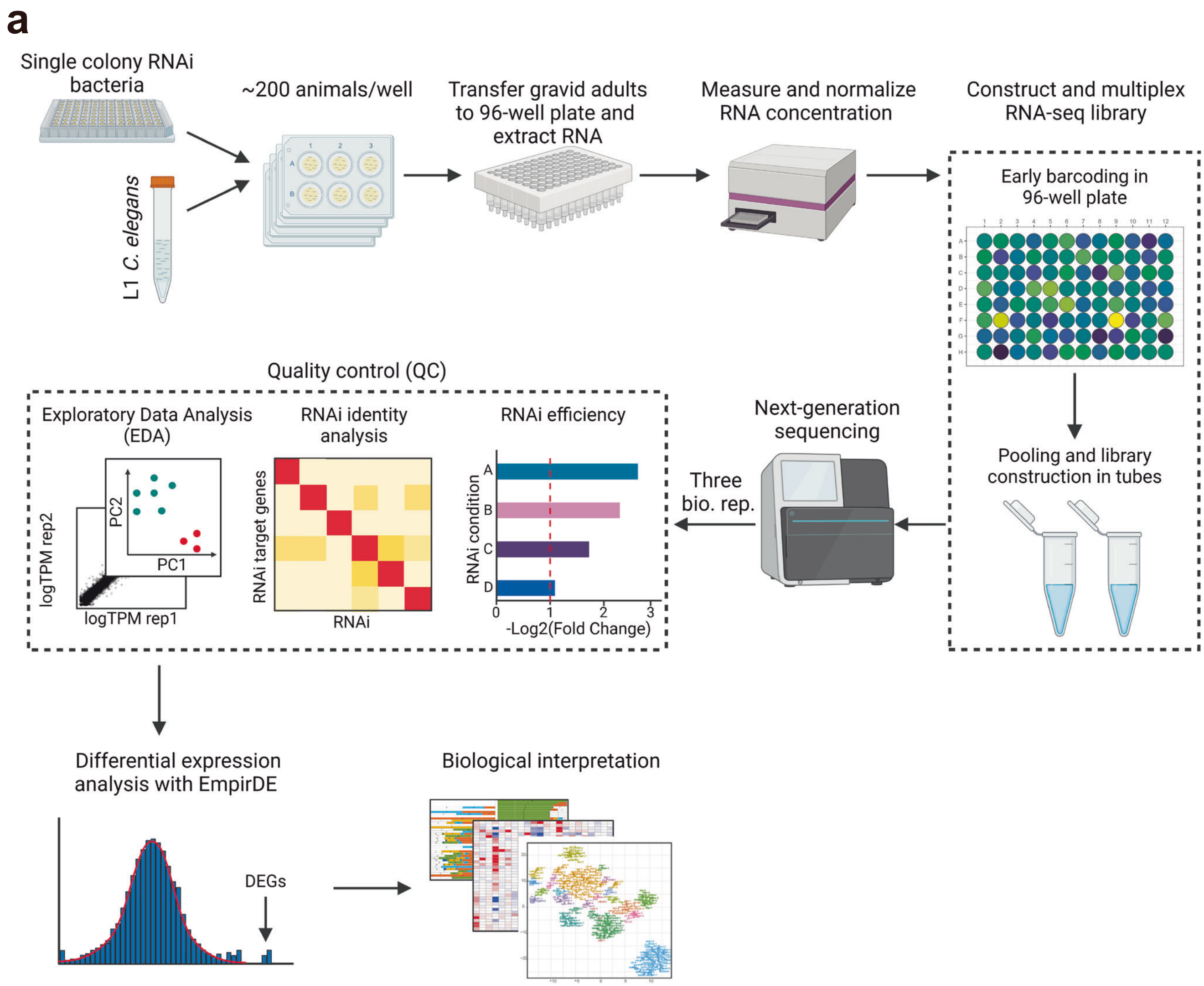
Several experimental steps of WPS were developed and optimized, including growing animals, harvesting RNA, and generating, pooling, and sequencing of multiplexed libraries (Fig. 1b, Supplementary Protocols, Supplementary Data 1 and Supplementary Note 1). Notably, WPS introduces a high-throughput worm lysis method for RNA extraction in 96-well plates, which does not lyse eggs, making it suitable to use WPS for gravid adults (Supplementary Fig. 1a). We optimized the 3' end barcoding method CEL-Seq2[24], which was originally developed for single-cell RNA-seq[25], for multiplexing bulk RNA-seq libraries, with significant reduction of costly reagents (Fig. 1b and Supplementary Protocol).

The transcriptome of *C. elegans* changes greatly over the course of its lifetime; there are oscillatory expression profiles during development, and gene expression continues to change as the animals reproduce and age[26–28]. Therefore, if a knockdown has even a small effect on development, it can result in many DEGs that are secondary to the effect of the knockdown on development, rather than in response to the perturbed gene. We therefore opted to use a period in the animal's lifetime in which the transcriptome does not change to minimize developmental effects of knockdowns. Animals develop from L1 to gravid adults in ~58 h at 20 °C and we found that the gravid adult transcriptome was most stable between 60 and 68 h post-L1-plating (Fig. 2a, Supplementary Fig. 1b). Therefore, we used a time of ~63 h post-plating, which is in between the first egg laid at 58 h and the first egg hatched at 68 h, allowing enough time for sample collection and processing, and providing a buffer for perturbations that elicit a mild developmental day (Fig. 2a).

In other systems, it has been shown that most genes can be quantified with a relatively shallow read depth[29,30]. We performed down-sampling analysis of a dataset with sequencing depth ranging from 39.5 to 53.9 million reads in three biological replicates (Supplementary Fig. 1c and Supplementary Data 2). We used an average of 6 million reads per sample, with which 90% of genes with >4 transcripts per million (TPM) and 80% of genes with >2 TPM could be quantified (Fig. 2b, Supplementary Fig. 1d). In addition, we compared the gene detection sensitivity of this WPS setup with the conventional approach and found that the differences are negligible for significantly expressed genes (e.g., TPM > ~0.5) (Supplementary Fig. 1e). For library multiplexing, we pooled ~54 samples, which included 16 perturbations, each containing three biological replicates, together with six negative controls (empty vector RNAi) into one sequencing library. This design was intended to minimize batch effects by having all replicates of controls and perturbations in the same library. We first established a proof of concept by targeting 103 NHRs as discussed below and then extended this to the knockdowns ~900 metabolic genes in the metabolic network of *C. elegans*[31,32]. Here, we combined these two WPS datasets (> 4000 profiles collected in 80 libraries) for benchmarking analysis.

WPS quality control

Analysis of large-scale and high-throughput functional genomics experiments can be complicated by batch effects and low-quality samples[6,33,34]. We followed standard practices to ensure the quality of individual samples[35,36], and to identify and remove outlier replicates (Fig. 2c, Supplementary Fig. 1f, Supplementary Methods). We next developed two RNAi quality control (QC) analyses to verify the gene that was knocked down. First, the reduction of targeted gene expression can be directly read out. For instance, for 85% of genes that are expressed at a high level (TPM ≥ 30), we found a > 2-fold reduction in their mRNA levels when knocked down (Supplementary Fig. 1g). Second, due to abundant reverse strand reads that map to the gene body of knocked down genes, the identity of the perturbed gene can be directly identified from the WPS data (Fig. 2d, red reads). These reads are likely derived from unspecific reverse transcription of anti-sense RNA generated during the RNA interference process[37]. This latter RNAi identity verification is particularly useful for genes that are expressed at low levels and was able to verify the identity of almost all NHR RNAi clones that were also confirmed by Sanger sequencing (Fig. 2e). We next performed WPS using RNAi clones that were not confirmed a priori[31] and found three incorrect clones (Fig. 2f, g, indicated in red). Importantly, we could identify the actual target by mapping anti-sense sequences to the *C. elegans* genome (Fig. 2f, non-diagonal signals for the red RNAi conditions). Two clones that returned a hit in the search were subsequently confirmed by Sanger sequencing (Supplementary Fig. 1h). The other incorrect clone had a partial insert that did not target a transcribed gene and was considered a non-targeting perturbation (NTP). In the metabolic-gene WPS screen[31] we found that ~13% of the samples had a wrong RNAi identity, including 67 NTPs (Supplementary Data 3), showing the necessity of RNAi QC in large-scale screens. Taken together, WPS data can be directly used to validate that the gene that has been knocked down and clones that


a
Single colony RNAi bacteria
L1 C. elegans
~200 animals/well
Transfer gravid adults to 96-well plate and extract RNA
Measure and normalize RNA concentration
Construct and multiplex RNA-seq library
Early barcoding in 96-well plate
Pooling and library construction in tubes
Next-generation sequencing
Three bio. rep.
Quality control (QC)
Exploratory Data Analysis (EDA)
logTPM rep2
logTPM rep1
PC2
PC1
RNAi identity analysis
RNAi target genes
RNAi
RNAi efficiency
RNAi condition
A
B
C
D
0
1
2
3
-Log2(Fold Change)
Differential expression analysis with EmpirDE
DEGs
Biological interpretation

b

Optimization step	Highlights		
RNA extraction	**Scalable:** lyse animals by chemical method in 96-well plate	**Specific:** only break down animal bodies and not their eggs	**Robust:** identified a transcriptional steady state, 60-68 hours post L1, which allows for robust large-scale experimentation
RNA-seq library constructions	**Improved multiplexing:** pool individual samples after first-strand cDNA synthesis to improve robustness and cost efficiency	**Standardization:** optimize and standardize a protocol for bulk RNA-seq library construction based on CEL-Seq2 method	**Optimal read depth:** ~6M on average for optimal cost-benefit tradeoff
Quality control	**RNAi quality assurance:** identify reads from the anti-sense RNA generated during RNA interference to ensure the identity of RNAi target and quantify the efficiency of knockdown by the decrease of target gene expression		
Differential expression analysis	**Power:** three biological replicates for >95% RNAi conditions	**Rigorous false discovery control:** a two-pronged DE analysis framework, EmpirDE, that uses gene-specific models with empirical null distributions to rigorously identify DEGs at desired false discovery rate	

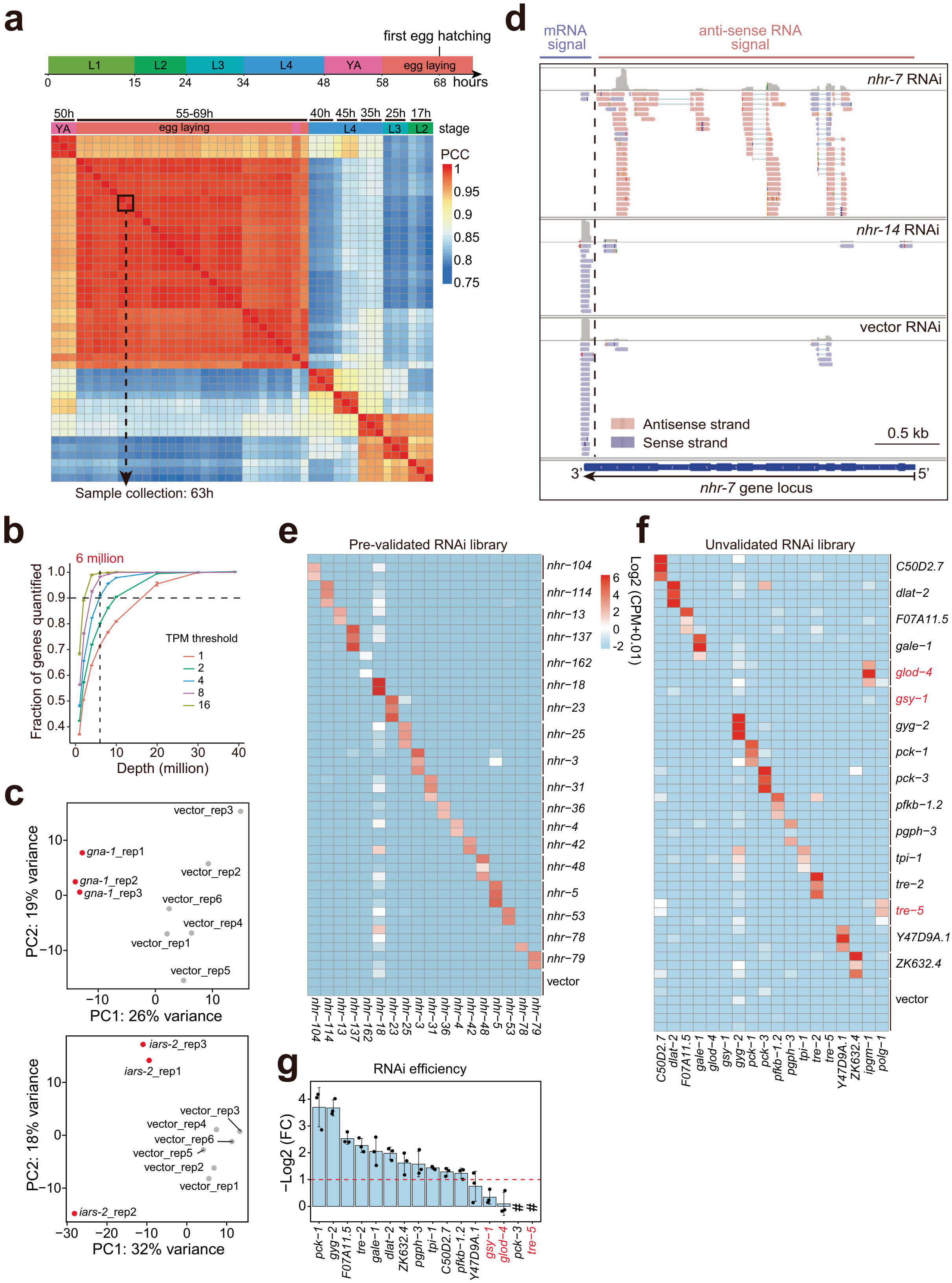
Fig. 1 | Worm Perturb-Seq (WPS) overview. a Overview of the WPS pipeline. This figure was created in BioRender. lee, y. (2025) https://BioRender.com/u29b568. **b** WPS optimization highlights.

are incorrect can simply be removed from the dataset or analyzed with the corrected target information.

Standard DE analysis results in high false discoveries

The 67 NTPs from the metabolic-gene WPS study[31] should have zero DEGs and can therefore be used to evaluate the actual FDR in WPS. We initially conducted DE analysis with DESeq2[38], by comparing each RNAi perturbation to vector controls from the same sequencing library. Surprisingly, this approach resulted in dozens of DEGs in both NTP and four randomly spike-in vector control conditions (Fig. 3a, $P_{adj} < 0.01$, fold change (FC) > 2, collectively referred to as NTPs hereafter). This suggests a high level of false discoveries despite stringent filtering by estimated FDR and FC thresholds.

False positive DEG calls are common in RNA-seq studies[39–42] and are potentially more profound when combined with large-scale screens because of systematic variations[6,40]. By comparing mRNA levels among vector control, NTP, and RNAi samples within the same sequencing library and across dozens of libraries, we discovered two



confounding issues. The first issue involved sequencing libraries in which a gene consistently behaved differently in the vector control samples compared to the RNAi samples in the same sequencing library (Fig. 3b, 'control-outlier gene', Supplementary Fig. 2a). For instance, *swt-3* expression was lower in all RNAi samples when compared to the vector control in the same batch (Fig. 3b), resulting in *swt-3* being identified as a DEG in all these conditions, including in the NTP. This problem is more likely an effect associated with confounded controls in large-scale experiments, especially the array-based screens like in WPS. The second issue involved genes with highly variable mRNA levels across the WPS dataset (Fig. 3c, see Supplementary Fig. 2b for the entire dataset, 'noisy genes') that were frequently called as DEGs in

Fig. 2 | Development of WPS and data quality control. **a** Comparison of the *C. elegans* transcriptome across developmental stages. The Pearson correlation coefficient (PCC) was calculated by the WPS profiles of animals fed vector control bacteria and collected at different time points post L1. **b** Subsampling analysis of WPS profiles, combining data from 62 to 65 h for each replicate shown in Fig. 2a. The plot shows the fraction of genes quantified versus sequencing depth. Genes whose expression levels in subsampling fall within roughly ±30% interval of the reference value were considered as quantified (for detailed definition, see Supplementary Methods). Error bar shows the mean values (± s.d.) from three replicates of the subsampling profile. **c** Representative Principal Component Analysis (PCA) results for gene expression profiles of perturbations without (*gna-1* on top panel) and with (*iars-2* in bottom panel) a low-quality outlier replicate. Red and gray dots indicate RNAi and control samples, respectively. **d** An example showing reads mapped to the reverse strand of the RNAi target gene (*nhr-7*) and a decrease in mRNA reads at the 3′ end. The reads mapped to *nhr-7* gene locus were visualized by Integrative Genomics Viewer (IGV[89]). *nhr-7* RNAi was compared to vector control RNAi and another RNAi condition (*nhr-14*). Quantification of anti-sense RNA reads in a Sanger-sequenced (**e**) and an unvalidated (**f**) WPS sequencing library. Row names represent the intended RNAi gene (three replicates each) and column names represent the actual knocked down genes. Row names in red indicate wrong RNAi clones. Values are the log2(Count-Per-Million (CPM) + 0.01) of the reads mapped to the reverse strand of each gene in the columns. **g** Log2(Fold Change (FC)) of the RNAi targeted gene expression for the WPS sequencing library shown in (**f**). Pound key (#) indicates not detected. Each bar represents the mean (±s.d.) and each dot represents one biological replicate ($n = 3$). Source data are provided as a Source Data file.

both RNAi perturbations and NTPs by DESeq2. We next sought to systematically address these and other possible effects that resulted in the seemly systematic, anti-conservative *P* values and unexpectedly high false discoveries.

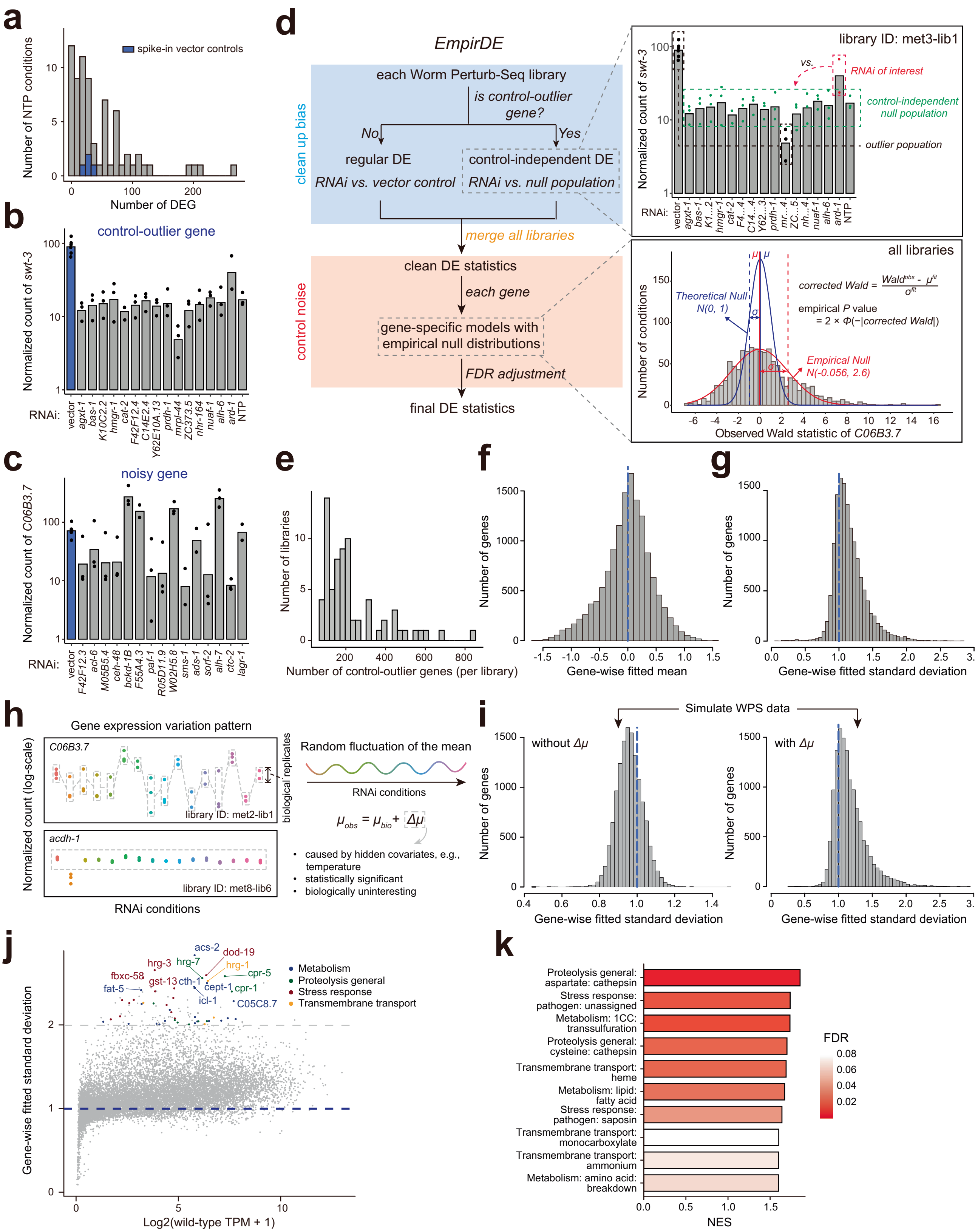
EmpirDE: an empirical null-based, gene-centered method to rigorously analyze differential expression

A major challenge in transcriptomics is that successful DE analysis hinges on reasonable model assumptions and accurate parameter estimation. However, it is impossible to accurately estimate parameters, such as gene-specific variance or to identify model misspecification in small-sample-size experiments[43–46]. To systematically identify and correct for false positive DEGs, we developed EmpirDE, which leverages the power of having hundreds of conditions assessed in a uniform experimental setup, enabling the rigorous identification of true DEGs that are elicited by a specific knockdown. EmpirDE uses a two-pronged approach, first at the level of individual sequencing libraries and second at the level of an entire dataset (Fig. 3d). The first step performs DE analysis within each sequencing library (~16 conditions in our experiments) using DESeq2. However, instead of simply comparing an RNAi condition to the control, this step identifies control-outlier genes and treats these differently using a control-independent DE analysis procedure (Supplementary Note 2). Briefly, for each control-outlier gene, this step empirically identifies a control-independent null population based on the distribution of the gene's expression levels in the sequencing library and compares the level for the gene in each RNAi condition to the newly defined null population. The second step of EmpirDE combines the DE results from all sequencing libraries (here ~1000 triplicate conditions) to correct anti-conservative *P* values based on gene-specific empirical null distributions of the DE test statistic (i.e., Wald statistic). Dozens to hundreds of conditions uniformly collected in WPS provided a unique opportunity to directly estimate the empirical null from the data in a gene-centered manner[47]. Assuming real effects are rare in large-scale experiments, the gene expression in most conditions can be viewed as 'unchanged', thus defining an empirical null population. By fitting the central peak of the distribution of the test statistic (i.e., the distribution of conditions that did not have a significant effect)[47], this step estimates the empirical n`ull and rescales the original test statistic (i.e., Wald statistic) accordingly to obtain a corrected Wald statistic on a gene-by-gene basis. This corrected statistic should follow a standard normal distribution and can be converted to a *P* value (referred to as empirical *P* value) to identify perturbations where differences in expression levels are statistically significant from the empirical null distribution (Fig. 3d).

EmpirDE identified fewer than 200 control-outlier genes in most sequencing libraries (Fig. 3e), indicating a relatively low but significant number of confounded genes (1.4% of 14,000 detected genes). In the scenario of a well-fitted DE model, the empirical null distribution of the Wald test statistic in DESeq2 analysis should adhere to its theoretical null, a standard normal distribution[38]. Surprisingly, we observed a systematic difference between the theoretical and empirical null for the ~14,000 detected genes in this dataset (Fig. 3f, g, Fig. 3d shows an example gene *C06B3.7*). While the mean of the empirical null distribution was symmetrically aligned around the mean of the theoretical Wald distribution (0) (Fig. 3f), the empirical null had systematically larger standard deviations than the theoretical expectation (1) (Fig. 3g). As a result, *P* values computed using the theoretical Wald distribution were anti-conservative, but this could be corrected by EmpirDE (Fig. 3d, empirical *P* value).

We next investigated the source of larger-than-expected standard deviations of empirical null. By inspecting mRNA levels across all conditions for each gene, we found a wide-spread fluctuation of the mean levels across perturbations (Fig. 3h, Supplementary Fig. 2b). This mean fluctuation is different from the random variation between replicates (i.e., dispersion), as replicates within the same condition behave consistently. The mean fluctuation is typically mild in its effect size, thereby distinguishing it from specific changes induced by RNAi (Fig. 3h, *acdh-1*). We hypothesized that the broad Wald statistic distribution is caused by such mean fluctuations, and that the mean of observed gene expression (μ_{obs}) is the sum of the actual biologically relevant expression change (μ_{bio}, caused by RNAi) and the aforementioned mean fluctuation ($\Delta\mu$) (Fig. 3h). Such fluctuation can be driven by experimental confounders, such as subtle differences in temperature in different positions in the culture plates, that are known hidden covariates in large-scale, array-based experiments[6]. As hidden covariates may be unknown and/or fully confounded with the covariate of interest (RNAi), their effects artificially contribute to the effect from the RNAi treatment they covary with and can be misinterpreted as biological signal from the RNAi treatment in WPS, resulting in the systematic overestimation of *P* values in regular DESeq2 analysis. Importantly, although these fluctuations are often statistically significant, they should be biologically uninteresting, based on the empirical null principle[47].

To test the hypothesis that fluctuations of mean can introduce the observed test statistic inflation, we performed a simulation study. We used scDesign3[48] to simulate the metabolic-gene WPS dataset[31]. To mirror real data, we used the DEGs identified in the WPS analysis as the ground truth and synthesized a new dataset in which the mean expression of each DEG was altered based on the DEG fold change while remaining consistent otherwise. To test the role of $\Delta\mu$, we further introduced a random fluctuation of the mean for each gene in each condition, based on parameters estimated from the real data (Supplementary Fig. 2c). Consistent with our hypothesis, we found that the standard deviations of the empirical null distributions in simulated data matched the inflated test statistic of real data when a random $\Delta\mu$ was added, while being very close to the theoretical distribution when removing $\Delta\mu$ (Fig. 3i, Supplementary Fig. 2d–f). Therefore, the observed anti-conservative *P* values can be explained by a random mean fluctuation.



While the mean fluctuation is systematic, some genes have stronger test statistic inflation than others, such as the noisy genes we previously observed (Fig. 3d, g). We further investigated if there are any common features among these genes, and in general, we found that the test statistic inflation was not dependent on their expression levels. Instead, we noticed that the 'noisier' or highly fluctuating genes were often associated with stress and environmental responses[49,50] (Fig. 3j). This observation was substantiated by Gene Set Enrichment Analysis (GSEA)[51], using empirical-null standard deviation as the ranking metric (Supplementary Methods, Fig. 3k). Enriched categories are related to various nutrients and waste transport, as well as interactions with bacteria diet. While these responses might also result from

Fig. 3 | EmpirDE analysis framework reveals systematic anti-conservative *P* values caused by a deviation from the expected distribution of Wald test statistics. **a** Distribution of the number of DEGs in non-targeting perturbations (NTPs) identified by DESeq2 analysis (FC > 2, adjusted *P* value (P_{adj}) < 0.01). Examples of a control-outlier (**b**) and noisy (**c**) gene. Each bar plot shows the expression levels of the gene of interest in a WPS sequencing library that includes RNAi perturbations and vector control conditions. **d** Schematic illustrating the EmpirDE framework. The zoom-in windows show two example genes shown in (**b**, **c**). For (**b**–**d**), each dot in the bar plot represents one biological replicate ($n = 2$ or 3). **e** The number of control-outlier genes per WPS sequencing library. Distribution of fitted means (**f**) and standard deviations (**g**) for all genes in the empirical null modeling. The blue dashed line shows the values for theoretical null. **h** Examples showing the random fluctuation of the mean for two genes. **i** Distribution of fitted standard deviation using simulated WPS data with (right) and without (left) adding a random fluctuation of the mean in the simulation. **j** Scatter plot showing the fitted standard deviation of each gene against expression levels in wild-type condition. Genes exhibiting a standard deviation greater than 2 are colored based on their WormCat categories. **k** Gene Set Enrichment Analysis (GSEA) result using the fitted standard deviation as the ranking metric (Supplementary Methods). WormCat Level 3 was used for the analysis. NES Normalized Enrichment Score.

genetic perturbations, they are well-known, and more likely, to be influenced by environmental factors, such as hidden covariates in the experiment.

EmpirDE framework rigorously controls FDR

To benchmark the performance of EmpirDE, we first used simulation data to assess both FDR and power. As expected, both EmpirDE and regular DESeq2 correctly controlled the FDR on the simulated data without the addition of $\Delta\mu$ (Supplementary Fig. 3a, b). However, with $\Delta\mu$, only EmpirDE was able to rigorously control the observed false discovery proportion (FDP) at the expected FDR (Fig. 4a). Importantly, the power of EmpirDE is greater than that of DESeq2 at the same level of observed FDP (Fig. 4b), indicating a true increase in performance instead of nominal rescaling of *P* values. Such rigorous control of FDR depends not only on the optimized statistical modeling framework but also on the proper adjustment for multiple testing. WPS experiments involve both simultaneously testing the expression of thousands of genes in each perturbation (column-wise multiple testing) and hundreds of gene perturbations (row-wise multiple testing). Using the simulation data, we found that the worst-case adjusted *P* values in both column-wise and row-wise adjustments of a DE test aligned best with the expected FDR (Fig. 4c), while its loss of power was negligible (Supplementary Fig. 3c).

Next, we experimentally benchmarked EmpirDE performance. We first used the NTP experiments mentioned above (Fig. 3a) to compare the number of false positive DEGs between DESeq2 and EmpirDE analysis with different thresholds for both FDR and fold change (Fig. 4d). We found that the 90% quantile of the number of DEGs detected in the NTPs was much lower with EmpirDE compared to DESeq2 (Fig. 4d, Supplementary Fig. 3d). Specifically, at a FC of 1.5 and FDR < 0.1 we detected 4 and 435 false positive DEGs in the EmpirDE and DESeq2 analysis, respectively (Fig. 4d, white dashed line).

To further evaluate the performance of EmpirDE, we used the reproducibility of DEG calls to empirically evaluate the power and error of the DE analysis. We randomly selected and independently repeated, in triplicate, 36 RNAi experiments that yielded a broad range of DEGs (Fig. 4e). DEGs that were identified in one experiment but had no significant change in the other (FC < 1.1 or in reversed direction), are considered genuine false discoveries. We used the rate of such irreproducible DEGs to estimate the true FDR and found that EmpirDE showed a rate of irreproducible calls consistent with the FDR threshold (10%), regardless of the effect size (number of DEGs) (Fig. 4f). In contrast, DESeq2 analysis achieved the desired control of FDR only when the effect size was large. The rigorous control of false positives of the two-pronged EmpirDE approach can be further demonstrated by visually inspecting each of the 36 conditions (Fig. 4g, h, Supplementary Fig. 4). For instance, many DEGs in the *metr-1* RNAi experiment were control-outlier genes and, as expected, these did not replicate in the repeat experiment (Fig. 4g, right side). Although non-control-outlier DEGs were generally reproduced with both DESeq2 and EmpirDE, the latter still eliminated a few highly changed but unreproduced calls (Fig. 4g, left). Notably, the EmpirDE approach was critical when true positives were sparse and false positives identified by DESeq2 analysis masked the retrieval of true positives (Fig. 4h). Finally, these analyses also facilitated EmpirDE parameter optimization, for instance, selecting an optimal threshold to determine control-outlier genes in the first step of the framework (Supplementary Methods, Supplementary Fig. 3e, f).

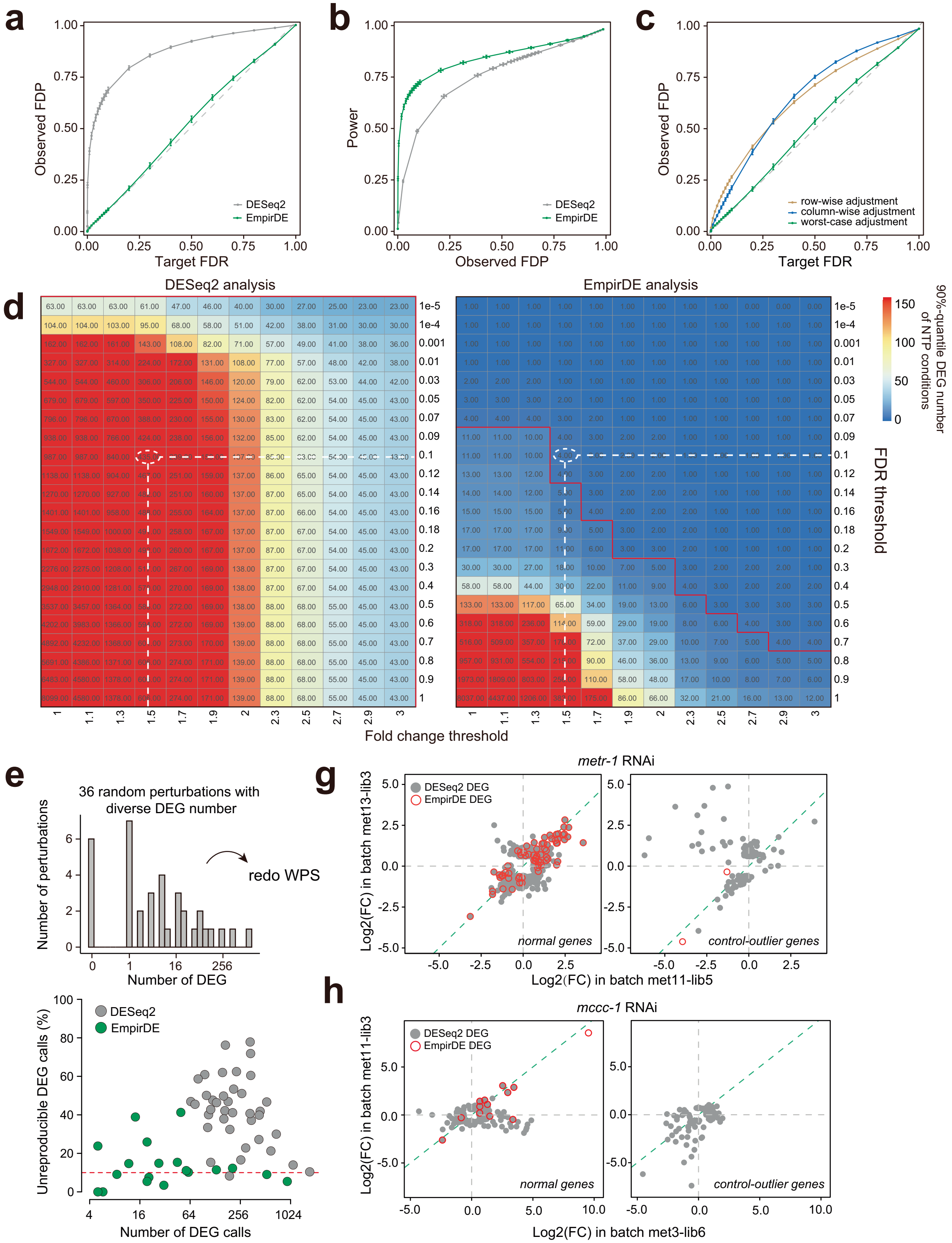
DESeq2 is widely used and therefore we asked what could drive its relatively poor performance in our benchmarking analysis. First, we noted that the mean fluctuations that drive the inflated test statistic generally have small effect sizes (Fig. 3h, Supplementary Fig. 2c). Consistently, when a commonly used and more stringent threshold was applied (FDR < 0.01 and FC > 2), the reproducibility of DEGs from regular DESeq2 analysis was increased (Supplementary Fig. 5a). However, a substantial portion of these DEGs still remain unreproducible. We reasoned that the remaining false positives, which had large effect sizes, might involve genes influenced by the confounded controls (Fig. 3b). Thus, we further applied the first step of EmpirDE framework to regular DESeq2, i.e., cleaning up control-outlier genes using control-independent DE analysis (Fig. 3d). This time the regular DESeq2 analysis also resulted in low false discoveries (Supplementary Fig. 5b). Similarly, using NTP benchmarking, we observed that cleaning up the control-outlier genes decreased the false positives by half under all thresholds, however, only the two-pronged EmpirDE achieved near-complete elimination of false positives (Supplementary Fig. 5c).

Taken together, with the EmpirDE framework that uses gene-specific empirical null models, WPS can robustly assign DEGs for large numbers of perturbations with high signal-to-noise ratio and rigorously controlled FDR for perturbations eliciting from a few to thousands of DEGs.

A proof-of-principle of WPS with 103 NHRs

NHR transcription factors (TFs) play important roles in various physiological processes including metabolism, development, and homeostasis[52]. The *C. elegans* genome is predicted to encode more than 250 NHRs, making it the largest TF family. In contrast, the human genome encodes only 48[53–55]. Although many NHRs have been studied in *C. elegans*[21,56–60], more than half remain completely uncharacterized[52,61]. We analyzed the expression levels and patterns of all 288 predicted *C. elegans* NHRs and selected 103 for WPS that are expressed at relatively high levels both in the whole body and in the intestine and/or hypodermis, tissues highly suitable for RNAi[62] (Fig. 5a, b).

WPS analysis of these NHRs yielded a gene regulatory network (GRN) comprising 6778 interactions between 101 perturbations and 3,673 genes, with in- and out-degrees following expected distributions[63,64] (Fig. 5c, d, Supplementary Data 4). We found that ~80% of perturbed NHRs (81) were responsive (≥5 DEGs, a conservative threshold compared with the false positives in the NTP analysis, Fig. 4d). This rate is much higher than that reported in whole-genome single-cell Perturb-seq experiments (~30%)[13], and double than what we found in metabolic gene screens (40%)[31], indicating that the majority of the 103 NHRs tested are actively regulating gene expression in adult animals. Most NHR knockdowns resulted in a moderate number of DEGs (5–100) (Fig. 5d) and the magnitude of gene expression changes



was also modest (Supplementary Fig. 6a). Notably, 70% of DEGs identified in more than one perturbation changed in the same direction (up or down, Supplementary Fig. 6b). As expected, genes that were mostly down-regulated are expressed at higher levels than those that are mostly up-regulated in NHR perturbations (Supplementary Fig. 6c).

We used WormCat[65] to identify biological processes enriched in the DEGs for each of the 54 NHRs that yielded more than 10 DEGs and found that many NHRs affected genes involved in stress response and metabolism, specifically pathogen response and lipid metabolism (Fig. 5e, f and Supplementary Fig. 6d, e). These observations indicate that several NHRs may function to establish and/or maintain metabolic

Fig. 4 | EmpirDE rigorously controls FDR. a, **b** Benchmarking the performance of EmpirDE analysis framework. The observed False Discovery Proportion (FDP) is compared to target FDR (**a**) and power (**b**). The full metabolic-gene WPS dataset (3691 samples) was simulated 10 times with random mean fluctuation ($\Delta\mu$) to produce the error bars of each metric. FDP and power were measured based on a pooled set of 117,096 simulated DE changes in all conditions (Supplementary Methods). **c** Benchmarking FDR control of different multiple testing adjustment strategies. The data points and error bars in (**a**–**c**) indicate mean ± s.d. from 10 simulations. **d** Evaluating false discoveries using NTP experiments. We estimated false discoveries using the 90% quantile of the numbers of DEG across 71 NTP conditions (shown in the heatmap color and numbers). The 90% quantile represents the value below which 90% of the data points fall, effectively capturing the upper range of typical DEG counts while excluding the most extreme outliers. The red lines show the threshold boundary for five false positive DEG calls. **e** Number of DEGs (defined by FDR < 0.1 and FC > 1.5 using EmpirDE) for 36 perturbations that were repeated by a second WPS experiment. **f** Fraction of unreproducible DEGs for EmpirDE versus DESeq2 analysis. Unreproducible DEGs were defined by genes that are called as DEG in one experiment (FDR < 0.1, FC > 1.5) but confidently non-DEG in the other (FC < 1.1 or show a different FC direction). The red dashed line shows the theoretical FDR (FDR = 0.1). Comparison of log2(FC) measured in two independent experiments for representative RNAi with either high (**g**) or moderate (**h**) number of DEGs. The green dashed line indicates the diagonal (y = x).

functions and to prime the animal to respond to different stressors. The knockdown of individual NHRs was associated with several other WormCat categories as well. Some of these were known, including the association of *nhr-31* with the lysosomal vATPase, *nhr-49* with lipid metabolism, and *nhr-10, 68, 114*, and *101* with the propionate shunt pathway[59,66–68] (Supplementary Fig. 6d, f). To assess the recall of known NHR's function more quantitatively, we also compared our WPS data with a published dataset for *nhr-25* perturbations[69] and observed a good concordance (Supplementary Fig. 6g).

We wondered whether the regulation of stress response and metabolic genes was due to a few hub genes that are affected by many NHR perturbations (Supplementary Fig. 6h). We identified a set of hub genes (in-degree >5) that are annotated in WormCat as either stress response or metabolic genes and found that different NHRs influenced the expression of distinct stress and metabolic genes (Fig. 5g, h). Together with the in-degree distribution that shows that most genes are regulated only by few NHRs (Fig. 5d), this analysis indicates that the enrichment for stress response and metabolic genes in the NHR GRN is not driven by a few common genes. Interestingly, we found that the GRN mostly consists of activating interactions, i.e., upon knockdown of an NHR, gene levels tend to go down. However, on average, we found that stress response gene expression was increased upon NHR knockdown (Fig. 5i, j, Supplementary Fig. 7a–f). These results indicate that NHRs activate the expression of metabolic genes, especially those involved in lipid metabolism, while they downregulate stress response genes, either directly or indirectly.

Of the 288 *C. elegans* NHRs, at least 269 are homologs of HNF4[70]. This observation raises the question whether these NHRs regulate similar targets, or whether they evolved distinct and diverse functions. To answer this question, we compared the target genes of the 81 NHRs in the GRN (Supplementary Methods) and found that these NHRs not only regulated various sets of targets, but also clustered into modules consisting of distinct NHR pairs, which we named 'pairwise modularity'. In fact, 52 of 81 NHRs (64%) shared a significant overlap only with one other NHR (Fig. 6a, b and Supplementary Figs. 8, 9a, b, Supplementary Data 5), and this pairwise modularity was statistically significant based on a randomization test (Fig. 6c, Supplementary Fig. 8). We also ensured that this observation was not due to off-target effects based on the gene expression changes and anti-sense RNA signals of the counterpart NHR (Supplementary Fig. 10a, b). We identified numerous NHR pairs for which functional relationships were not yet known and that provide hypotheses for further study (two examples in Supplementary Fig. 11a, b). Importantly, this observation was facilitated by the EmpirDE framework because it increased the signal-to-noise ratio compared to DESeq2 (Supplementary Fig. 8). Thus, even with a relatively low number of perturbations (~100 RNAi conditions), EmpirDE can effectively increase the interpretability of the data.

Interestingly, NHR sequence similarity only correlated with few pairs that shared target genes. Overall, protein sequences of the NHR DNA binding domains showed relative low similarity to each other (percent identity <0.5), with only limited number of clusters (Fig. 6d). Although the few pairs with high protein sequence similarity were more likely to share targets (e.g., NHR-10, NHR-68, NHR-114, and NHR-101) the protein similarity between most NHR pairs was lower and did not correlate with similarity in their target genes (Fig. 6e, f). Remarkably, even some NHRs from different evolutionary origins shared target genes (e.g., *nhr-107* and *nhr-41*, Fig. 6a). We found that the similarity among different NHRs correlated better with their expression patterns (Fig. 6g). Therefore, the pairwise modularity unveiled by WPS may be affected more by mechanisms involved in the regulation, and less by the biophysical properties (e.g., DNA binding domains), of these NHRs.

The pairwise modularity suggests that NHRs may form 'AND-logic gates' in regulating gene expression, where two NHRs are both required for downstream gene regulation. Indeed, we have previously discovered that *nhr-10* and *nhr-68* function in an AND-gated feedforward loop to detect the persistent accumulation of propionate[67]. Interestingly, we found that *nhr-10*, *nhr-68*, *nhr-101* and *nhr-114* clustered together in a module, with *nhr-68* and *nhr-101* being the most similar. Therefore, we hypothesize that the AND-logic may be extended to *nhr-68* and *nhr-101* (Fig. 6a). As a preliminary test, we performed WPS with double *nhr* knockdowns. We first confirmed the known AND-logic connection between *nhr-10* and *nhr-68*, based on the lack of additive effects on gene expression (Supplementary Fig. 11c, observed FC substantially lower than the additive FC). Next, we tested the double knockdown of *nhr-68* and *nhr-101*, which also showed a lack of additive effects, supporting the idea that these two genes also function in an AND-logic gate. Future studies based on mutant strains and other phenotypical readouts will provide further validations for both *nhr-68* and *nhr-101* and all other pairs identified in our data.

Discussion

In this study, we provide a WPS platform that combines strengths of multiplexed bulk RNA-seq with high-throughput whole-animal gene perturbations by RNAi. WPS is both efficient and cost-effective, e.g., a 2-week timeframe for collecting 96 perturbations in triplicate and more than 10-fold cost reduction compared with conventional methods (Supplementary Protocols), which enables replicate screens with full transcriptome readouts of hundreds of perturbations in a living animal. Future screens with additional RNAi libraries, different bacterial diets, and supplementation of metabolites or drugs will provide insights into how the animal responds to a variety of perturbations.

A key advantage of WPS is that it is based on whole-organism in vivo perturbations. While this is not feasible in mammals, it should be applicable to organisms amenable to large-scale RNAi screens, such as *Drosophila*. However, we do envision that WPS-like screens will be feasible in bulk in tissue culture cells, especially when smaller sub-libraries of genes (~100) are selected for perturbations. Another key feature of WPS is the EmpirDE framework that uses an empirical null for each detected gene and that can be applied due to the scale of WPS experiments, and which can alleviate systematic errors such as confounding experimental covariates. Although the concept of empirical

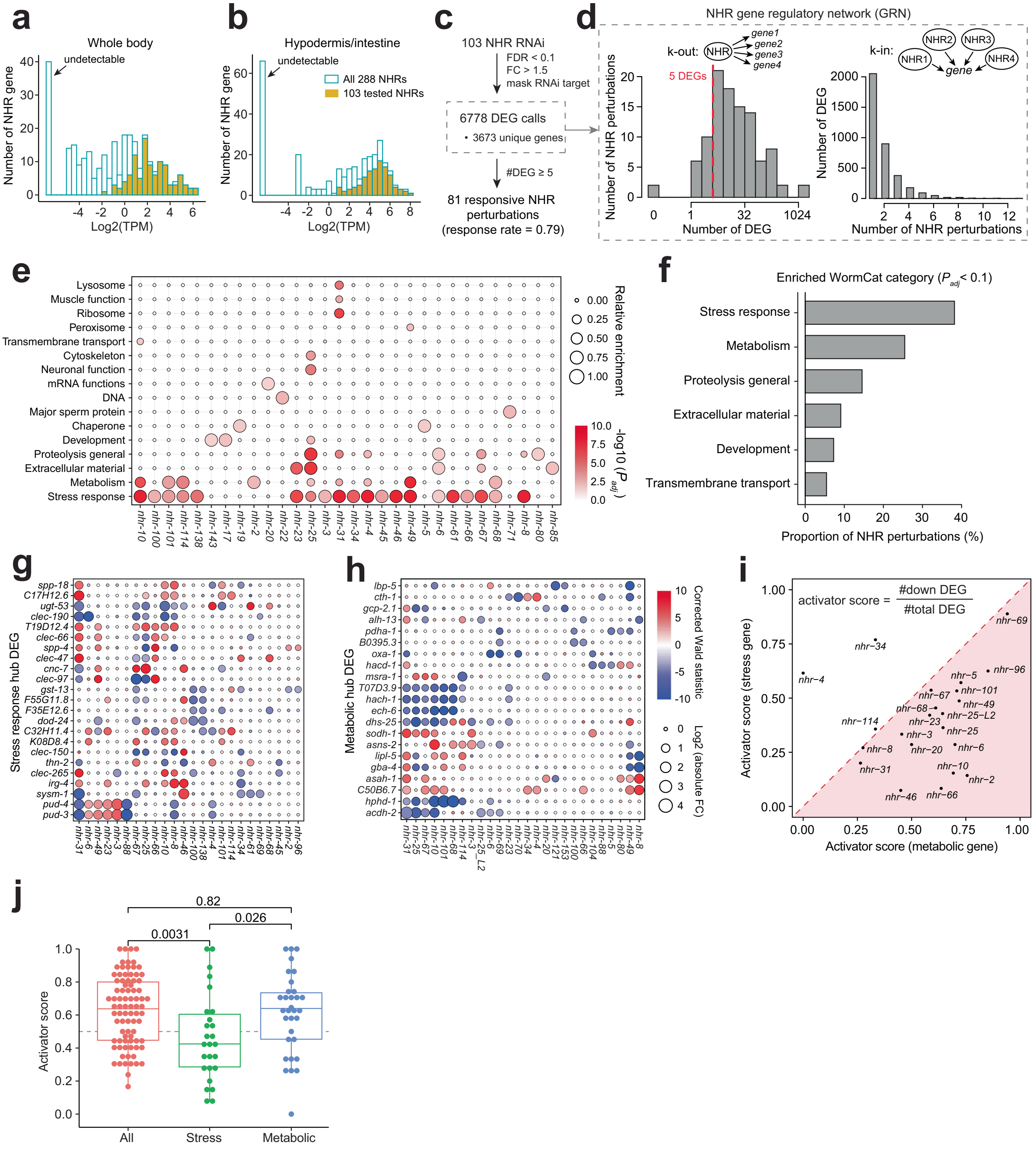


Fig. 5 | A NHR gene regulatory network. a Expression levels of *nhr* genes in adult animals. The TPM was quantified by the reference profile used in Supplementary Fig. 1c. **b** Tissue expression of *nhr* genes based on an adult stage single-cell RNA-seq data[90]. The maximal TPM of hypodermis and intestine expression is shown in the plot. **c** Summary of NHR WPS experiments. **d** Distribution of the number of DEGs in NHR perturbations (i.e., k-out) and the number of NHRs regulating the same gene (i.e., k-in). The red dashed line indicates the responsiveness threshold (≥ 5 DEGs). **e** Functional enrichment analysis of NHR RNAi conditions. Only the 54 NHRs (55 perturbations because *nhr-25* was perturbed at both L1 and L2 stages) with more than 10 DEGs were analyzed to ensure power. Conditions without significant ($P_{adj} > 0.05$) enrichment are not shown in the plot. The relative enrichment is defined as the −log10(P_{adj}) normalized by its maximum in a given RNAi condition. **f** Proportion of testable RNAi (i.e. >10 DEGs) that showed significant enrichment ($P_{adj} < 0.1$) in WormCat Level 1. The *P* values in (**e**, **f**) were derived from one-sided hypergeometric test with multiple testing adjusted by Benjamini-Hochberg method. Visualization of hub genes (i.e., regulated by >5 NHRs) of the stress response (**g**) and metabolic (**h**) categories. **i** Comparison of the activator score for metabolic (x-axis) and stress response (y-axis) DEGs. NHR perturbations with ≥5 DEGs in the corresponding categories were analyzed. **j** Overall comparison of activator scores for different DEG categories. Each dot represents an activator score for one NHR, calculated using DEGs from the category shown on the X-axis (All: $n = 83$; Stress: $n = 26$; Metabolic: $n = 32$). Boxes show the IQR (25th–75th percentiles) with median line; whiskers extend to 1.5×IQR. Two-tailed Wilcoxon tests were performed to calculate the *P* values. Source data are provided as a Source Data file.

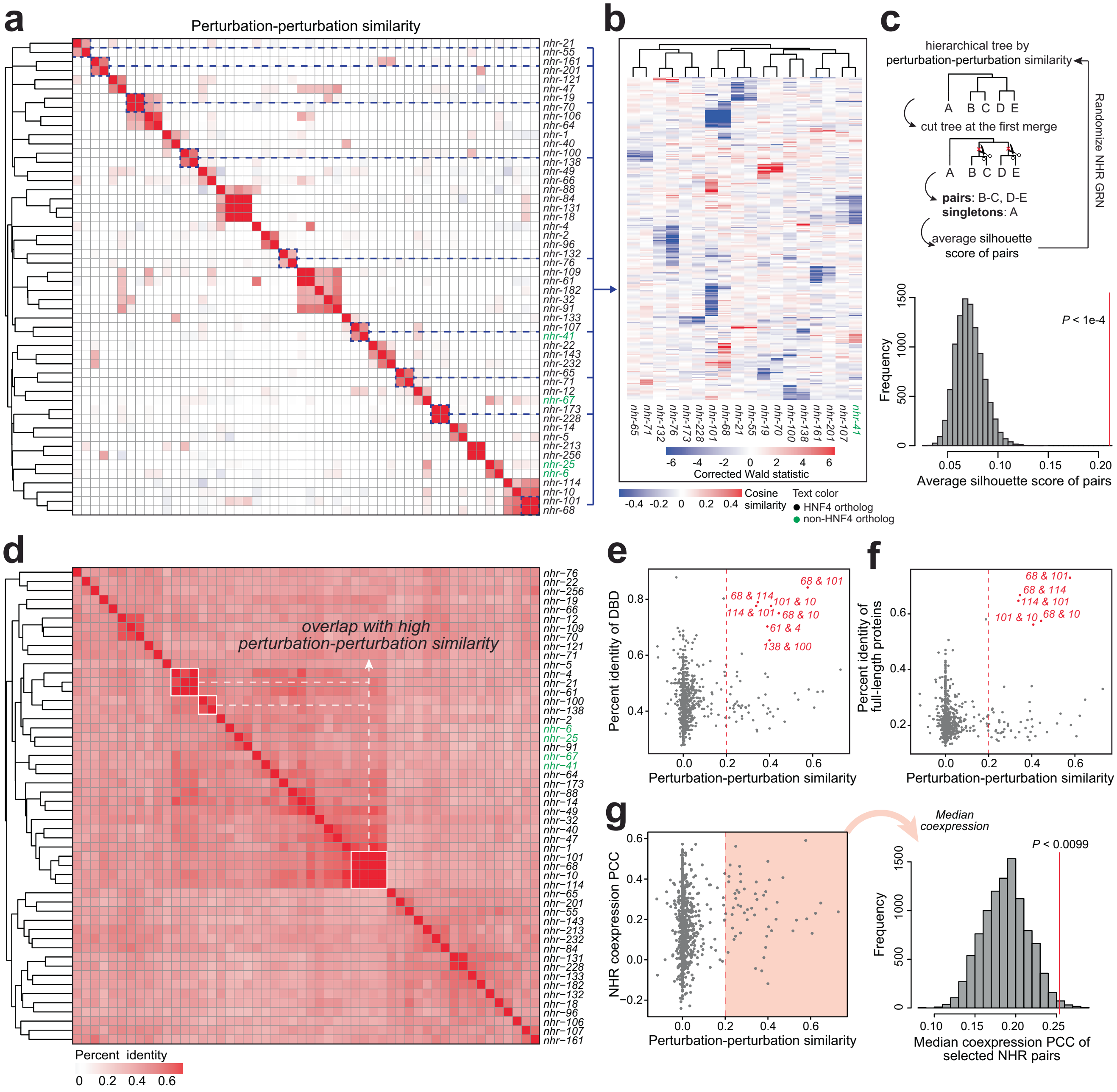


Fig. 6 | The pairwise modularity of NHRs. a Heatmap depicting perturbation-perturbation similarity of DEG profiles for the NHR perturbations. The perturbation-perturbation similarity was defined by cosine similarity of the filtered log2(FC) profile. The filtered log2(FC) was derived by masking the log2(FC) values of genes that are not called as DEGs (FDR < 0.1, FC > 1.5) to zero. **b** Visualization of gene expression changes in selected NHR pairs. The gene expression change was measured by the corrected Wald statistic. Rows are the union DEGs of these selected NHR perturbations. **c** Randomization test of the pairwise modularity of NHR gene family. The schematic shows the design of the randomization test. Histogram shows the average silhouette score of pairs in 10,000 randomizations. The red line indicates the observed score from real data. The NHR GRN was randomized by swapping the network edges while preserving the network structure and properties, such as in- and out-degrees (Supplementary Methods). The gene-gene correlation was not preserved in this randomization to fully randomize the GRN. **d** Heatmap depicting protein sequence similarity (percent identity) for the DNA binding domain (DBD) of NHRs. The heatmap was clustered using distance matrix generated by Clustal Omega online tool from EMBL-EBI[87] (Supplementary Methods). Scatter plots showing the comparison between perturbation-perturbation similarity and sequence similarity of DBD (**e**) and full-length protein (**f**). Each data point indicates a pair of NHRs and selected pairs are labeled. **g** Scatter plot and randomization test for the associations between perturbation-perturbation similarity and NHR coexpression. Coexpression was measured based on the median Pearson Correlation Coefficient (PCC) of *nhr* gene expression in a compendium of *C. elegans* gene expression data across various conditions[68] (Supplementary Methods). The median coexpression level of pairs with a cosine similarity greater than 0.2 (red region in (**g**)) is calculated and compared with that from randomized data (Supplementary Methods). The histogram shows that the median from real data (red line) is significantly greater than that from randomized data, indicating a statistically significant association between NHR coexpression and perturbation-perturbation similarity. Source data are provided as a Source Data file.

null has been widely applied in genomics[71–73], to the best of our knowledge, it has not been used to directly model the test statistic distribution at the level of individual genes (features), possibly due to the lack of large systematic data like those generated with WPS. EmpirDE exploits the unique power of having many conditions (>100), each with three replicates, to achieve rigorous statistical analyses. Conventionally, such level of statistical rigor is only achievable with a high number of replicates (e.g., 8–12)[74].

By combining large-scale data, NTPs, and repeated experiments, we systematically identified two sources of false positives in DE analysis. The first source, control outlier genes (Fig. 3b), arise from confounded control samples and results in false DEGs that cannot be easily filtered out by thresholding FDR or FC. While this is an intrinsic confounder for arrayed experiments, this issue can also result from any specific treatment of controls, whether intentional or unintentional. The second source is systematic fluctuations in mRNA levels across conditions (Fig. 3h, i), and is likely related to hidden covariates in the experiment. These effects are typically small in magnitude and can often be filtered by thresholding FC (e.g., FC > 2). However, they systematically compromise the statistical rigor of DE analysis, resulting in anti-conservative *P* values. Notably, recent studies by other groups have also reported a prevalence of anti-conservative *P* values when parametric DE models are used, even in regular experiments[41,42]. We believe that while both of the two issues could be more specific to high-throughput screens, they might have been simply overlooked in regular, small-scale experiments, where these false positives may not be identified in the first place.

Our benchmark analysis (Fig. 4) should not be interpreted as a challenge to the well-established statistical foundation of DE analysis, such as the negative binomial model or generalized linear model (GLM) used in DESeq2. As discussed above, we demonstrate that most uncontrolled false discoveries stemmed from confounding effects in the experiments. Unlike regular batch effects, which are orthogonal to the variables being tested and can be corrected using a GLM in DESeq2, these confounding effects are entangled with the biological effect of interest, cannot be directly corrected, and result in high level of false discoveries. EmpirDE complements DESeq2 in handling these issues to achieve bona fide FDR control. Importantly, EmpirDE is agnostic to the source of confounders, effectively reducing both false positives and false negatives, demonstrated by simulation and by experimental benchmarking. This statistical rigor allows us to confidently identify DEGs, even for those with small effect sizes. Therefore, EmpirDE provides a robust and rigorous DE solution that should be broadly applicable to large-scale studies.

By applying WPS to more than 100 NHR perturbations, we discover a pairwise modularity in which two or more NHRs regulate the expression of overlapping sets of genes, which cannot be explained by protein (and presumably binding site) similarity. Instead, this pairwise modularity suggests that 'AND-logic gates' are a common mechanism of gene regulation in *C. elegans*. Future studies with other TFs will be important to see if this is a general principle, or if it is a specific feature of NHRs. Knockdown of many NHRs affected only few genes, suggesting that these TFs may either not be active under the conditions tested, or are truly specialized in their regulatory function. In two companion studies[31,75], we further validated WPS by perturbing ~900 metabolic genes. These studies generated high-quality, highly interpretable datasets, providing tremendous insights into metabolic wiring and rewiring at a systems level. Notably, WPS interrogates gene functions in vivo, thus linking genes to their native physiological roles. For instance, using metabolic gene WPS data, we identified an unconventional central carbon metabolism that consumes ribose, rather than glucose, from dietary RNA and through the pentose phosphate pathway. Together, we envision that WPS-style in vivo functional genomics will provide a powerful tool to uncover gene functions in living organisms.

Methods

C. elegans strains and maintenance

N2 strain was used as the wild-type strain. Animals were maintained at 20 °C on solid nematode growth media (NGM)[76] and fed *E. coli* HT115 containing the empty RNAi vector L4440[18].

RNA interference

To construct WPS RNAi libraries, each RNAi bacteria strain was cherry picked from a parent library (e.g., the metabolic RNAi library[23], or TF RNAi library[21]) and streaked onto LB agar plate with 50 µg/mL ampicillin to produce single colonies. A single colony for each RNAi was used in the WPS RNAi library. RNAi was performed accordingly as described with slight modifications[62]. Briefly, bacteria were cultured overnight at 37 °C in 1 mL LB supplemented with 50 µg/mL ampicillin in a 96-deep well plate. 100 µL of each culture was then diluted 50-fold using fresh LB medium with 50 µg/mL ampicillin in a well of a 24-deep well plate. After incubating for 4 h at 37 °C, bacteria were centrifuged at 3000 *g* for 20 min in a Beckman Coulter Avanti® J-26XP High-Performance Centrifuge with a JS-5.3 Swing Bucket Centrifuge Rotor, and the pellet was resuspended in 200 µL M9. The resuspended bacteria were then transferred to 6-well NGM plates containing 50 µg/mL ampicillin and 2 mM Isopropyl β-d-1-thiogalactopyranoside (IPTG, Fisher Scientific) for induction of double-stranded RNA (dsRNA) expression. Plates were dried in a hood and incubated overnight at room temperature.

For developmental stage time course experiments, all animals were fed bacteria with vector control RNAi. Approximately 2500 synchronized L1 animals were plated for collecting L2 animals; ~1000 synchronized L1 animals were plated for collecting L3 animals; ~500 synchronized L1 animals were plated for collecting L4 animals; and ~200 synchronized L1 animals were plated for young adult and gravid adult samples.

For all other WPS experiments, approximately 200 synchronized L1 animals were plated into each well, followed by incubation at 20 °C for ~63 h. In the cases where RNAi feeding led to a developmental delay phenotype, synchronized L1 animals were initially fed with vector control RNAi. After a period of 17 h post-plating (i.e., at the L2 stage) or 25 h post-plating (i.e., at the L3 stage), animals were transferred to the corresponding RNAi plates to circumvent RNAi-associated developmental delay. Animals usually develop normally after such delayed RNAi exposure. Only RNAi conditions without notable developmental delay were sequenced.

RNA extraction in 96-well plate

We developed a 96-well RNA extraction method for *C. elegans* tissues while leaving all eggs intact. Please refer to Supplementary Protocol for details.

WPS sequencing library construction

We adapted the CEL-Seq2 single-cell RNA-seq library construction protocol[25] for WPS sequencing library construction. We meticulously optimized each step of the protocol to ensure robustness and reproducibility. As part of this optimization, we modified the adaptor sequences of the CEL-Seq2 primers to ensure compatibility with both Illumina and BGI platforms for sequencing. For a comprehensive description of the modified protocol and primer sequences, please refer to the Supplementary Protocol.

WPS sequencing library design

We used an Illumina NextSeq sequencer capable of providing ~350 million reads (or its equivalent from BGI), which allowed us to pool ~50 samples to obtain an average coverage of ~7 million raw reads/sample. Typically, a library includes 15-16 RNAi conditions in triplicate and 6 vector control samples. We conducted three biological replicates for all RNAi conditions on different days. In each different-day replication, we included two vector control RNAi samples to minimize the chance of failing in vector control experiment, which would impact data analysis for all RNAi conditions in the same batch. One vector control sample was prepared side-by-side with RNAi conditions, while the other one was independently prepared in the same day, using separate bacteria culture and bleaching *C. elegans* from a distinct parental animal batch. The latter case aligned with criteria for biological replication (fresh material) except that the experiment was conducted on the same day, thus we refer to them as 'same-day replicates'. We noted that the expression variability between different-day and same-day replicates was similar, therefore, all six vector control samples were treated as biological replicates in differential expression analysis to enhance the statistical power (further details are provided in the following sections).

Next generation sequencing

Most WPS sequencing libraries were sequenced on the BGISEQ-500 next-generation sequencer platform with 100-bp paired-end reads. A subset of the libraries was sequenced using an Illumina NextSeq 500 sequencer with a NextSeq 500/550 High Output Kit v2.5 (75 Cycles). For Illumina sequencing, paired-end sequencing was performed with 14 cycles for read 1 and 75 cycles for read 2.

WPS raw data processing

The pair-end reads data were either received from BGI directly or produced through standard bcl2fastq procedure with illumina platform. Reads were processed by an in-house dolphinNext pipeline[77] to generate a gene-by-sample read count matrix. The pipeline includes the following steps: (1) raw reads were demultiplexed by a homemade python script that extracts the barcode information from read 1 and combines that with read 2. (2) The processed reads were passed to Trimmomatic (v0.32)[78], to remove polyA and adaptor sequences. (3) next, reads were aligned to the *C. elegans* genome (WormBase WS279) by STAR[79] (parameter: *--runThreadN 4 --alignIntronMax 25000 --outFilterIntronMotifs RemoveNoncanonicalUnannotated*). (4) Finally, the output bam file was processed by ESAT[80] to obtain the read counts of genes. In ESAT, we used an extension window of 1000 bp and the 'proper' method of multiple mappings. Unique Molecular Identifier (UMI) features were not used by setting *umiMin = 1*. Read counts, rather than UMI counts, were used as the gene expression quantity. We did not observe significant PCR duplicates during the development of our method (i.e., read counts highly correlate with UMI counts, data not shown), which is consistent with the low number of PCR cycles for sequencing library construction (Supplementary Protocol). Therefore, we directly used the read counts regardless of the presence of UMI in our sequencing library. The dolphinNext pipeline also includes a few quality control (QC) procedures for sequencing library and alignment quality and is interactive through the online portal[77]. The pipeline processes each WPS sequence library individually and produces a read count table for the sequencing library. The pipeline can be downloaded at https://github.com/XuhangLi/WPS.

Reads count tables were used as the input for all downstream analyses. Reads from ribosomal RNA (i.e., mapped to ribosomal genes) were discarded. The sequencing library depth was measured with the sum of read counts of each sample after ribosomal gene removal. Samples with depth lower than 1 million were removed.

WPS RNAi identity QC and dsRNA decontamination

We discovered that the reads from dsRNA in RNA interference (mostly in anti-sense strand) could be used to determine the identity of the RNAi clone used. However, these reads might potentially confound the quantification of the RNAi target gene expression since some map to the sense strand at the 3′ end of the gene. In rare cases, they can also influence the quantification of other genes when their transcripts extend to regions containing dsRNA reads. Therefore, we developed a python script to both quantify the dsRNA (anti-sense RNA) signals and the expression levels of the dsRNA-influenced genes. This is feasible because dsRNA signals are confined to the coding region of the target gene, while the mRNA signals predominantly reside at the 3′-UTR of the transcript. We achieved it by identifying genomic regions covered by dsRNA signals and re-quantifying genes that were influenced by only counting the reads in the clean regions.

We performed the dsRNA analysis on a library-by-library basis, ensuring that any re-quantification of gene expression was uniformly applied to all samples within a sequencing library. To identify possibly dsRNA-influenced genes, we searched for genes whose transcripts overlapped with the exons of any RNAi target gene in the sequencing library. This gave a set of genes to be corrected for potential dsRNA contamination. Next, we identified the genomic regions contaminated by thresholding the reads mapped to the complementary strand of the mRNA for each RNAi-targeted gene. Finally, the read counts of all potentially contaminated genes were recounted using the clean (not contaminated) regions only.

dsRNA signals were quantified by counting reads mapping to the complementary strand of the mRNA(s) for each RNAi-targeted gene. This dsRNA quantification procedure was applied to all metabolic genes to identify the potential cross-contamination and sample swaps. When applicable, the procedure was applied to a control bam file that was made from an RNA-seq library of animals treated with only vector control, establishing the background level of reads mapping to the complimentary strand for each gene and was used to calculate the enrichment of dsRNA (anti-sense RNA) signal in the RNAi identity QC. The control sequencing library used in our study was the developmental stage sequencing library (see the corresponding section below for details).

Since the dsRNA-influenced genes were quantified solely by reads mapped to the clean region, it may significantly reduce the total reads (depth) for a gene, potentially resulting in a loss of power. Therefore, we applied such dsRNA de-contamination only to genes whose loss of depth was less than 50% (recounted read counts in vector controls were greater than or equal to 50% of the original read counts). Genes that were not corrected were noted, and additional scrutiny was applied when evaluating their RNAi efficiency. The dsRNA (anti-sense RNA) analysis is available in WPS data analysis pipeline (https://github.com/XuhangLi/WPS).

WPS RNAi efficiency QC

We performed QC of RNAi efficiency based on two complementary criteria: the reduction of reads for the targeted gene and/or the detection of target anti-sense RNA. A reduction in reads for the targeted gene may not always be observed even if the RNAi is successful because the gene is lowly expressed or if the expression quantification is influenced by dsRNA and cannot be decontaminated (see above). Therefore, we considered an RNAi-condition to pass QC when there was either a two-fold decrease in reads of the targeted gene and/or a greater than 10-fold increase in anti-sense RNA signals corresponding to the targeted gene. To simplify this quantification, we calculated the fold change simply by dividing the TPM of the targeted gene in the RNAi condition by that in the vector control condition within the same batch. Similarly for anti-sense RNA signals, we divided the anti-sense RNA count-per-million (CPM) in a RNAi condition by the background anti-sense RNA CPM based on a vector-control-only sequencing library that was described in the previous RNAi identity QC section.

We also used the anti-sense RNA signal to identify potential cross-contaminations (i.e., one condition contains anti-sense RNA mapping to two genes). Such cross-contaminated samples were rare and were either labeled as 'MULTIPLE' in the sample metadata and included in the dataset or removed. Together, the QC pipeline outputs a list of failed-QC conditions and evaluation figures (such as the heatmap of anti-sense RNA) for manual interpretation. All QC results were carefully inspected to ensure the quality of the dataset.

For any RNAi conditions that did not pass RNAi QC, we performed Sanger sequencing of the RNAi clone. We found these fail-QC RNAi carried plasmids that (1) contain an insert lacking at least 100 consecutive base pairs targeting to a *C. elegans* gene ('SHORT); or (2) contain an insert that do not target to any *C. elegans* genomic region ('VECTORLIKE'); or (3) contain a recombined vector that lacks the T7 promoter, therefore deficient in expressing dsRNA ('RCBVECTOR'); (4) contain an insert that targets to multiple *C. elegans* gene ('MULTIPLE'); (5) undefined RNAi identity because the Sanger sequencing did not return a signal ('NOSIGNAL'), or (6) contain the RNAi insert that targets to another *C. elegans* gene. The last were relabeled in the metadata table and included in the final dataset. The erroneous RNAi such as short inserts were relabeled with specific prefix (e.g. 'SHORT_') in the sample name and was used in the analysis when applicable (e.g. forming the set of non-targeting perturbation (NTP)).

As a showcase of the frequency of these fail-QC perturbations, we found among the 3784 samples generated in the metabolic WPS experiment, 76 (2.0%) were removed due to low depth (<1 million) or bad quality (see below), 89 'SHORT' (2.4%), 71 'VECTORLIKE' (1.9%), 38 'RCBVECTOR' (1.0%), 33 'MULTIPLE' (0.9%), 20 'NOSIGNAL' (0.5%) and 254 (6.7%) swapped to targeting another *C. elegans* gene. Together, the on-target pass-QC rate for a large-scale WPS is expected to be ~85% (3203/3784) including vector control samples.

WPS sample quality QC via exploratory data analysis (EDA)

To identify the 'outlier' samples, we performed library-level EDA based on a serial manual inspection of plots based on Principal Component Analysis (PCA), Euclidean distance and Pearson correlation, which is a common practice for RNA-seq analysis (http://bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2.html). We consider a sample to be problematic ('bad sample') if it displayed high distance to other replicates (usually > 50), existed as a clear outlier in PCA plots, and/or showed poor sample-sample correlation (r-squared < 0.95) or had large set of outliers in the inter-replicate gene expression scatter plot. A bad sample usually satisfies most or all of these criteria. We automated the generation of these QC plots but did not automate the identification of bad sample. We reasoned that samples may go wrong in different ways such that the thresholds for one study/experiment may not be applied to another. For instance, in the metabolic WPS data, we noticed that a sample could be an outlier in PCA plot and show significant distance with other replicates, however, displayed good correlation (i.e., r^2 > 0.98) with other replicates. Further investigation found that this is due to the difference in their sequencing depths (i.e., the one is less than 2 million). Therefore, we did not consider such samples as bad samples. Since being interactive is the nature of EDA, this part of QC was designed to require manual inspection of the data by the researcher. This is not a speed limiting step of the data processing as inspecting the plots of one sequencing library usually only takes a few minutes. Of note, bad samples are rare in WPS routine, for instance, only 48 (1.3%) samples were identified in the metabolic WPS dataset.

Control-dependent differential expression (DE) analysis

Typically, a sequencing library includes 15–16 RNAi conditions in triplicate and six vector control samples. As mentioned above, these six vector control samples were collected over three different-day replicates, each compromising two independently cultured, same-day replicates. We initially analyzed the gene expression variance level within the two same-day replicates and found it was similar to that among the different-day replicates (Supplementary Protocol). In addition, we noted that DE analysis solely based on different-day replicates often produced slightly more DEGs compared with using all six samples (data not shown). This may be because it is less prone to underestimating variations with a greater sample size. Together, we reasoned that since using six replicates practically generates more conservative results, and theoretically can be more powerful because of increased sample size, we decided to use the six control samples as six biological replicates in our WPS DE analysis. We acknowledged that the two kinds of replicates may behave differently in the hand of another researcher, so we advise WPS users to carefully evaluate before deciding on using only different-day replicates versus all the six (see Supplementary Protocol).

The control-dependent DE analysis was performed using DESeq2 (v1.26.0[38]). Given that sequencing library construction can introduce batch effects, we conducted DE analysis on a per-library-basis using roughly 50 samples in each run of DESeq2. To mitigate potential batch effects between replicates, replicate batch information (i.e., rep1, rep2, and rep3) was incorporated into the DE model (*design ~ replicate_batch_label + RNAi_condition_label*). Genes with fewer than 10 read counts across all samples in a sequencing library were excluded from the DE analysis. We disabled independent filtering (*independentFiltering = F*) and instead employed a custom filter (see below) for consistency across sequencing libraries. We produced two log-fold-change estimates, including *log2FoldChange* estimates from DESeq2[38] (referred to as raw fold change) and the shrinkage estimates from *apeglm* (*apeMethod = 'nbinomC'*) (referred to as shrunk fold change), which were compared and utilized as descried in the section *WPS analysis parameter selection*.

Together, this control-dependent DESeq2 analysis follows standard procedures of DESeq2 and is also referred to as DESeq2 approach (as compared with EmpirDE approach) or conventional DE analysis in this paper. The outputs here form a foundation for further test statistic modeling in the EmpirDE analysis.

Control-independent differential expression analysis

The idea of control-independent DE analysis is to perform DE analysis by comparing an RNAi condition against all the other samples within the same sequencing library. Given that DE is typically sparse and condition-specific in large scale screening, we expect most if not all genes will be affected, and hence exhibit differential expression, in only a limited number of conditions within a sequencing library (i.e., DE call percentage <30%, meaning the frequency of DE call for any gene is less than 5 out of 16 conditions in a library). Consequently, most conditions in a sequencing library can serve as a null population for DE analysis that does not rely on control labels.

However, if a gene is truly differentially expressed in multiple conditions within the same sequencing library, the power to detect DEGs will be reduced when one condition is compared directly with all others. A more effective approach is to compare an RNAi condition only with a true null population, in which the gene of interest is not differentially expressed. This approach, however, poses a challenge of identifying the main (null) population based on gene expression data. We used AdaTiss[81] for robust fitting of the gene expression across all samples and to exclude samples in which the gene's expression was extreme with respect to this fit. With size-normalized and batch-corrected expression levels (corrected using the *removeBatchEffect* function in *limma* package), we applied AdaTiss to fit the mean and variance for each gene in each sequencing library, one at a time (example command: *out = AdaReg(model.matrix(~1,data = as.data.frame(y)), y), where y represents the expression level vector*). A fit was deemed successful if $\text{pi}_0 \geq 0.7$ (at least 70% of samples were in the main population), and was then used to calculate z-scores for each condition (*z = (y-out$beta.rob.fit)/sqrt(out$var.sig.gp.fit)*). Overall, the rate of successful AdaTiss fitting is usually around 95%. In the case of unsuccessful fitting, we used simple statistics as a surrogate. For genes lowly expressed (median normalized read count ≤10), we used the mean and standard deviation to calculate z-score. For those highly expressed ones (median count greater than 10), we used the median and mad (median-absolute-deviation). The rationale for using mean/sd for lowly expressed genes and median/mad for highly expressed ones is to mitigate the high variance for lowly expressed genes (thus, mean/sd provides a more conservative estimation of the main population) while maximizing the power for highly expressed genes.

To define the null population based on the fitting, we used a z-score cutoff of 2.5 (equivalent to a *P* value of approximately 0.01). For each gene, any condition (including vector controls but excluding the specific RNAi under analysis) with a median z-score (across three replicates) below 2.5 or above −2.5 was included in the null population. Conditions not meeting this criterion were categorized as 'outlier' population. In the uncommon event where over 50% of conditions in a sequencing library were identified as outliers, making it likely that many RNAi affected the gene, we conservatively designated all conditions within that library as the null population to reduce the risk of false positives in such scenarios.

To streamline DE analysis with DESeq2, i.e. to build a DE model using a single gene expression matrix, we replaced the outlier expression values with imputed values derived from inliers (the null population). This strategy was adopted to circumvent the need to run DESeq2 separately for each gene because of different null populations, which would be computationally impractical without modifying the DESeq2 package. A bootstrap strategy was used for imputation. For each expression value in outlier conditions, we randomly selected a corresponding expression value from inlier conditions in the same biological replicate. To do so, we use the DESeq2 normalized counts for inlier samples and multiply the sampled value by the outlier sample size factor, rounded to an integer, to get an imputed count value. This procedure resulted in a new read count matrix, wherein the values for outlier conditions of each gene were replaced with these imputed counts, based on inliers identified through robust-fitting z-scores.

Like the control-dependent DE, genes with fewer than 10 read counts across all samples in a sequencing library were excluded from the analysis of that library. To manage potential single outlier samples within the null population, we enabled the outlier replacement function in DESeq2 by setting *minReplicatesForReplace = 7*. Other DE parameters were identical to those used in vector-dependent DE analysis. DE results were derived by contrasting the targeted RNAi condition against the defined null population.

Combining control dependent and independent DE analysis results in the EmpirDE framework

The EmpirDE analysis integrates results from both control-dependent and independent DE analyses to resolve problems caused by control-outlier genes (Fig. 3d). This is necessary because control-independent DE analysis can be unreliable when the null population is inaccurately estimated. Therefore, we combined the control-dependent and independent DE results to optimize power and error rates in EmpirDE. The approach involves applying control-independent DE analysis solely for control-outlier genes.

We developed a single parameter, the outlier threshold (*P_out*), a *P* value cutoff, to determine whether a gene should be regarded as a control-outlier gene. The control-outlier genes were identified based on two criteria: (1) within a sequencing library, this gene was unidirectionally (i.e., either all increased or decreased) differentially expressed in at least 50% of RNAi conditions in control-dependent DE analysis with a *P* value below the threshold of *P_out*; and (2) concurrently, at least 75% of RNAi conditions were coherently differentially expressed under a relaxed threshold of *P_out * 10*. By managing the 50% and 75% quantiles, this approach pinpointed genes where the overall RNAi conditions shifted up or down in gene expression compared to the vector control. By default, EmpirDE used a *P_out* of 0.005, whose determination is described in the following section *EmpirDE parameter selection*.

In each sequencing library, we applied control-independent DE results to all identified control-outlier genes. There were a few additional considerations. First, if any gene was found to be differentially expressed with substantially greater statistical significance in control-independent DE analysis – defined by *P* values at least 100 times lower and a higher fold-change – control-independent DE results were used to enhance the power of DEG discovery. Second, to maintain consistent empirical null modeling (see details below), genes marked as control-outlier genes in more than 25% of libraries (e.g., for metabolic WPS, this is 72 libraries * 0.25 = 18) had control-independent DE results applied across all libraries. This was regardless of whether they were identified as outlier genes in each individual library.

Empirical null modeling of DE test statistic

The empirical null was modeled individually for each gene by combining all conditions in a WPS experiment (dataset). In a standard WPS application, at least 96 conditions are experimented, providing a substantial sample size for building the empirical null.

We used the fitting function of the *locfdr* package in R to model the empirical null. For each gene, its Wald statistics generated by DESeq2 across all experimental conditions (>100) were used as the input for the *locfdr* function. The command used was: *locfdr(target_gene_wald_statistics, bre = brk, plot = 0, type = 0)*, where *brk = length(target_gene_wald_statistics) %/% 8*. This break size (*bre*) formula was empirically determined based on what gave the best fit in manual inspections. Extreme outliers in the Wald statistic (defined as greater than the 99% quantile plus 3 MAD (Median Absolute Deviation) or less than the 1% quantile minus 3 MAD) were excluded from the fitting, as the presence of such outliers could cause the program to fail. Occasionally, *locfdr* would exit with an error due to issues in fitting the distribution. In these cases, we incrementally increased the break size (bre = brk + 1, 2, 3,...) until a successful fit was achieved. In rare situations where fitting could not be completed after 100 increments, we used the median and MAD of the Wald statistic distribution to estimate the empirical null. Upon determining the null's parameter estimates, we rescaled the Wald statistic to compute a corrected Wald statistic and subsequently calculated the new *empirical P* values (Fig. 3d).

To compute the empirical FDR, we applied a bi-directional multiple testing correction to conservatively control the FDR. This strategy was also benchmarked through simulations (see below for details). To increase power and exclude very lowly expressed genes (akin to independent filtering in DESeq2), we first filtered the genes with median normalized counts in both vector control and RNAi samples of 30 or less (individually for each DE comparison). These filtered genes were assigned with an adjusted *P* value of *NA*. To adjust for multiple testing, a row-wise adjusted *P* value was calculated using the Benjamini-Hochberg (BH) method across all conditions for a given gene. Simultaneously, a column-wise adjusted *P* value was calculated using BH method across all pass-filter genes for a given condition. We defined the empirical FDR as the maximum of the row-wise and column-wise adjusted *P* values. This worst-case FDR approach ensures that the rate of false DE calls among all genes for a given RNAi condition, and the rate of false calls among all conditions for a given gene, are both below the desired threshold.

WPS data simulation

To mimic the real metabolic WPS dataset collected from 72 WPS sequencing libraries across 12 RNAi plates, we used scDesign3[48] to simulate each batch, i.e., each sequencing library, individually. A typical sequencing library contains 16 RNAi conditions in triplicates and 6 vector control samples. In the simulation, we first removed lowly expressed genes with a maximum read count of 10 or less. The read count matrix was then used to estimate simulation parameters via *fit_marginal* function in scDesign3. We incorporated the RNAi condition as the sole covariate in fitting mu (*mu_formula = 'condition'*) and bypassed the marginal distribution fitting of standard deviation (*sigma_formula = '1'*). The canonical negative binomial model was used throughout (*family_use = 'nb'*). Marginal distribution estimates were subsequentially input into *fit_copula (copula = 'gaussian')* to determine gene correlation parameters. These together established the simulation parameters for a sequencing library.

To simulate differential expression, we defined the ground truth for DEGs using a total of 117,782 DEGs identified in real data by the default WPS data analysis method (FDR < 0.1, FC > 1.5). This ground truth table decides which genes in which RNAi conditions should be simulated as DEGs and their desired fold changes. Next, we reconstructed the mean estimate matrix from the parameter estimation step to reflect the DEGs to be simulated. To achieve this, we first calculated the average fitted mean for each gene using the mean matrix to define its *reference expression level*. For genes designated as differentially expressed, we defined their new means in the reconstructed mean matrix as their reference expression levels multiplied by the desired fold changes from the ground truth table. For other genes, their new means were simply defined as the reference expression level, simulating no differential expression. Finally, a simulated sequencing library was generated using *simu_new* function, employing the reconstructed mean matrix and other estimated parameters as input.

In simulations incorporating $\Delta\mu$, we added random noises ($\Delta\mu$) to the reconstructed mean matrix before generating simulated data. We empirically determined the individual level of random noise for each gene based on comparisons between real data and standard-NB simulated data (the simulation data generated without adding $\Delta\mu$, as stated above). Specifically, we first identified inflated genes that required delta μ addition to align with real data (Supplementary Fig. 2c). These are genes whose log2(FC) variation across all conditions (also see below) was greater in real data than in standard-NB model simulations (Supplementary Fig. 2c, $\sigma_{real} > \sigma_{NB}$, referred to as inflated genes). A random delta μ was then added to these inflated genes using a heuristic formula that best captured gene-specific mean fluctuation (Supplementary Fig. 2c, Eqs. 1–2). Notably, this random delta μ was added to the means of inflated genes, irrespective of their differential expression status.

$$\log 2(FC_{\Delta\mu}) = \mathrm{N}\left(0,\ \left(0.8 \times \sqrt{{\sigma_{real}}^2 - {\sigma_{NB}}^2}\right)^2\right) \tag{1}$$

$$\Delta\mu = \mu_0 \times (FC_{\Delta\mu} - 1), \mu_0 is\ the\ reference\ expression\ level \tag{2}$$

Benchmarking EmpirDE using simulation data

The DE analysis on simulated data follows the same procedures as those used for real data analysis stated above. There were a few modifications because of the differences between real and simulated data. Firstly, control-independent DE analysis was omitted because control-outlier effects were not simulated, given that it is irrelevant to the analysis of empirical null modeling. Therefore, for empirical null fitting, we directly used DE results from control-dependent DE analysis. Secondly, we excluded DE analysis for two sequencing libraries (met2_lib5 and met3_lib4) because they either lacked three replicates pooled in the same library (met2_lib5) or were without vector controls (met3_lib4) (for details on these two libraries, see ref. 31). This reduced the total number of DEGs used for benchmarking analysis to 117,096. Thirdly, we implemented a simplified DE model using only the RNAi condition as covariate, as batch effects in replicates were not simulated. Other procedures remained consistent for both DESeq2 and EmpirDE analysis.

Using DE analysis outputs, we analyzed the distribution of log2(FC) in both simulated and real metabolic WPS data to derive parameters (σ_{real} and σ_{NB}) for delta μ modeling. The log2(FC) distribution for each gene was fitted following the same method as that for fitting the empirical null with the Wald statistic, to produce the corresponding σ. To increase the robustness of this parameter estimation, we performed standard NB simulations of WPS data 10 times and averaged the σ generated for each gene to determine its parameter σ_{NB}.

We benchmarked EmpirDE performance with the pool of 117,096 ground truth DEGs across the entire dataset. For each FDR threshold (ranging from 1e-30 to 1), we calculated the actual False Discovery Proportion (FDP, FP/(FP + TP)) and power (TP/(TP + FN)) based on DEGs solely defined by the FDR threshold. The metabolic WPS data were independently simulated 10 times to evaluate variability in FDP and power. To compare multiple testing adjustment strategies, we applied BH adjustments post-filtering of lowly expressed genes, as stated above. The lowly expressed genes were those with median normalized counts below 30 in both control and the RNAi condition (approximately equivalent to 5 in TPM). 'Column-wise' adjustment refers to adjusting multiple tests for genes within a single condition, using *P* values of all pass-filter genes in that condition. 'Row-wise' adjustment was for multiple tests across different conditions for a specific gene, using that gene's *P* values across 1078 conditions in this simulation. The worst-case adjustment took the higher adjusted *P* values from these two approaches.

Benchmarking EmpirDE using non-targeting perturbations (NTPs)

We identified 71 RNAi conditions, including the four spike-in controls, as non-targeting perturbations (NTP) based on the RNAi identity QC and Sanger sequencing. These RNAi clones failed RNAi identity QC, lacking substantial dsRNA detection and targeted gene knockdown, and were found to be either 'SHORT', 'VECTORLIKE', or 'RCBVECTOR' in Sanger sequencing. Thus, they should not effectively target any *C. elegans* genes, and can serve as independent negative controls (Supplementary Data 3). We used the number of DEG calls in these conditions to evaluate false discoveries. Given an FC and FDR threshold, we used the 90% quantile of the number of DEGs in these 71 conditions as an estimate of false positive calls. Choosing the 90% quantile, rather than the median, mean or max, aims to be conservative ('overestimating' false positives) while also allowing for a few outlier conditions that might not act as true negative controls.

Benchmarking EmpirDE using reproducibility

A total of 36 RNAi conditions were independently experimented 2-3 times (each with three replicates), serving to assess the reproducibility, which in turn provided a proxy of false positives. The strictly unreproducible DE calls were considered as empirical false positives. These were defined as DEGs in one experiment (FDR < 0.2, FC > 1.5 for Supplementary Fig. 3e and FDR < 0.1, FC > 1.5 elsewhere; using a slightly relaxed FDR threshold in Supplementary Fig. 3e was to include more DEG calls for better evaluating parameters) but the gene showed no significant change in expression in another (FC < 1.1) or exhibited a reverse expression change (e.g., one increased while the other decreased). This approach offers a qualitative assessment of DEG reproducibility, thus serving as a proxy for false discoveries in DE analysis. The unreproducible DEG rate was calculated as the number of strictly unreproducible DEG calls divided by the total DEG number of that experiment. The average rate of the two repeats is shown in Fig. 4f.

Of note, two of the 36 conditions had more than two (i.e., three) independent repeats. We conducted all pairwise comparisons among the three (resulting in three combinations). Therefore, a total of 40 pairs were used and displayed.

EmpirDE parameter selection

We systematically evaluated parameters in the EmpirDE framework, using NTPs and independent repeats. These included the *P_out* threshold for control-outlier gene selection, two FC estimates (log2(FC) from DESeq2 and the shrunk log2(FC) from *apeglm*), and thresholds to define DEGs (FC and FDR thresholds).

Evaluation using the 71 NTP conditions showed that false positives were insensitive to the type of FC estimates when the FDR cutoff was adequate (e.g., FDR < 0.1, data not shown). A similar observation was made with the 36 repeats. Given that FC estimates from DESeq2 yielded slightly more DEG calls (~5%) and are simpler, we opted for this FC in our standard EmpirDE analysis. Accordingly, this FC was also used when reporting the result of DESeq2 analysis (Fig. 3a).

Regarding *P_out*, we evaluated thresholds ranging from 0 to 1, and found that the values between 0.001 and 0.01 were optimal when evaluated with NTP conditions (Supplementary Fig. 3e). A similar optimal range (0.0025–0.0075) was observed with the 36 repeated conditions (Supplementary Fig. 3f). Thus, we selected 0.005 as the *P_out* parameter, which is ten-fold lower than the common statistical threshold of 0.05 but is reasonable because this threshold is on *P* values without multiple testing adjustments.

Regarding the thresholds for defining DEGs, we reasoned that an FDR of 0.1, expecting 10% false discoveries, is ideal to balance power and error, given that benchmark analyses demonstrate the FDR in EmpirDE approach is statistically rigorous (Fig. 4). We then selected a fold-change threshold (1.5) that controlled false positives in NTPs below five under this FDR threshold (Fig. 4d). This fold-change threshold can further eliminate false discoveries and uninteresting DEGs that only changed subtlety.

Together, we settled on a parameter combination of *P*_out = 0.005, empirical FDR < 0.1 and DESeq2 fold change estimate greater than 1.5 to define the final DEGs. This parameter set demonstrated stringent FDR control and established a responsiveness cutoff of 5 DEGs.

Alterations in experimental setup of NHR WPS

The NHR WPS dataset was generated during the early stages of this project, and due to historical reasons, has a different sequencing library design. Consequently, data processing was slightly altered to accommodate.

The primary distinction lies in the arrangement of the sequencing library. In the NHR experiment, samples from the same biological replicate were pooled together in a sequencing library, comprising 47 RNAi conditions and 2 vector controls. Accordingly, the three replicates of a RNAi condition were sequenced in three separate sequencing libraries, which is different from the metabolic WPS. Most NHR data were collected from an experiment using a single 96-well RNAi plate, producing two sequencing libraries that contained vector control samples derived from the same total RNA extraction. Consequently, the vector controls in the two sequencing libraries were technical replicates and should not be included simultaneously in DE analysis. The majority of NHR data were from 6 sequencing libraries (3 replicates x 2 libraries each), and there were two supplementary libraries created to incorporate extra NHR conditions and to redo the experiments for some conditions that failed quality control. This led to a total of 104 unique NHR RNAi conditions with 100 out of the 104 conditions assessed in triplicate or more (up to five biological replicates). The remaining four were in duplicate.

Lastly, the construction of NHR sequencing library followed an earlier version of WPS library construction protocol. A notable difference is the use of SuperScript™ II instead of SuperScript™ III that is used in standard WPS (Supplementary Protocol). However, we did not notice obvious differences in the detection sensitivity of genes driven by this change of reverse transcriptase (data not shown).

NHR WPS QC

Quality control (QC) for NHR WPS dataset was conducted in a manner similar to the metabolic sequencing library. However, it is important to note that the NHR sequencing library was collected in the initial phase of the project, a period when many methodology optimizations were still ongoing, the pass-QC rate for the NHR sequencing library is somewhat lower than that observed for the metabolic sequencing library. Out of 392 samples, a total of 353 (~90%) passed quality control. The fail-QC samples include 19 bad quality samples (5%), a frequency notably higher than metabolic sequencing library (2%), and 20 failed RNAi identity QC due to a pipetting error that caused cross-contamination.

NHR WPS DE analysis

The DE analysis for the NHR sequencing library posed unique challenges due to its distinct library arrangement, necessitating a different DE strategy rather than the within-library analysis. We divided all samples (across 8 sequencing libraries) into four DE groups to maximize sample pooling, which aids in better estimating dispersion, and to prevent the co-occurrence of technical replicates of vector controls. The four DE groups were defined as follows: Group 1 included most conditions from the first half of the 96-well plate, along with additional conditions from the supplementary library #1. Group 2 comprised the remaining conditions from the first half of the 96-well plate, plus their extra replicates sequenced in supplementary library #2. Group 3 contained most conditions from the second half of the 96-well plate, plus their extra replicates from supplementary library #1. Group 4 involved the remaining conditions from the second half of the 96-well plate, plus their extra replicates from supplementary library #2. This grouping strategy was complex and a compromise, reflecting the less refined experimental design at the early stages of the project. This issue was unique to the NHR dataset.

The standard EmpirDE analysis was applied to the NHR dataset, with control-dependent and -independent DE performed using samples from the four DE groups separately. The parameter choices remained consistent with those used in the rest of the project. Additional modifications specific to the NHR dataset are as follows:

1. We noted that the NHR sequencing library exhibited a significantly lower incidence of control-outlier genes compared to the metabolic WPS (~10-fold less, data not shown). We suspect a relevance to the preparation of bacterial diet: for NHR WPS, control bacteria were cultured in the same 96-well plate with the RNAi bacteria. However, for metabolic WPS, control bacteria were cultured in a second 96-well plate because of the inclusion of more RNAi conditions in each experiment (usually ~120 conditions, Supplementary Protocol). Based on this observation, we proposed an optimized experimental design in the Supplementary Protocol for future users.

2. A comparison of empirical nulls between the metabolic and NHR WPS revealed distinct, albeit moderately correlated, standard deviations (data not shown), suggesting each experiment exhibits unique noises to address. Therefore, the gene-specific random fluctuations can be influenced by experimental batches, and the empirical null needs to be reconstructed in each individual WPS study for its best outcome.

Transcriptional profiling of animals in different developmental stages

We profiled the transcriptome of animals at different developmental stages, ranging from L2 larvae to adult, using the WPS sequencing library construction method. This includes 51 samples from animals at 17, 25, 35, 40, 45, 50, 55, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69 h post L1 plating in biological triplicate. Animals were fed on HT115 bacteria expressing empty vector control (L4440), i.e., the dsRNA expression was induced as in regular WPS experiments. This is to be aligned with the dietary condition of WPS study. The raw data were processed following standard WPS procedures to obtain the read count matrix.

For methodological development purposes, we constructed two technical replicates of this sequencing library using different reverse transcriptases (SuperScript™ II and III). The library generated by SuperScript™ III was used as the non-RNAi control in the RNAi identity analysis of standard WPS (for RNAi identity analysis, see above sections). Analysis of the two sequencing libraries did not find notable differences in detecting gene expression other than batch effects. Therefore, for downstream analysis, we aggregated data from the two replicates by adding up read counts. In five samples (55h_rep1, 68h_rep1, 67h_rep1, 63h_rep3, and 55h_rep2), we noted a slight decorrelation with their biological replicates, a deviation from the correlation levels observed in other samples. This decorrelation can be related to variations in development rate or sample quality in these specific samples. To ensure the integrity of data interpretation, these five samples were excluded from downstream analysis. Sample-to-sample Pearson correlation coefficients were calculated using variance stabilized read counts obtained through the *vst* function in DESeq2 package[38]. These coefficients were then visualized using the *pheatmap* from pheatmap package in R.

The raw and processed sequencing data are available at Gene Expression Omnibus (GEO) session GSE255865. Sequencing libraries constructed with SuperScript™ II and III were made available separately for reference.

Subsampling analysis for optimal sequencing depth

To generate a reference profile for subsampling analysis, we aggregated WPS data (i.e., adding up read counts) from one replicate of the developmental stage samples collected at 62, 63, 64 and 65 h. This produced a transcriptome profile with a sequencing depth of ~50 million reads. We used R package *subSeq*[82] to subsample this reference data 10 times at 9 different depths (39 M, 30 M, 20 M,10 M, 8 M, 6 M, 4 M, 2 M and 1 M reads). The coefficient of variation (CV) for each gene at each depth was then calculated. Genes with a CV < 0.15 at a specific depth were classified as quantified at that depth. This CV cutoff represents that the majority of sampling population (~95%, two standard deviations) is within a ± 30% interval of the true value, identifying the genes that can be accurately quantified. At each sampling depth, we calculated the fraction of genes quantified at varying TPM threshold (greater than 1, 2, 4, 8 or 16) to determine the sensitivity of gene detection. This analysis was repeated individually using each of the three biological replicates in the developmental stage experiments, whose standard deviations defined the error bars in Fig. 2b. A similar analysis with error bars defined by the three replicates was included in Supplementary Fig. 1e. The combined data of the first replicate was used as reference profile for analyses that need a reference TPM profile of the wild-type animal (Fig. 5a and Supplementary Fig. 1c, g).

We opted for sampling final read counts to better compare effects of different final depths rather than raw read depth, however, a similar result was also observed in the subsampling of fastq files (data not shown).

Regular RNA-seq profile for benchmarking gene detection

A transcriptional profile of adult *C. elegans* (64 hours post L1 seeding) was generated using regular RNA-seq approach to benchmark gene detection sensitivity of WPS (Supplementary Fig. 1e). The total RNA sample from one replicate of the developmental stage time-course experiment was sent to BGI for Transcriptome Resequencing Pure-sequencing service, which is a commonly used commercial RNA-seq service similar to Illumina TruSeq RNA. Data were processed using standard RSEM processing pipeline (v1.7.0) in Via Foundry platform[77], which aligns the reads to genome by Bowtie and estimate gene-level counts using RSEM. The resulting gene-level TPM profile was used for the benchmark analysis.

Gene set enrichment analysis (GSEA) of noisy genes

Gene Set Enrichment Analysis (GSEA) was performed using the original Java program[83] executed via command line. To construct a ranking metric, we used the standard deviations of empirical null for each gene. Specially, we subtracted one from these standard deviations, thereby calculating the difference between empirical and theoretical nulls. These adjusted values were used as ranking metric for the analysis. The "GSEAPreranked" method was used with parameters "*-scoring_scheme weighted -nperm 10000*". The analysis was performed on WormCat annotations Category 3.

Functional enrichment analysis of NHR GRN

We performed the functional enrichment analysis for NHR perturbations with more than 10 DEGs. DEGs of the RNAi targeted gene were excluded from the analysis. Functional enrichment analysis was performed by *enricher* function from *clusterProfiler* package[84] in R. The universe (*universe* parameter) of the enrichment analysis was defined by all genes (16,245) analyzed in the differential expression analysis. WormCat v2 (Nov. 11, 2021)[85] was used for the gene sets.

NHR perturbation-perturbation similarity

Perturbation-perturbation similarity was calculated following the same methodology used in metabolic WPS analysis[31]. In brief, we used filtered log2(FC), derived by masking log2(FC) of RNAi target genes and non-DEGs to zero, to compute cosine similarity values. These cosine values were used to quantify the perturbation-perturbation similarity. Only responsive perturbations were included in this analysis.

To define the *nhr* pairs based on cosine similarity, we constructed a hierarchical tree by *hclust* function in R using 1-cosine similarity as distance input and 'complete' linkage method. The tree was then cut at its first merge of the leaves, yielding either singleton leaves or clustered pairs (Fig. 6c). Using these pairs as clusters, we calculated the silhouette score of each data point by *silhouette* function from *cluster* package. The average silhouette score for all pairs formed a metric measuring the fitness of assigning *nhr* into pairs. To determine the statistical significance, we randomized the NHR GRN 10,000 times using edge swapping approach (50–100x coverage each randomization, i.e., the edges were swapped 50–100 × *the total number of edges* times each randomization)[86], producing a distribution of average silhouette score with random networks. An empirical *P* value was calculated based on this distribution.

Comparing NHR perturbation-perturbation similarity with protein sequence similarity

We analyzed the protein sequence of the longest transcript for 52 *nhr* genes sharing a significant overlap of DEGs with another NHR. These sequences were used as input for Clustal Omega online tool provided by EMBL-EBI[87]. Default parameters were used except for choosing 'yes' for 'DISTANCE MATRIX' and 'No' for 'mBed-like Clustering Guide-tree'. The output distance matrix and percent identity matrix were downloaded and used for the downstream analysis. We also analyzed the DNA biding domain (DBD) sequence similarity of the same 52 *nhr* genes by the same method. The DBD of NHRs were predicted by Conserved Domain Search Service (CD Search) from NCBI with the default settings.

The distance matrix was used to build a hierarchical tree with which the percent identity matrix was visualized in Fig. 6d. The percent identity matrix was also used to compare with cosine similarity directly (Fig. 6e, f).

Compare NHR WPS perturbation-perturbation similarity with gene coexpression

To comprehensively characterize *nhr* gene coexpression, we used a *C. elegans* gene expression compendium that comprises 4796 samples across 177 datasets[68]. Given that genes can be coexpressed in one dataset but not in another, we first calculated the Pearson Correlation Coefficient (PCC) for gene-gene correlation within each individual dataset. The median PCC in these 177 datasets was then used to quantify the overall strength of coexpression between pairs of *nhr* genes. We compared this median PCC with the corresponding perturbation-perturbation cosine similarity. To quantify the concordance of these two quantities, we focused on the NHR pairs with high cosine similarity (>0.2) and calculated the median level of their overall strength of coexpression. This observed value was compared to its distribution by random, generated by shuffling the name label of NHR genes in cosine similarity matrix for 10,000 times, to produce an empirical *P* value.

Validation of potential AND-logic gate between NHR pairs

The pairwise modularity observed in NHR perturbations suggests the potential presence of AND-logic gates within the NHR gene regulatory network. We reasoned that such motifs could be observed by by comparing the effect of double NHR knockdown on the transcriptome to single RNAi conditions. We performed new WPS experiments for two NHR pairs: *nhr-10/68*, a known AND-logic gate pair[67], and *nhr-68/101*, a newly identified pair. For each pair, we performed single-gene RNAi for each NHR and double-gene RNAi targeting both genes simultaneously.

The double-gene RNAi constructs were generated by inserting ~400 bp cDNA fragments of each target gene into a single L4440-Dest-RNAi vector, which was linearized using *HindIII* and *BglII* restriction enzymes[88]. The cloning was carried out using the Gibson assembly method. WPS experiments were performed as described above.

Statistics and reproducibility

No statistical methods were used to predetermine sample size. All data were analyzed without exclusion, except for low-quality RNA-seq samples (see Fig. 2c, Supplementary Fig. 1f and corresponding section in Methods for details). The experiments were not randomized, and investigators were not blinded to group allocation during experiments or outcome assessment. For most RNA-seq experiments, three biological replicates were used. Differential expression analysis methods are detailed in the corresponding Methods section. Statistical methods for other experiments are specified in the relevant figure panels or legends.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Source codes and data for reproducing all results related to core computational analyses (simulation, benchmarking and DE analysis) have been deposited in Zenodo for full reproducibility [https://doi.org/10.5281/zenodo.15223779]. Other source data are provided with this paper. Raw and processed data in this study are available in Gene Expression Omnibus (GEO) under accession code GSE255865. Downloadable read count data together with detailed documentation of experimental metadata are also available at the WPS portal hosted in our WormFlux website [https://wormflux.umassmed.edu/WPS]. Source data are provided with this paper.

Code availability

The WPS data analysis pipeline is available at our GitHub repository [https://github.com/XuhangLi/WPS] under MIT license, including detailed procedures of raw data processing, quality control and EmpirDE analysis. A walkthrough of the pipeline can be found in the repository. A standalone R package for EmpirDE can be found at GitHub repository [https://github.com/XuhangLi/EmpirDE] under MIT license.

References

1. Hughes, T. R. et al. Functional discovery via a compendium of expression profiles. *Cell* **102**, 109–126 (2000).
2. Giaever, G. et al. Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418**, 387–391 (2002).
3. Przybyla, L. & Gilbert, L. A. A new era in functional genomics screens. *Nat. Rev. Genet.* **23**, 89–103 (2022).
4. Ideker, T. et al. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* **292**, 929–934 (2001).
5. Kemmeren, P. et al. Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors. *Cell* **157**, 740–752 (2014).
6. Baryshnikova, A. et al. Quantitative analysis of fitness and genetic interactions in yeast on a genome scale. *Nat. Methods* **7**, 1017–1024 (2010).
7. Adamson, B. et al. A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. *Cell* **167**, 1867–1882.e1821 (2016).
8. Jaitin, D. A. et al. Dissecting immune circuits by linking CRISPR-Pooled screens with single-. *Cell RNA-Seq. Cell* **167**, 1883–1896.e1815 (2016).
9. Dixit, A. et al. Perturb-Seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell* **167**, 1853–1866.e1817 (2016).
10. Datlinger, P. et al. Pooled CRISPR screening with single-cell transcriptome readout. *Nat. Methods* **14**, 297–301 (2017).
11. Gasperini, M. et al. A genome-wide framework for mapping gene regulation via. *Cell. Genet. Screens Cell* **176**, 377–390.e319 (2019).
12. Schraivogel, D. et al. Targeted Perturb-seq enables genome-scale genetic screens in single cells. *Nat. Methods* **17**, 629–635 (2020).
13. Replogle, J. M. et al. Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. *Cell* **185**, 2559–2575.e2528 (2022).
14. Jin, X. et al. In vivo Perturb-Seq reveals neuronal and glial abnormalities associated with autism risk genes. *Science* **370**, eaaz6063 (2020).
15. Zheng, X. et al. Massively parallel in vivo Perturb-seq reveals cell type-specific transcriptional networks in cortical development. *bioRxiv*. https://doi.org/10.1101/2023.09.18.558077 (2023).
16. Santinha, A. J. et al. Transcriptional linkage analysis with in vivo AAV-Perturb-seq. *Nature* **622**, 367–375 (2023).
17. Fire, A. et al. Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* **391**, 806–811 (1998).
18. Timmons, L. & Fire, A. Specific interference by ingested dsRNA. *Nature* **395**, 854 (1998).
19. Kamath, R. S. et al. Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature* **421**, 231–237 (2003).
20. Rual, J.-F. et al. Toward improving *Caenorhabditis elegans* phenome mapping with an ORFeome-based RNAi library. *Genome Res.* **14**, 2162–2168 (2004).
21. MacNeil, L. T. et al. Transcription factor activity mapping of a tissue-specific gene regulatory network. *Cell Syst.* **1**, 152–162 (2015).
22. Horowitz, B. B., Nanda, S. & Walhout, A. J. M. A transcriptional cofactor regulatory network for the *C. elegans* intestine. *G3***13**, jkad096 (2023).
23. Bhattacharya, S. et al. A metabolic regulatory network for the *Caenorhabditis elegans* intestine. *iScience* **25**, 104688 (2022).
24. Yanai, I. & Hashimshony, T. CEL-Seq2-single-cell RNA sequencing by multiplexed linear amplification. *Methods Mol. Biol.* **1979**, 45–56 (2019).
25. Hashimshony, T. et al. CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.* **17**, 77 (2016).
26. Spencer, W. C. et al. A spatial and temporal map of *C. elegans* gene expression. *Genome Res.* **21**, 325–341 (2011).

27. Hendriks, G. J., Gaidatzis, D., Aeschimann, F. & Grosshans, H. Extensive oscillatory gene expression during *C. elegans* larval development. *Mol. Cell* **53**, 380–392 (2014).
28. Meeuse, M. W. et al. Developmental function and state transitions of a gene expression oscillator in *Caenorhabditis elegans*. *Mol. Syst. Biol.* **16**, e9498 (2020).
29. Heimberg, G., Bhatnagar, R., El-Samad, H. & Thomson, M. Low Dimensionality in gene expression data enables the accurate extraction of transcriptional programs from shallow sequencing. *Cell Syst.* **2**, 239–250 (2016).
30. Liu, Y., Zhou, J. & White, K. P. RNA-seq differential expression studies: more sequence or more replication?. *Bioinformatics* **30**, 301–304 (2014).
31. Li, X. et al. Systems-level design principles of metabolic rewiring in an animal. *Nature* https://doi.org/10.1038/s41586-025-08636-5 (2025).
32. Zhang, H. et al. A systems-level, semi-quantitative landscape of metabolic flux in *C. elegans*. *Nature* **640**, 194–202. (2025).
33. Taub, M. A., Corrada Bravo, H. & Irizarry, R. A. Overcoming bias and systematic errors in next generation sequencing data. *Genome Med.* **2**, 87 (2010).
34. Tung, P. Y. et al. Batch effects and the effective design of single-cell gene expression studies. *Sci. Rep.* **7**, 39921 (2017).
35. Love, M. I., Anders, S., Kim, V. & Huber, W. RNA-Seq workflow: gene-level exploratory analysis and differential expression. *F1000Res* **4**, 1070 (2015).
36. Conesa, A. et al. A survey of best practices for RNA-seq data analysis. *Genome Biol.* **17**, 13 (2016).
37. Sijen, T. et al. On the role of RNA amplification in dsRNA-triggered gene silencing. *Cell* **107**, 465–476 (2001).
38. 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).
39. Consortium, S. M.-I. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat. Biotechnol.* **32**, 903–914 (2014).
40. Li, S. et al. Detecting and correcting systematic variation in large-scale RNA sequencing data. *Nat. Biotechnol.* **32**, 888–895 (2014).
41. 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).
42. Pall, T., Luidalepp, H., Tenson, T. & Maivali, U. A field-wide assessment of differential expression profiling by high-throughput sequencing reveals widespread bias. *PLoS Biol.* **21**, e3002007 (2023).
43. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
44. Hart, S. N., Therneau, T. M., Zhang, Y., Poland, G. A. & Kocher, J. P. Calculating sample size estimates for RNA sequencing data. *J. Comput, Biol.* **20**, 970–978 (2013).
45. Poplawski, A. & Binder, H. Feasibility of sample size calculation for RNA-seq studies. *Brief. Bioinform.* **19**, 713–720 (2018).
46. Schurch, N. J. et al. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?. *RNA* **22**, 839–851 (2016).
47. Efron, B. Large-scale simultaneous hypothesis testing - the choice of a null hypothesis. *J. Am. Stat. Assoc.* **99**, 96–104 (2004).
48. Song, D. et al. scDesign3 generates realistic in silico data for multimodal single-cell and spatial omics. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-023-01772-1 (2023).
49. Chen, C., Samuel, T. K., Sinclair, J., Dailey, H. A. & Hamza, I. An intercellular heme-trafficking protein delivers maternal heme to the embryo during development in C. elegans. *Cell* **145**, 720–731 (2011).
50. Sinclair, J. et al. Inter-organ signalling by HRG-7 promotes systemic haem homeostasis. *Nat. Cell Biol.* **19**, 799–807 (2017).
51. Subramaniam, N., Treuter, E. & Okret, S. Receptor interacting protein RIP140 inhibits both positive and negative gene regulation by glucocorticoids. *J. Biol. Chem.* **274**, 18121–18127 (1999).
52. Tao, L. J., Seo, D. E., Jackson, B., Ivanova, N. B. & Santori, F. R. Nuclear hormone receptors and their ligands: metabolites in control of transcription. *Cells* **9**, 2606 (2020).
53. Reece-Hoyes, J. S. et al. A compendium of *C. elegans* regulatory transcription factors: a resource for mapping transcription regulatory networks. *Genome Biol.* **6**, R110 (2005).
54. Antebi, A. in *WormBook, The C.elegans Research Community, ed. WormBook*. http://www.wormbook.org (Nov 21, 2005). (ed I. Greenwald) (2005).
55. Scholtes, C. & Giguere, V. Transcriptional control of energy metabolism by nuclear receptors. *Nat. Rev. Mol. Cell Biol.* **23**, 750–770 (2022).
56. Arda, H. E. et al. Functional modularity of nuclear hormone receptors in a *C. elegans* gene regulatory network. *Mol. Syst. Biol.* **6**, 367 (2010).
57. Reece-Hoyes, J. S. et al. Extensive rewiring and complex evolutionary dynamics in a *C. elegans* multiparameter transcription factor network. *Mol. Cell* **51**, 116–127 (2013).
58. Fuxman Bass, J. I. et al. A gene-centered *C. elegans* protein-DNA interaction network provides a framework for functional predictions. *Mol. Syst. Biol.* **12**, 884 (2016).
59. Van Gilst, M. R., Hajivassiliou, H., Jolly, A. & Yamamoto, K. R. Nuclear hormone receptor NHR-49 controls fat consumption and fatty acid composition in *C. elegans*. *PLoS Biol.* **3**, e53 (2005).
60. Asahina, M., Valenta, T., Silhankova, M., Korinek, V. & Jindra, M. Crosstalk between a nuclear receptor and beta-catenin signaling decides cell fates in the *C. elegans* somatic gonad. *Dev. Cell* **11**, 203–211 (2006).
61. Antebi, A. Nuclear receptor signal transduction in C. elegans. *WormBook*, 1–49 https://doi.org/10.1895/wormbook.1.64.2 (2015).
62. Conte, D. Jr, MacNeil, L. T., Walhout, A. J. & Mello, C. C. RNA Interference in *Caenorhabditis elegans*. *Curr. Protoc. Mol. Biol.* **109**, 26 23 21–26 23 30 (2015).
63. Albert, R. Scale-free networks in cell biology. *J. Cell Sci.* **118**, 4947–4957 (2005).
64. Deplancke, B. et al. A gene-centered *C. elegans* protein-DNA interaction network. *Cell* **125**, 1193–1205 (2006).
65. Holdorf, A. D. et al. WormCat: an online tool for annotation and visualization of *Caenorhabditis elegans* Genome-Scale Data. *Genetics* **214**, 279–294 (2020).
66. Hahn-Windgassen, A. & Van Gilst, M. R. The *Caenorhabditis elegans* HNF4alpha Homolog, NHR-31, mediates excretory tube growth and function through coordinate regulation of the vacuolar ATPase. *PLoS Genet.***5**, e1000553 (2009).
67. Bulcha, J. T. et al. A persistence detector for metabolic network rewiring in an animal. *Cell Rep.* **26**, 460–468 (2019).
68. Nanda, S. et al. Systems-level transcriptional regulation of *Caenorhabditis elegans* metabolism. *Mol. Syst. Biol.* **19**, e11443 (2023).
69. Ward, J. D. et al. Defects in the C. elegans acyl-CoA synthase, acs-3, and nuclear hormone receptor, nhr-25, cause sensitivity to distinct, but overlapping stresses. *PLoS ONE***9**, e92552 (2014).
70. Robinson-Rechavi, M., Maina, C. V., Gissendanner, C. R., Laudet, V. & Sluder, A. Explosive lineage-specific expansion of the orphan nuclear receptor HNF4 in nematodes. *J. Mol. Evol.* **60**, 577–586 (2005).
71. van Iterson, M., van Zwet, E. W., Consortium, B. & Heijmans, B. T. Controlling bias and inflation in epigenome- and transcriptome-wide association studies using the empirical null distribution. *Genome Biol.* **18**, 19 (2017).

72. Cao, X., Wu, B. & Hertz, M. I. Empirical null distribution based modeling of multi-class differential gene expression detection. *J. Appl Stat.* **40**, 347–357 (2013).
73. Schmitt, A. D., Hu, M. & Ren, B. Genome-wide mapping and analysis of chromosome architecture. *Nat. Rev. Mol. Cell Biol.* **17**, 743–755 (2016).
74. Halasz, G. et al. Optimizing murine sample sizes for RNA-seq studies revealed from large-scale comparative analysis. *BioRxiv* (2024).
75. Zhang, H. et al. Worm Perturb-Seq: massively parallel whole-animal RNAi and RNA-seq. *Nat. Commun.* (2025).
76. Brenner, S. The genetics of *Caenorhabditis elegans*. *Genetics* **77**, 71–94 (1974).
77. Yukselen, O., Turkyilmaz, O., Ozturk, A. R., Garber, M. & Kucukural, A. DolphinNext: a distributed data processing platform for high throughput genomics. *BMC Genom.***21**, 310 (2020).
78. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
79. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
80. Derr, A. et al. End Sequence Analysis Toolkit (ESAT) expands the extractable information from single-cell RNA-seq data. *Genome Res.* **26**, 1397–1410 (2016).
81. Wang, M., Jiang, L. & Snyder, M. P. AdaTiSS: a novel data-Adaptive robust method for identifying Tissue Specificity Scores. *Bioinformatics* **37**, 4469–4476 (2021).
82. Robinson, D. G. & Storey, J. D. subSeq: determining appropriate sequencing depth through efficient read subsampling. *Bioinformatics* **30**, 3424–3426 (2014).
83. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102**, 15545–15550 (2005).
84. Wu, T. et al. clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *Innovation***2**, 100141 (2021).
85. Higgins, D. P., Weisman, C. M., Lui, D. S., D'Agostino, F. A. & Walker, A. K. Defining characteristics and conservation of poorly annotated genes in Caenorhabditis elegans using WormCat 2.0. *Genetics* **221**, iyac085 (2022).
86. Milo, R. et al. Network motifs: simple building blocks of complex networks. *Science* **298**, 824–827 (2002).
87. Madeira, F. et al. Search and sequence analysis tools services from EMBL-EBI in 2022. *Nucleic Acids Res.* **50**, W276–W279 (2022).
88. Min, K., Kang, J. & Lee, J. A modified feeding RNAi method for simultaneous knock-down of more than one gene in *Caenorhabditis elegans*. *Biotechniques* **48**, 229–232 (2010).
89. Robinson, J. T. et al. Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
90. Roux, A. E. et al. Individual cell types in *C. elegans age* differently and activate distinct cell-protective responses. *Cell Rep.* **42**, 112902 (2023).

Acknowledgements

We thank members of the Walhout lab, Mike Lee, and Chad Myers for discussion and critical reading of the manuscript. We thank former Garber lab members, Kyle Gellatly and Rachel Murphy, for their help in the early stage of this project. This work was supported by grants from the National Institutes of Health GM122502 and DK068429 to A.J.M.W., U01HG012064 to M.G., and NSF DBI-1846216 and NIGMS R35GM140888 to J.J.L.

Author contributions

H.Z., X.L., and A.J.M.W. conceived the project and wrote the manuscript. H.Z. and X.L. jointly developed the WPS technology and analyzed the data. H.Z. conducted the experiments. X.L. wrote the codes. D.S. helped design the simulation study under the supervision of J.J.L. D.S. also provided critical discussions that led to the idea of modeling the empirical null. O.Y. and A.K. helped with the dolphinNext pipeline for WPS data processing. S.N. provided the NHR coexpression data. M.G. and A.J.M.W. supervised the study. The co-first authorship order was determined by a coin flip. X.L. and H.Z. contributed equally and reserve the right to list their name first in their resumes.

Competing interests

M.G. and A.K. are co-founders of Via Scientific, Inc., a UMass Chan Medical School spin-off. A.K. is a board member of the company. M.G., A.K., and O.Y. have equity in the company. They ensure that steps have been taken to prevent these affiliations from affecting analysis integrity and are dedicated to upholding research transparency and integrity. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-60154-0.

Correspondence and requests for materials should be addressed to Manuel Garber or Albertha J. M. Walhout.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints

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