Published online 2 June 2021 *Nucleic Acids Research, 2021, Vol. 49, No. 14* ***e83***
https://doi.org/10.1093/nar/gkab433

noisyR: enhancing biological signal in sequencing datasets by characterizing random technical noise

Ilias Moutsopoulos[1], Lukas Maischak[2], Elze Lauzikaite[1], Sergio A. Vasquez Urbina[2], Eleanor C. Williams[1], Hajk-Georg Drost[2] and Irina I. Mohorianu[1,*]

[1]Wellcome-MRC Cambridge Stem Cell Institute, University of Cambridge, Cambridge CB2 0AW, UK and [2]Computational Biology Group, Department of Molecular Biology, Max Planck Institute for Developmental Biology, Max-Planck Ring 1, 72076 Tübingen, Germany

Received February 18, 2021; Revised April 16, 2021; Editorial Decision May 02, 2021; Accepted May 06, 2021

ABSTRACT

High-throughput sequencing enables an unprecedented resolution in transcript quantification, at the cost of magnifying the impact of technical noise. The consistent reduction of random background noise to capture functionally meaningful biological signals is still challenging. Intrinsic sequencing variability introducing low-level expression variations can obscure patterns in downstream analyses. We introduce *noisyR*, a comprehensive noise filter to assess the variation in signal distribution and achieve an optimal information-consistency across replicates and samples; this selection also facilitates meaningful pattern recognition outside the background-noise range. *noisyR* is applicable to count matrices and sequencing data; it outputs sample-specific signal/noise thresholds and filtered expression matrices. We exemplify the effects of minimizing technical noise on several datasets, across various sequencing assays: coding, non-coding RNAs and interactions, at bulk and single-cell level. An immediate consequence of filtering out noise is the convergence of predictions (differential-expression calls, enrichment analyses and inference of gene regulatory networks) across different approaches.

INTRODUCTION

High-throughput sequencing (HTS) became a new standard in most life science studies yielding unprecedented insights into the complexity of biological processes. The increase in sequencing depth and number of samples, across both bulk and single cell experiments, facilitated a greater diversity in biological questions (1), at the same time allowing a higher sensitivity for the detection of perturbations in gene expression levels between samples (2). This increased accuracy greatly assists with the biological interpretation of results such as identification and characterization of differential expression (DE) at tissue and cellular levels (3) or the inference and characterization of gene regulatory networks (4). However, HTS may exhibit high background noise levels resulting from non-biological/technical variation, introduced at different stages of the RNA-seq library preparation, or from amplification/sequencing bias (5) to random hexamer priming during the sequencing reaction (6). These technical alterations of signal can affect the accuracy of the downstream DE results or create spurious patterns biasing downstream interpretations. Statistical methods developed to date (7,8), focused mainly on batch/background correction, normalization, and evaluation of DE, have been designed to mitigate the impact of these biases on DE analyses (9). A noise filter for pre-processing of the data before these steps would ensure a reduction of further amplification of these biases. Here, we introduce a new high-throughput noise filter to remove random technical noise from sequencing data and illustrate the downstream information consistency that is achieved.

While technologies may exhibit different technical biases, the sequencing bias across an experiment was expected to be uniform. This expectation was based on the assumption that sequencing reads would uniformly cover the expressed transcripts, with the algebraic sum of reads from each gene being proportional to the expression of that gene (10). However, in practice we observe a reproducible, yet uneven distribution of signal across transcripts (10); moreover, highly abundant genes show a higher consistency of transcript-coverage than lower abundance genes. This coverage bias of lower abundance genes is one of the main origins of technical noise (11). The latter can be attributed to the stochasticity of the sequencing process, the limits of sequencing depth, and alignment inaccuracies during the mapping procedure. To further explore the coverage bias of lower abundance genes, we define genes whose quantification is characterized by such a lack of coverage-uniformity as “noisy”.

The presence of noise in HTS data has been widely acknowledged, and there have been several attempts to un-

[*]To whom correspondence should be addressed. Tel: +44 1223 767800; Email: iim22@cam.ac.uk

derstand and quantify it. A recent study (12) presented a variety of common experimental errors that may increase sequencing noise and proposed ways to alleviate their effect such as using a mild acoustic shearing condition to minimise the occurrence of DNA damage. Fischer-Hwang and colleagues (13) presented a denoizing tool that can be applied on aligned genomic data with high fold-coverage of the genome to improve variant calling performance. The recent prevalence of single-cell sequencing technologies has further highlighted the issue of noise, as the lower sequencing depth per cell leads to more uncertainty of the quantification of (low abundance) genes. Efforts have been made to reduce the noise levels experimentally, such as by utilizing a different barcoding approach (14).

On the computational side, several imputation and denoizing algorithms have been proposed, e.g. a machine learning (ML) based deep count autoencoder (15). Other tools focus on DE analysis, such as TASC (16), which uses a hierarchical mixture model of the biological variation. However, successful methods usually rely on assumptions specific to the biological experiment and are tailored to particular settings or model systems, thus leaving most large-scale sequencing efforts, lacking such specific experimental design, exposed to random technical noise. To our knowledge, there is little focus on bulk experiments, where technical noise still exists at low abundances, independent of biological assumptions; for these experiments the low number of replicates hinders imputation-based approaches.

Existing approaches for calling DE genes mitigate to various extents the presence of noise, however these are not designed to identify and assess the impact of genes showing random, low-level variation. As a result, some of these are detected by the DE analyses, biasing the biological interpretation of the results. In addition, the choice of tools used for pre-processing steps may influence the relative transcript expression estimation accuracy (17). These analytical biases mainly arise from differences in the detection and handling of transcript isoforms or processing of unmapped and multi-mapping reads (3). Such variation in abundance estimation can in turn strongly affect the downstream analyses (18).

We developed ***noisyR***, a denoizing pipeline to quantify and exclude technical noise from downstream analyses, in a robust and data-driven way. The approach underlines consistency of signal over a user-defined threshold. ***noisyR*** is applicable on either the original, un-normalized count matrix, or alignment data (BAM format). Noise is quantified based on the correlation of expression across subsets of genes for the former, or distribution of signal across the transcripts for the latter, in different samples/replicates and across all gene abundances (Methods). We illustrate the approach on bulk and single cell RNA-seq datasets and highlight the impact of the noise removal on refining the biological interpretation of results.

MATERIALS AND METHODS

Materials

The bulk mRNA-seq used to illustrate ***noisyR*** was generated by Yang *et al.* (19). The dataset comprises 16 samples across 8 time points [0–72 h post stem cell induction]. The raw data (fastq files and metadata) were downloaded from GEO (accession numbers GSE117896, GSM3314677–GSM3314692).

Next, sRNA data was retrieved from Paicu *et al.* (20) for the plant dataset (2 samples, a wildtype and DCL1 knockdown, with three biological replicates each, in *Arabidopsis thaliana*, GSM2412286–GSM2412291) and from Wallach *et al.* (21) for the animal dataset, 6 samples generated for the identification of microRNAs as TLR-activating molecules in *Mus musculus* (PMID: 31940779, GSE138532, GSM4110737 - GSM4110742). For both datasets, the reads were aligned to mature and hairpin miRNAs, downloaded from miRBase (22) and TEs, downloaded from TAIR and Ensembl, for *M. musculus*.

For assessing the impact of noise on direct biological interpretations and predictions, such as the interaction of miRNAs and mRNAs, we selected a PARE (parallel analysis of RNA ends, also known as degradome sequencing) dataset, consisting of three biological replicates (GSE113958) presented in Thody *et al.* (23).

The single-cell mRNA-seq dataset used to illustrate ***noisyR*** was generated by Cuomo et al (study of stem cell differentiation) (24). The data is available on ENA, ERP016000–PRJEB14362. The six donors with the highest number of cells (hayt, naah, vils, pahc, melw, qunz) were selected, cells in time point 3 were included.

The reference genomes used for alignment were: Homo_sapiens.GRCh38.98 (Ensembl version 98), Mus_musculus.GRCm38.98 (Ensembl version 98) and *A. thaliana* (25).

Methods, bulk mRNAseq data

Data pre-processing and quality checking. Initial quality checks were performed using fastQC (version 0.11.8) and summarized with multiQC (version 1.9) (26). Alignments to reference genomes were performed using STAR (version 2.7.0a) with default parameters (27); the count matrices were generated using featureCounts (version 2.0.0) (28) against the *M. musculus* exon annotations obtained from the Ensembl database (genome assembly GRCm38.p6). Additional quality checks included density plots, (comparable distributions are a necessary but not sufficient condition for comparability), MA plots for the sufficiency check (expected to have a funnelling shape; observed outliers are candidates for differentially expressed transcripts), incremental dendrograms and PCA plots to evaluate the similarity of distributions (11,29).

Data post-processing and biological interpretation of results. The differential expression analysis was performed after quantile normalization of the count matrix using the standard functions from edgeR, version 3.28.0 (30) and DESeq2, version 1.26.0 (7). The thresholds for DE were $|\log_2(FC)| > 1$ and adjusted P-value < 0.05 (Benjamini–Hochberg multiple testing correction). The enrichment analysis was performed using g:profiler (R package gprofiler2, version 0.2.0) (31), against the standard GO terms, and the KEGG (32) and reactome (33) pathway databases. The observed set consisted of the DE genes, the background set comprised all expressed genes, using the full or denoised count matrix respectively.

To assess the effect of noise correction across the multiple options of mRNA quantification, the sequencing reads were aligned to the reference genome using Bowtie2 (version 2.4.2) (34) and HISAT2 (version 2.1.0) (35). Aligners were run both with default parameters and with parameters set to match the STAR functionality of searching for up to 10 distinct, valid alignments for each read ("bowtie2 -end-to-end -k 10" and "hisat2 -q -k 10"). The transcript expression was quantified using featureCounts. The robustness of the quantification was assessed by investigating the overlap between edgeR and DESeq2 analyses. The genes with adjusted P-value <0.05 (Benjamini–Hochberg multiple testing correction) and $|\log_2(\text{FC})| > 1$ were considered before and after noise correction.

Gene regulatory network inference. To assess the implications of the noise filter on downstream biological interpretations, we used the bulk and single-cell datasets as inputs for various gene regulatory network (GRN) inference tools and compared the results for filtered and unfiltered inputs. For this purpose, we selected several gene subsets, ranging in size from 49 to 996 genes for the bulk dataset and from 57 to 246 genes for the single-cell dataset, based on enrichment analyses performed on the DE genes according to their inclusion in annotated pathways (Supplementary Table S1).

We chose a subset of the GRN inference tools benchmarked by BEELINE (36): GENIE3 (37), GRNBoost2 (38), and PIDC (39). We packaged the tools as Singularity containers (https://github.com/drostlab/network-inference-toolbox) and then assembled them into a custom pipeline (https://github.com/drostlab/network-inference-pipeline).

This pipeline extracts the subsets of genes corresponding to selected pathways and uses them as inputs for the GRN inference tools. The results are rescaled, binarized and compared using the *edgynode* package (v0.3.0, https://github.com/drostlab/edgynode). The edge weights and node degree distributions for all genes across the selected subsets are then visualized.

In detail, the similarity assessment of network topologies was performed using the *edgynode* function network_benchmark_noise_filtering() and was visualized using plot_network_benchmark_noise_filtering(). For this purpose, the inferred networks were converted to a binary format (presence/absence of an edge) using the overall median edge weight per network as a threshold. In network_benchmark_noise_filtering() four different types of matrices are used as input: a weighted adjacency matrix returned by a network inference tool where (i) no noise filter and no quantile normalization (original) was performed (denoted in the figures as −F −N), (ii) a noise filtering but no quantile normalization was performed (+F −N), (iii) no noise filtering but a quantile normalization was performed (−F +N) and (iv) both, noise-filtering and quantile normalization were performed (+F +N).

In a pairwise all versus all comparison, for each gene, the Hamming distance over the binary edge weight vectors was computed using the hamming.distance() function from the R package e1071 v1.7-4, yielding a distribution of distances, which captures how many genes gained or lost their connection with other genes. A Kruskal–Wallis Rank Sum Test was performed using the stats::kruskal.test() function in R to assess whether comparisons of Hamming distance distributions between original, noise-filtered, and normalized combinations were statistically significantly different. Furthermore, visualizing these distributions across comparisons and for all network inference tools facilitated an evaluation of the overall change of network topologies driven by the network inference tool or the normalization/noise-filtering that was applied. These visualizations were then used to assess the impact and robustness of our noise-filter on the interpretation of biological network topologies. We applied the pipeline, including *edgynode*, with the same parameter configurations to both bulk (Yang *et al.*) and single-cell (Cuomo *et al.*) data to retrieve comparable results for direct comparisons. Computationally reproducible analysis scripts to perform all inference steps, data transformations, and visualizations, including the ones used in this study can be found at https://github.com/drostlab/network-inference-pipeline.

Methods, sRNAseq data

The six *A. thaliana* sRNA samples were assessed using multiQC version 1.9 (26). Next, the sequencing adapters (both standard and HD) were trimmed using Cutadapt (version 3.2) (40) and the UEA sRNA Workbench (41). The larger three samples were subsampled without replacement to 8M reads (11); the smaller three samples were left unchanged. The read/sRNA-length distributions were bimodal with peaks at 21nt and 24nt, corresponding to miRNAs and TE sRNAs, respectively. These sRNAs were aligned (using STAR (version 2.7.0a) (27)) to both microRNA hairpins (miRBase Release 22.1) (22) and TEs (obtained from TAIR10) (25).

The six *M. musculus* sRNA samples were processed in a similar way as the plant samples and subsampled without replacement to 3.5M sequences (11). The distribution of read lengths was bimodal with peaks at 22nt and 30nt corresponding to microRNAs and piRNAs respectively. The sRNAs were aligned to microRNA hairpins (miRBase Release 22.1) (22) and TEs (Ensembl release 101).

Methods, PARE data

The three *A. thaliana* PARE samples (GSE113958) were QCed (multiQC version 1.9) (26) and the reads trimmed to 20nt; next, all samples were randomly subsampled without replacement to 25M (11). The subsampled reads were aligned to the reference genome (obtained from TAIR10 (25)) using STAR (using STAR (version 2.7.0a) (27)), with default parameters. The reads aligned to each position along a transcript were grouped on sequence and summarized by frequency. Each summarized fragment was matched (as reverse complement) to *A. thaliana* miRNAs. To visualize the distribution of signal across transcripts, *t*-plots were created, where each point corresponds to a summarized PARE fragment; the points for which a corresponding miRNA was identified were highlighted using the miRNA label (23).

Methods, single cell data

For the single cell SmartSeq2 data, the cellranger software version 3.0 (42) was used for pre-processing, initial quality checks, and to generate the count matrix (it internally uses the STAR aligner). Further quality checks included distribution plots for the number of features, counts, mitochondrial and ribosomal reads per cell; significant outliers were removed during pre-processing. Dimensionality reduction and clustering were performed with the Seurat R package version 3.2 (43). The UMAP reduction method (44) was used for visualization and assessment of results.

Methods, noise quantification

Two approaches were implemented for the identification of noise. (i) The "count matrix approach" is a simple, fast way to obtain a threshold utilizing solely the un-normalized count matrix (m genes x n samples). (ii) The "transcript approach" is more refined, as it takes into account the distribution of signal across the transcript obtained by summarizing the aligned reads from the BAM alignment files. For both approaches, a variety of correlation and distance measures are used to assess the stability of signal across samples (45). Most results were obtained using Pearson Correlation Coefficient (the default); similar results are obtained with other similarity or inverted dissimilarity measures such as Spearman Correlation, Euclidean distance, Kulback-Leibler divergence, and Jensen-Shannon divergence.

Count matrix approach. For each sample in the count matrix, the genes are sorted, in descending order, by abundance. A sliding window approach is used to scan the sorted genes (genes with similar abundances are grouped into "windows"). The window length is a hyper-parameter that can be user-defined or a single value inferred from the data using a Jensen–Shannon entropy based approach (Supplementary methods 1). The sliding step can be varied to reduce computational time at the cost of reducing the number of data points and potentially losing accuracy. For each window, the correlation of the abundances of the genes from the sample of interest and all other samples is calculated and averaged using the arithmetic mean. Per sample, the variation in correlation coefficient (y-axis) is represented versus the average window abundance (x-axis). A correlation threshold (as a hyper-parameter) is used to determine a corresponding abundance threshold as a cutoff—the noise threshold. The correlation threshold is inferred from the data to minimize the variance of noise thresholds across the different samples. Several available approaches are based on the (smoothed) line plot or a binned boxplot of abundance against correlation (Supplementary methods 2). Genes with abundances below the sample specific noise thresholds across samples were excluded from downstream analyses; the average of the thresholds were added to the count matrix, to avoid further biases. By increasing the minimum values in the count matrix from zero to the noise threshold, methods that are based on fold-changes will not emphasise small differences in abundance at very low values, which becomes especially problematic for genes that are seemingly absent in some samples but present and lowly expressed in others. This effect is particularly striking in single-cell data.

Transcript approach. Using the transcript coordinates of the aligned reads as input, the expression profile for each individual transcript was built as an algebraic point sum of the abundances of reads incident to any given position (46); if the alignment was performed per read, the corresponding abundance for every entry was set to +1. For each sample j, and for each transcript T, the point-to-point Pearson Correlation between the expression profile in j and the one in all other samples is calculated. The noise detection is based on the relative location of the distribution of the point-to-point Pearson Correlation Coefficient (p2pPCC) versus the abundances of genes and is specific for each individual sample. For low abundance transcripts the stochastic distribution of reads across the transcript leads to a low p2pPCC; the aim of the approach is to determine the range where the distribution of correlation coefficients (used as proxy for the distribution of reads across a transcript) are above a user-defined threshold; to approximate the signal-to-noise threshold a binning on the abundances was performed. For all examples presented in this study, the binning was done on log2 ranges; the signal-to-noise thresholds were defined as the abundance above which the first quartile of the p2pPCC distribution consistently remains >0.25 (IQR method - see Supplementary methods 2). Once a noise threshold was determined for each sample, the original count matrix was then filtered analogous to the count matrix approach. The BAM files can also be filtered directly by removing all genes which fall below the noise threshold in every sample. Downstream analysis that is not based on the count matrix, such as alternative splicing analysis can also be informed by the noise threshold by setting a lower bound of expression acceptance.

To benchmark the *noisyR* pipeline, for both the count-matrix and transcript-based approaches, we used a server with 32 cores, specifying 16 and 32 cores as *noisyR* parameters, respectively. The specification of the server used for benchmarking is: Kernel: Linux 4.19.0-6-amd64 #1 SMP Debian 4.19.67-2+deb10u2 (2019-11-11), Hardware: Intel(R) Xeon(R) CPU E5-2630 v3 @ 2.40GHz; RAM: 756GB

The benchmarking focused on several covariates: the number of samples from two to 16, the number of exons used for the transcript approach and the sequencing depth/number of mapped reads for the transcript approach.

RESULTS

Noise quantification in bulk RNA-seq data

To exemplify the impact of denoizing on the biological interpretations from bulk RNA-seq experiments, we applied *noisyR* on mRNA-seq and smallRNA-seq data. First, we illustrated the advantages of using the pipeline on a subset of mRNA-seq samples from a 2019 study by Yang *et al.* (19). To assess the distributions of signal we used density plots (Figure 1A) and summaries of Jaccard similarity indices (Figure 1B) across all samples. For the former, we observed a multi-modal distribution that suggests a signal to

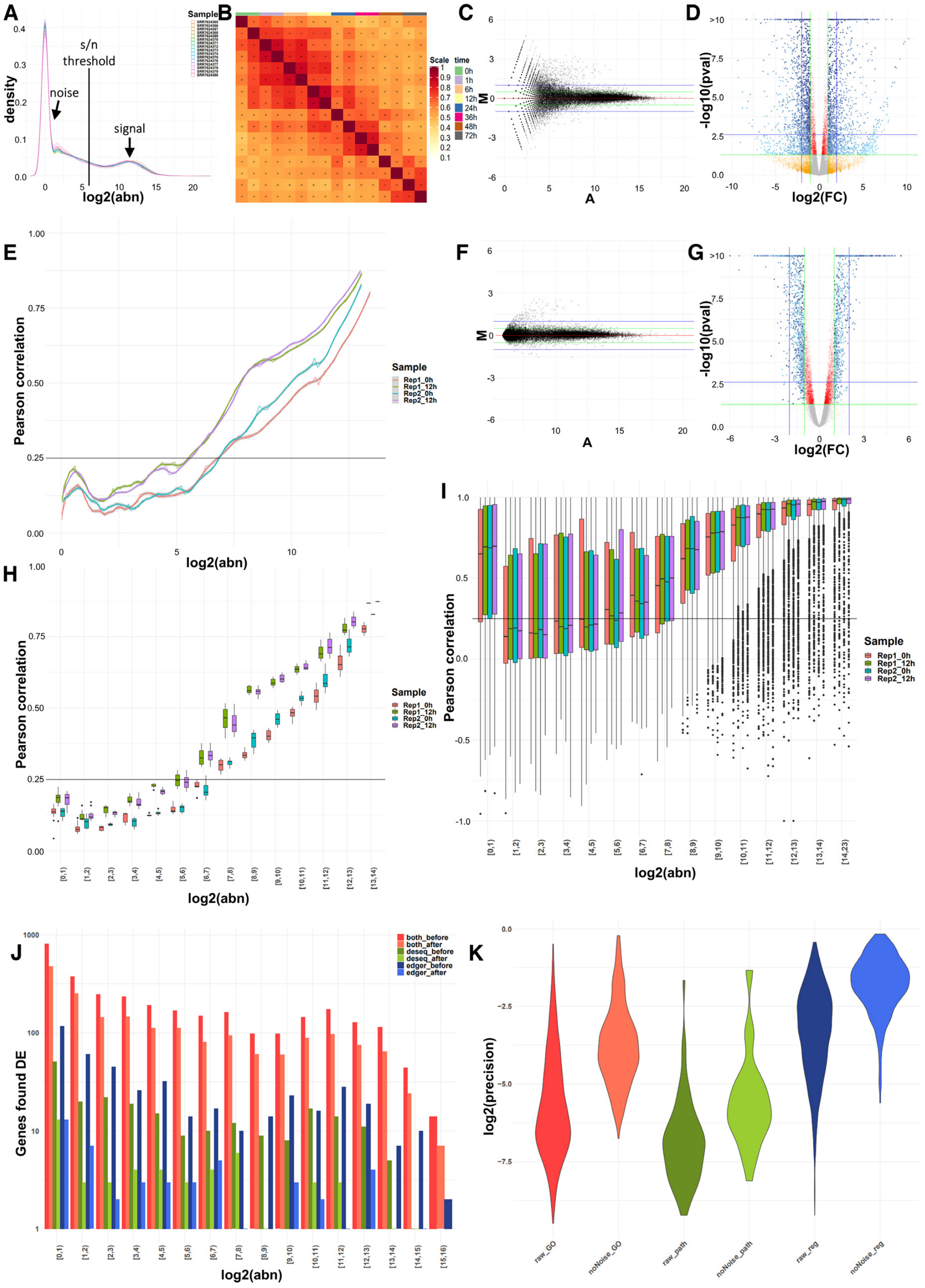

Figure 1. Overview of QC measures and original versus denoised outputs on standard components of an mRNA-seq pipeline. (**A**) Distributions of gene abundances by sample; the RHS distribution corresponds to the biological signal, the LHS distribution to the technical noise; the aim of *noisyR* is the identification of biologically meaningful values for the signal/noise threshold in between. (**B**) JSI on the 100 most abundant genes per sample; the replicates, and consecutive time points share a larger proportion of abundant genes. (**C**) MA plot of the raw abundances for the two 12h biological replicates; a larger proportion of low abundance genes exhibit high fold-changes, potentially biasing the DE calls. (**D**) Volcano plot of differentially expressed genes on the original, normalized count matrix; the colour gradient is proportional to the gene abundance. (**E**) Line plot of the PCC calculated on windows of increasing average abundance for the count matrix-based noise removal approach. (**F**) MA plot of the denoised abundances for the two 12 h biological replicates; the low-level variation is significantly reduced. (**G**) Volcano plot of differentially expressed genes on the denoised count matrix. (**H**) Box plot of the PCC binned by abundance for the count matrix-based noise removal approach. (**I**) Box plot of the PCC binned by abundance for the transcript-based noise removal approach. (**J**) Histogram of the differentially expressed genes found by applying DESeq and edgeR on the original and denoised count matrix respectively, binned by abundance; counts are on a log-scale for visualization. (**K**) Violin plot of the precision (intersection size divided by the query size) for the results of the enrichment analysis performed on the differentially expressed genes found for the original (*raw*) and denoised (*noNoise*) matrices (log-scale). In the Gene Ontology set (*GO*) the terms from Biological Process, Cellular Component and Molecular Function were grouped; in the Pathway set (*path*) the *Kegg* and *Reactome* terms were grouped; in the Regulatory terms (*reg*) the enriched Transcription Factors and microRNA entries were grouped.

noise transition range between [3,7] on $\log_2$ scale; for the latter, the high similarity along the diagonal mirrors the temporal component of the time series. To reduce the number of low abundance, high fold change DE calls (Figure 1C, Supplementary Figure S1A for sample similarity and the secondary DE distribution visible in Figure 1D and Supplementary Figure S1C), we used first the *noisyR* count-based pipeline, on default parameters: window length = 10% × #genes and sliding step = 5% × window length (Figure 1, E and H, Supplementary Figure S1E). We used a correlation threshold of 0.25 and the boxplot median method, a combination of hyper-parameters producing the smallest coefficient of variation across abundance thresholds for the considered samples (Methods); the interquartile ranges (IQRs) of noise thresholds for the different samples ranged between 39 and 63, with an average of 58, for sequencing depths varying between 58M and 82M (Figure 1E, Supplementary Figure S1E). We detected an outlier with a low threshold of 18 (corresponding to a sequencing depth of ∼77M) and three with values of over 100, corresponding to sequencing depths of 73M, 71M and 96M respectively. Next, we applied the transcript approach focusing on the correlation of the expression profiles across exons/transcripts (Methods); despite the higher runtime compared to the count-based approach, the transcript-approach was more robust, as illustrated by the lower variance in signal/noise thresholds across samples (Figure 1I). The parameters that minimized the coefficient of variation were: correlation threshold = 0.26 and the boxplot median method; the resulting noise threshold IQRs ranged between 64 and 79, with an average of 75 and one outlier at 104. The signal/noise thresholds were similar for the two options, with an increased level of detail for the transcript-based approach.

These thresholds were used to exclude noisy genes from the count matrix (∼44k genes were excluded out of ∼56k genes expressed); the number of retained genes were 19.7k and 15.6k for the counts and transcript approaches, respectively. As a DE pre-processing step, the averaged noise threshold was added to all entries in the count matrix (Methods). The effect of the noise removal is illustrated by the narrower distribution in the MA plots (Figure 1F, Supplementary Figure S1B). Next, we performed a DE analysis between the 0 h and 12 h samples of the Yang dataset using the denoised matrix. Following the noise correction, we saw a 46% reduction in the number of DE genes - from 3,607 to 1,952. A large number of low abundance genes with spuriously high fold-changes were no longer called DE (11). Moreover, when comparing the outputs of two standard DE pipelines, edgeR (30) and DEseq2 (7), we noticed that the number of genes identified as DE by both methods only marginally decreased when the noise corrected input is used, whereas the number of DE genes called only with edgeR or only with DeSeq2 decreased significantly (Figure 1J, Supplementary Figure S1F); therefore we observed an increase in output consistency across methods when the noise filtered inputs were used. Moreover, the fold-changes and *P*-values of denoised genes correlated better and we no longer saw a large set of DE genes with (adjusted) *P*-values marginally below the DE threshold (Figure 1D versus G; Supplementary Figure S1C versus D). This step was followed by a functional enrichment analysis focusing on the DE genes, with the genes expressed (post filtering) as background set (31). The number of enriched terms was lower in the denoised data, 1108 versus 4671 in the original analysis; ∼24% of the terms were retained and the terms found with the denoised dataset were approximately a subset of the ones found without the noise correction (∼99.6% of terms found after denoizing were also found prior to noise removal). In addition, the noise-correction terms corresponded to a higher percentage of genes assigned per pathway (Figure 1K). Thus, applying *noisyR* focused the interpretation of results on the enrichment terms with highest confidence, ensuring biological relevance.

The *noisyR* transcript approach was also applied on two small RNA (sRNA) datasets, from plants (*A. thaliana*) and animals (*M. musculus*), respectively. In contrast to the mRNAseq data, sRNAs samples had different correlation vs abundance distributions. Overall low abundance sRNA transcripts/loci contained more noisy entries (47). Also, we observed a sharper increase to high correlation entries highlighting the transition from degraded transcripts to precisely excised sRNAs (48, 49). For both model organisms, miRNA hairpins and transposable elements (TEs) were analysed separately. For the former, we observed overall higher correlations than for mRNAs, likely because of the precise cleavage of the mature duplex and the lack of signal outside the duplex region (50); this characteristic is stronger for the animal case (Supplementary Figure S2C). For both animals and plants, the increasing distribution was clearly detectable (Supplementary Figure S2, A and C). The TE distributions also reflected the characteristics of the underlying sRNAs; for the animal example (Supplementary Figure S2D) we saw a sharper increase along the abundance bins, specific for the piRNAs (51), whereas in plants (Supplementary Figure S2B), the distribution of signal (expressed siRNAs) mirrored the biogenesis of heterochromatin siRNAs (52).

Effect of noise on single cell (smartSeq) data

To illustrate the broad applicability of *noisyR* on different HTS data, we present its output on single cell (smartSeq2) sequencing output focusing on a subset of samples from the dataset presented by Cuomo and colleagues (24); we focused on 6 donors, and one time-point, the number of cells per donor varied between 45 and 107. A common difficulty in single-cell experiments is that due to the higher number of samples/cells, the runtime is much higher if the pipeline is applied without modification, making the transcript approach intractable in practice, for higher number of cells; we also assessed whether the inferred signal/noise threshold was informative.

First, we applied *noisyR* using the count matrix approach on all cells with default parameters; we observed that correlation values rose to a weakly positive plateau (0.2–0.4) and remained stable over a wide range of abundances (Figure 2A). Our interpretation of this result is that lower sequencing depths and higher resolution of smart-seq compared to bulk data induces more dissimilarity across medium abundance values. To alleviate this effect, we grouped cells into a small number of pseudo-samples (similarly as in 53,54), both randomly selected and according to the sample origin

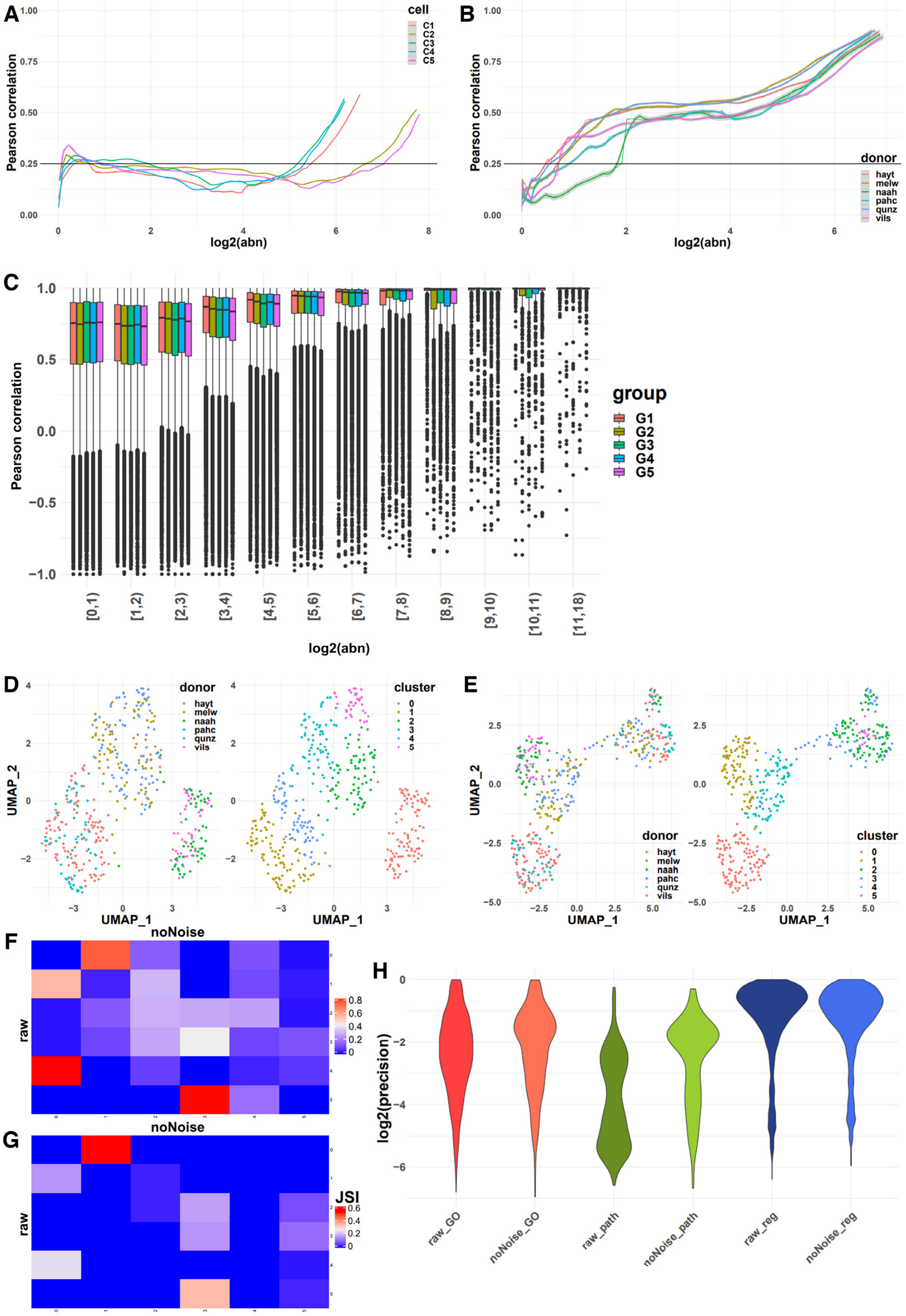

Figure 2. Overview of noise filtering on smartSeq data and impact on biological interpretation of results. (**A**) PCC calculated on windows of increasing average abundance for the count-matrix based noise removal approach applied to the full count matrix of all cells (four cells shown). (**B**) PCC calculated on windows of increasing average abundance for the count-matrix based noise removal approach applied to the "pseudo-samples" formed by grouping all cells from each donor. (**C**) Box plot of the PCC binned by abundance for the transcript-based noise removal approach applied to five groups of five cells each obtained by concatenating the corresponding BAM files. (**D**) UMAP representation of the cells using the raw count matrix grouped by donor (left) and by inferred cluster (right). (**E**) UMAP representation of the cells using the denoised count matrix grouped by donor (left) and by inferred cluster (right) (**F**) Contingency matrix of the clusters formed before and after the noise removal; the shade of each tile represents the proportion of the cluster from the raw matrix (row) that belongs to the corresponding cluster of the denoised matrix (column). (**G**) Heatmap of the Jaccard similarity index between the 50 most significant markers identified for each cluster on the raw matrix (rows) and denoised matrix (columns). (**H**) Violin plot of the precision (intersection size divided by the query size) for the results of the enrichment analysis performed on the marker genes found for each cluster of the raw and denoised matrix respectively (log-scale).

(i.e. donor). In the random grouping, each pseudo-sample summarized the expression of 87 cells, using arithmetic means (five pseudo-samples were generated from 435 cells); for the grouping by donor, the number of cells summarized per pseudo-sample varied between 45 and 107, i.e. all cells from a donor were summarized into a pseudo-sample.

For each pseudo-sample, we applied the count-based approach on the averaged expression of genes. In the resulting *noisyR* output, we observed a clearer step in the abundance-correlation plot (Figure 2B, Supplementary Figure S3A), especially when the cells were grouped by donor. This indicates that an effect of the summarization is a reduction in cell-to-cell variability which also focuses the noise identification procedure. The thresholds obtained via pseudo-sample summarization and count-based noise identification varied between 2 and 4 with an average of 2.6 (corresponding to a sequencing depth per pseudo-sample between 590K and 689K, representative of the average sequencing depth per cell of 640K); these were used in a similar manner as for the bulk data, to produce a denoised count matrix.

As the transcript approach is more computationally intensive, we applied it on a subsampled set of 25 cells. The subsamples were chosen randomly, and the process was reiterated five times, with the requirement that the summarized cells originate from the same donor. Formatting the data for *noisyR* was achieved by concatenating the BAM files for the selected cells and treating them as one sample. Whereas for the count approach, the results were highly variable between the cells, with several instances of low or negative correlations observed even at high abundances (Figure 2A), the results obtained using the transcript approach with the concatenated BAM files were more consistent, with an expected increasing trend in the distribution of correlations (Figure 2C). The correlation distributions were high, even at low abundances, which may be a consequence of the summarization; a suitable threshold may be selected on the median, IQR, or 5–95% range to infer a signal to noise threshold, as the distributions are stable for low values and increase as the abundance increases above ~ 2 on a $\log_2$-scale.

To assess the impact of *noisyR* on the biological interpretation of results, we performed some downstream analyses before and after the noise removal and compared the results. In this study, we focus on the structure and mathematical characteristics of the outputs, rather than specific biological interpretations. The gene abundances were normalized and the cells were clustered using the Seurat R package (43) (see Methods). The different clusterings were visualized using the UMAP (non-linear) dimensionality reduction (44) (Figure 2D and E, Supplementary Figure S3B and C). We observed that cells clustered into three groups of two donors each when original data was used, suggesting a batch effect; However, cells corresponding to the four donors are mixed across clusters, when the denoised data was analysed, suggesting that some of the putative initial batch effect may had been alleviated with the noise correction. We also observed a better separation of clusters in the denoised data, especially on the first UMAP component, which may be an indication of robustness. We further assessed the similarity of the two clustering results using a cell-centred contingency table (Figure 2F). We observed a good correspondence between the original and denoised matrices i.e. clusters 1 and 4 largely merged into cluster 0, and cluster 0 remains intact and turns into cluster 1. While the total number of clusters remained the same (under default parameters), the partitioning of cells was altered, which led us to believe that the results obtained with the original and denoised matrices may be qualitatively different, potentially affecting the downstream biological interpretations. To evaluate the changes in interpretation, we compared the clusters obtained prior to and post noise filtering by identifying the (positive) markers and computing the JSI between the top 50 markers of each cluster (Figure 2G, Supplementary Figure S3D, E). Similarly as for the contingency table, the JSI heatmap shows an analogous correspondence between clusters, albeit weaker. Finally, we performed a functional enrichment analysis of the markers identified pre/post noise filtering. Similarly to the bulk results, there were fewer DE genes (markers per cluster) identified in the denoised dataset, with the precision being higher on average across the different GO terms, pathways, and regulatory terms (Figure 2H). This strengthens our conclusion that the noise filtering process can add focus to the downstream biological analysis without significantly altering the overall composition of the data.

Effects of noise filtering on the biological interpretation of regulatory interactions

One of the main aims of high-throughput sequencing projects, besides the identification of differentially expressed genes (the effect), is to infer the complex interactions of genes that lead to biological functions, the cause (e.g. disease, development or stress response). Understanding these interactions between genes and the corresponding regulatory elements (at transcriptional level, such as transcription factors (55, 56), or post-transcriptional, small RNAs (57)) allows us to unveil the molecular mechanisms encoding phenotypic outcomes, including causes of diseases.

Effect on PARE data on predicting regulatory miRNA/mRNA interactions. First, we sought to understand the effect of noise removal on the identification of miRNA/mRNA interactions. We applied the *noisyR* transcript approach to a Parallel Analysis of RNA Ends Sequencing (PAREseq) dataset (23). The distribution of degraded fragments across transcripts showed the same distribution of correlation vs abundance as we earlier observed for the bulk RNAseq data (Figure 3A). Using a correlation threshold of 0.25, we determined a signal/noise threshold of 60 for this dataset. Next, we matched the highly abundant reads to known miRNAs (Methods, Figure 3B) and illustrated that by removing the noisy reads, having abundance less than the noise threshold (Figure 3C and D), the prediction of interactions is simplified (58), i.e. for most genes only a few peaks were left. In some cases (e.g. Figure 3C), only a very clear peak was retained after the noise removal, while for other transcripts some secondary interactions were kept. These results illustrate that noise-filtering is a crucial step for producing biologically meaningful mRNA/targeting predictions.

Effect on the inference and interpretation of gene regulatory networks. Characterizing direct interactions between reg-

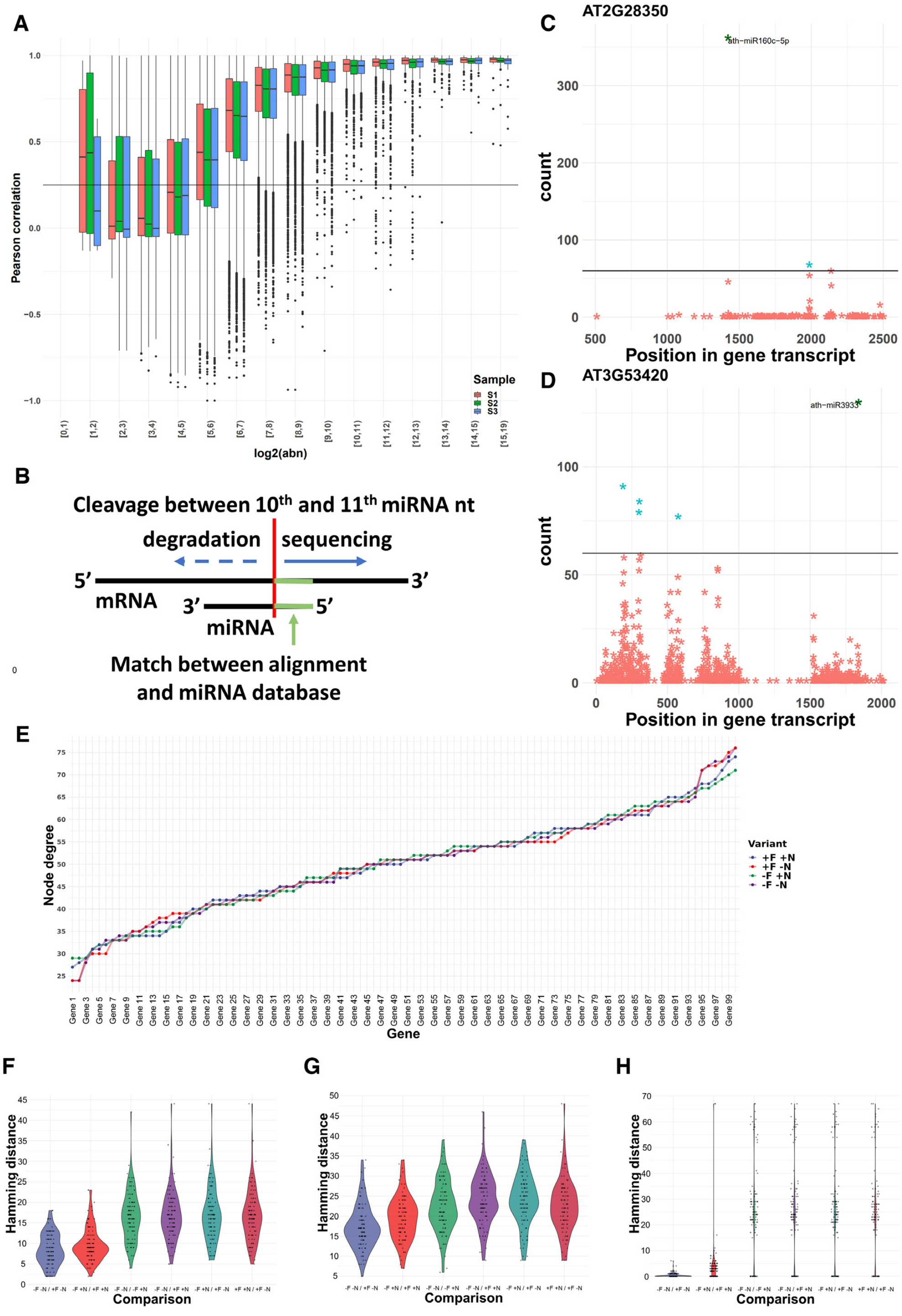

Figure 3. Effect of *noisyR* on PARE-Seq and GRN inference. (**A**) Box plot of the PCC binned by abundance for the transcript-based noise removal approach applied to PARE-Seq data. (**B**) Schematic overview of the microRNA/mRNA interaction; cleavage of the mRNA transcript occurs between the 10th and 11th nucleotide of the microRNA; (**C, D**) PARE *t*-plot illustrating the distribution of degradation products (each point) across the transcripts AT2G28350 and AT3G53420, respectively. All reads with summarised abundance less than the signal/noise thresholds are represented in red; degradation products corresponding to the signal, consistently identified across replicates, are represented in blue. The ones potentially generated by miRNAs are labelled. (**E**) node degree distributions (total number of edges connected to a node/gene) of 102 genes assigned to the neuron differentiation pathway from the Yang et al dataset. The four input data variants are shown: original (–F –N, purple); not noise-filtered but normalized (–F +N, green); noise-filtered but not normalized (+F –N, red); and noise-filtered and normalized (+F +N, blue) sorted by increasing values using −F −N as sorting key. (F–H) Pairwise hamming distance comparisons for each gene between all combinations of original (–F –N), noise-filtered (+F), and normalized (+N) input datasets using 102 Neuron differentiation genes from the bulk RNAseq (Yang et al.) dataset (Methods) show a comparable pattern across different gene regulatory network inference tools: (**F**) GENIE3; (**G**) GRNBoost2; (**H**) PIDC. The results consistently show that across network inference tools, noise-filtering has refining effects on the inferred network topologies in original or normalized data, further illustrating the advantages of noise-filtering to magnify biological signals by reducing technical noise.

ulatory elements and their targets is only feasible for a limited set of interactions (such as the miRNA/mRNA interaction in plants, leading to mRNA degradation (23)). To capture more of the vast complexity of gene interactions, for thousands of genes in tandem, Gene Regulatory Networks (GRNs) have been proposed as a systems biology tool to infer (direct and indirect) regulatory interactions from high-throughput sequencing data (expression data). In a Gene Regulatory Network, nodes represent individual genes (e.g. transcription factors) and edges denote the regulatory interaction between connected genes. When edge-weights are considered, they encode the relative strength of the modeled interaction between two genes. After the network inference step, the resulting topology of GRNs can be used as a proxy for capturing the underlying biological and regulatory complexity of the studied process which in combination with enrichment analyses based on various Gene Ontologies generates a comprehensive model of the investigated process.

We evaluate the impact of noise-filtering on the inference of GRNs on particular network modules (subnetworks), associated with annotated pathways; we quantify the impact of random noise in altering network topologies and subsequent biological interpretations. To achieve this, we run our Network Inference Pipeline (NIP) and *edgynode* network analytics package (Materials and Methods) on bulk RNA-seq datasets using non-noise-filtered original, non-noise-filtered normalized, and noise-filtered normalized count matrices. Bulk RNAseq data has been widely used despite its well-known effect to dilute expression signals of individual cells or tissue types. However, in the context of technical noise, the averaging across cells and tissues may buffer the noise effect on general patterns while reducing the possibility to detect weak, but biologically meaningful, expression signals (e.g. transcription factor (59) or transposable element expression (60)).

Using the Yang dataset (19) in four different setups (original, –F(iltered) –N(normalized); noise-filtered but not normalized, +F –N; not filtered but normalized, –F +N; and noise-filtered and normalized, +F +N) and subsetted on five biological pathways (Placenta development, 46 genes; Neuron differentiation, 102 genes; Cell differentiation, 249 genes; Phosphorus metabolic process, 493 genes; and Multicellular organism development 996 genes), we ran NIP to infer GRNs using three inference approaches GENIE3, GRNBoost2, and PIDC, detailed in Methods. The inferred weighted correlation networks were imported into *edgynode* and rescaled to the range [0, 100] to allow comparisons across inference tools.

Next, all rescaled weight matrices (Supplementary Figure S4E and F) were converted to binary format, using the median value over the entire weight matrix as threshold; a zero was assigned if the weight was below the median value, and a one, if the weight was above the median value. The resulting binary adjacency matrices were then used as input to compute the gene-specific node degrees and to calculate the pairwise Hamming distances for each gene between combinations of original, noise-filtered, and normalized datasets (Figure 3F, Supplementary Figure S4A–D) (Materials and Methods). This per-gene Hamming distance is a direct assessment of the number of edges that differ between inferences and captures both edge gain and loss. A low Hamming distance illustrates a robust network, whereas a high Hamming distance is proportional to large changes in the overall GRN topology. Panels Figure 3F–H illustrate pairwise comparisons between all combinations of input datasets: (i) original –F –N; (ii) not noise-filtered but normalized –F +N; (iii) noise-filtered but not normalized +F –N; and (iv) noise-filtered and normalized +F +N exemplified for 102 genes corresponding to the neuron differentiation pathway and shown for all three network inference tools (GENIE3, Figure 3F; GRNBoost2, Figure 3G; and PIDC Figure 3H). For all network inference tools, a common pattern is the refining effect of noise-filtering on the overall network topologies. Interestingly, the normalization step has, in most cases, much greater impact on the network topology than noise-filtering. This result implies that the filtering procedure can detect and remove technical noise without disrupting the global network topology.

In addition, (Supplementary Figure S4E and F) shows a comparison between rescaled weight matrix distributions for an original and a noise-filtered and normalized network inferred with GENIE3. In this analysis, most genes had a large number of low-weight values within their edge-weight distributions that would result in thousands of biologically meaningless, weakly supported, connections with other genes. Noise-filtering in this bulk RNAseq dataset allows the exclusion of noisy genes as these fall below the median-threshold level which results in a more refined and biologically meaningful network topology after binarization was applied (Methods).

Together, these results suggest that across network inference tools noise-filtering has refining effects on the inferred network topologies in original or normalized data, further illustrating the advantages of noise-filtering to magnify biological signals by reducing technical noise (58).

noisyR package

The *noisyR* package is available on CRAN (https://CRAN.R-project.org/package=noisyr) and comprises an end-to-end pipeline for quantifying and removing technical noise from HTS datasets (Figure 4). The three main pipeline steps are (i) similarity calculation across samples, (ii) noise quantification and (iii) noise removal; each step can be finely tuned using hyper-parameters; optimal, data-driven values for these parameters are also determined. Functions for individual steps are available and fully documented in the package. Also included is a black-box function, *noisyr::noisyr( )* that performs all processing steps, with default parameters. The package is written in the R (version 4.0.3) programming language and is actively maintained on https://github.com/Core-Bioinformatics/noisyR. Also available is a detailed vignette, illustrating the functionality of the pipeline: https://core-bioinformatics.github.io/noisyR/articles/vignette_noisyr_counts.html.

For the sample-similarity calculation, two approaches are available. The count matrix approach uses the original, unnormalized count matrix, as provided after alignment and feature quantification; each sample is processed individually, only the relative expressions across samples are compared. Relying on the hypothesis that the majority of genes are not DE, most of the evaluations are expected to point

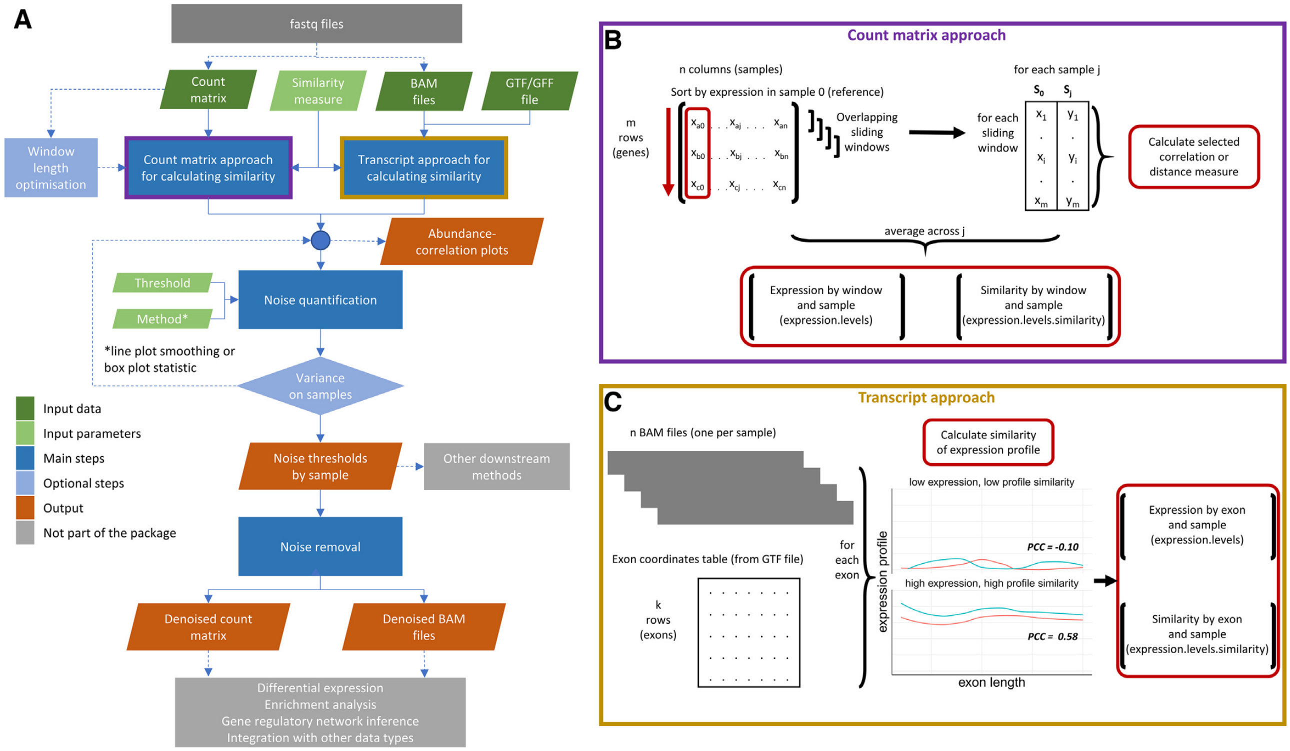
A
fastq files
Count matrix
Similarity measure
BAM files
GTF/GFF file
Window length optimisation
Count matrix approach for calculating similarity
Transcript approach for calculating similarity
Abundance-correlation plots
Threshold
Method*
Noise quantification
*line plot smoothing or box plot statistic
Variance on samples
Input data
Input parameters
Main steps
Optional steps
Output
Not part of the package
Noise thresholds by sample
Other downstream methods
Noise removal
Denoised count matrix
Denoised BAM files
Differential expression
Enrichment analysis
Gene regulatory network inference
Integration with other data types
B
Count matrix approach
n columns (samples)
Sort by expression in sample 0 (reference)
m rows (genes)
Overlapping sliding windows
for each sliding window
for each sample j
Calculate selected correlation or distance measure
average across j
Expression by window and sample (expression.levels)
Similarity by window and sample (expression.levels.similarity)
C
Transcript approach
n BAM files (one per sample)
Calculate similarity of expression profile
Exon coordinates table (from GTF file)
k rows (exons)
for each exon
expression profile
low expression, low profile similarity
PCC = -0.10
high expression, high profile similarity
PCC = 0.58
exon length
Expression by exon and sample (expression.levels)
Similarity by exon and sample (expression.levels.similarity)

Figure 4. Workflow diagram of the *noisyR* pipeline. (**A**) Workflow diagram describing the series of steps comprising the noisyR pipeline. Individual algorithms, finely tuned through hyper-parameters, are highlighted in blue. Optional steps are indicated through higher transparency. Common data pre- and post- processing steps not included in the package are indicated in grey. The steps for the count-matrix- and transcript- approaches are sketched in detail in subplots (**B**) and (**C**).

towards a high similarity across samples. Choosing from a collection of >45 similarity metrics (45), users can select a measure to assess the localized consistency in expression across samples (11). A sliding window approach is used to compare the similarity of ranks or abundances for the selected features between samples. The window length is a hyperparameter, which can be user-defined or inferred from the data (Supplementary methods 1). The transcript approach uses as input the alignment files derived from read-mappers (in BAM format). For each sample and each exon, the point-to-point similarity of expression across the transcript is calculated across samples in a pairwise all-versus-all comparison. The output formats for the two approaches are the same; the number of entries varies, since the count approach focuses on windows, whereas for the transcript approach we calculate a similarity measure for each transcript.

The noise quantification step uses the abundance-correlation (or other similarity measure) relation calculated in step (i) to determine the noise threshold, representing the abundance level below which the gene expression is considered noisy, e.g. if a correlation threshold is used as input then the corresponding abundance from a (smoothed) abundance-correlation line plot is selected as the noise threshold for each sample. The shape of the distribution can vary across experiments; we provide functionality for different thresholds and recommend the choice of the one that results in the lowest variance in the noise thresholds across samples. Options for smoothing, or summarizing the observations in a box plot and selecting the minimum abundance for which the interquartile range (or median) is consistently above the correlation threshold are also available. Depending on the number of observations, we recommend using the smoothing with the count matrix approach, and the boxplot representation with the transcript option. A detailed overview of benchmarking analyses is presented in Supplementary Table S2.

The third step uses the noise threshold calculated in step (ii) to remove noise from the count matrix (and/or BAM file). The count matrix can be calculated by exon or by gene; if the transcript approach is used, the exon approach is employed. Genes/exons whose expression is below the noise thresholds for every sample are removed from the count matrix. The average noise threshold is calculated and added to every entry in the count matrix. This ensures that the fold-changes observed by downstream analyses are not biased by low expression, while still preserving the structure and relative expression levels in the data. If downstream analysis does not involve the count matrix, the thresholds obtained in step (ii) can be used to inform further processing and potential exclusion of some genes/exons from the analysis.

A direct side-by-side comparison with other methods is challenging since some are based on wet-lab validations (e.g. qPCR validations that are restricted to the data presented in the original manuscript) (12–14) and the other methods are mainly focusing on single-cell data with the aim of imputing noise and missing values (15,16). The single cell dataset currently presented is smartSeq, with characteristics similar to mini-bulk samples. We have not found a noise removal method that can be applied to bulk data with a similar aim as *noisyR*.

In addition, a pivotal characteristic of *noisyR* is the calculation of the signal/noise threshold per sample, then used to remove all genes with counts less than the threshold. The second part of the filtering is the averaging of thresholds across samples and adding the resulting value to the whole matrix (to simulate the effect of an offset (61)). This approach marks the difference between this pipeline and the usage of a fixed threshold across all samples. To illustrate this, we compared the consistency of DE calls obtained using the standard pipelines, *edgeR* and *DESeq2*, and the subsequent enrichment results obtained with the *noisyR*-threshold and using a fixed threshold, varied between 0 (no noise correction) and 100. The results are presented in Supplementary Figure S5A and B. The DE results obtained using *noisyR*-thresholds were in line with the results of using comparable, fixed noise thresholds. In addition, we observed a local minimum of the specific differences of *edgeR* and *DESeq2*, achieved using the *noisyR* approach. The larger intersection between the two methods also led to more enriched terms found for the ***noisyR***-corrected data.

DISCUSSION

User-defined or data-driven options for the hyperparameters

noisyR hyperparameters can be used to finely tune the identification of the signal/noise thresholds. To optimize the noise filtering procedure and dampen the stochastically induced differences between samples (e.g. derived from variation in sequencing depth or sample read-complexity) the noise removal step is performed by adding the average of the signal/noise thresholds across samples, on the raw count matrix. Nevertheless, comparable thresholds across the dataset are essential for a meaningful filtering; we recommend the use of consistency and robustness checks throughout the pipeline to ensure that the input samples are comparable, coupled with the data-driven selection of threshold values for setting hyper-parameters. The option of user-defined values is available, however the selected values should be based on observations from the input dataset, rather than exclusively following default recommendations. Next, we discuss in detail the options available for selecting the hyperparameters for a more adaptive noise-filtering based on the structure of the input data.

For the count matrix approach, the length of the sliding windows plays a significant role in assessing the similarity across samples. Smaller windows require more computational time; however the intended level of detail may not always be preferable, as small gene expression fluctuations, from sample to sample, would reduce the across-sample similarity if the abundance range is not wide enough (Figure 5A). Even for medium-high abundances, expression or rank inconsistencies characterize smaller windows, indirectly leading to higher (and more variable across samples) signal/noise thresholds. If the window size is too large, less information is captured by the similarity measure and the accuracy of the noise threshold identification is also reduced (Figure 5B). We recommend medium-sized windows that cover the abundance range in small incremental steps as larger overlaps between windows result in a more robust estimation of similarity-variation. An intuitive approach for determining an informative window size for a dataset relies on monotony changes of the similarity measure, quantified as the number of times the derivative of the correlation (as a function of abundance) changes sign. On several datasets, this resulted in a window length of 1/10th of the total number of expressed genes and a sliding window step size of 1/20th of the total gene number. A different tactic, also implemented in *noisyR*, tackles this task from a different direction; it relies on optimizing the window length using an entropy-based approach with the Jensen-Shannon divergence to assess the stability achieved as the window length is increased (Supplementary methods 1). The shape of the distribution of correlations changes as the window length increases; however the change is less significant (evaluated using a *t*-test) for larger windows. The first point of stability is selected as the optimal window length, as it provides the largest possible granularity while maintaining robustness. The results from this approach are also consistent with earlier, empirical findings when applied to the Yang dataset (19).

Yet another hyperparameter is the similarity measure; we compared the results for different correlation and distance metrics. We aim to achieve a high consistency in quantifying the signal/noise thresholds that is independent of the similarity measure. We tested the standard parametric and non-parametric correlation measures as well as the ones implemented in the *philentropy* package (45), which provides a variety of >45 distance measures. Dissimilarity measures are inverted for comparison purposes (Figure 5C–F illustrates the Spearman correlation, Euclidean distance, Kulback-Leibler divergence, and Jensen–Shannon divergence). Some measures have fixed ranges (e.g. the correlation coefficients), while others are semi- or unbounded. This raises the question of how to choose a similarity threshold when the range of values resulting from the similarity measure is unknown. Inspired by the correlation threshold, which provides a good separation at 0.25 for many datasets, we focus, as a starting point, on the naive assumption to use a quarter of the full range of the observed similarity values as a first cut-off approximation. Picking a threshold in a data-driven manner is, however, preferable and, in this case, achievable. Selecting from a variety of threshold values that minimize the coefficient of variation (standard deviation divided by the mean) of the corresponding noise thresholds in different samples is an empirical approach that works in practice. If the samples are semantically grouped, e.g. replicates or time points, it may be better to minimize the variation in each individual group rather than across the full experimental design.

Effect of aligner choice on noise quantification

The choice of the read-aligner was shown to influence the downstream DE analyses when the same quantification model was applied (17). To assess the effect of different alignment approaches on the quantification and observed levels of noise, mRNA quantification using feature-

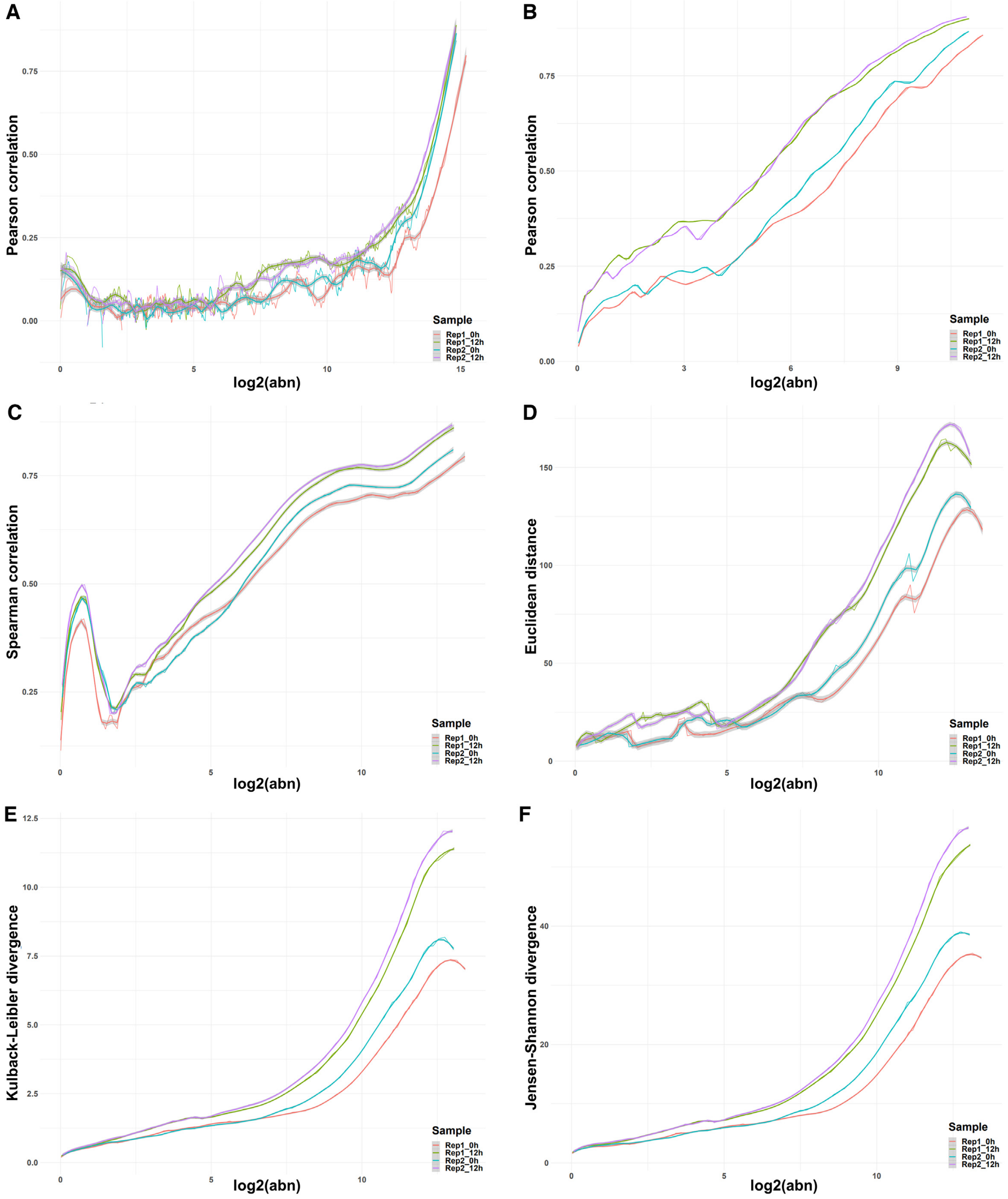

Figure 5. Effects of hyperparameter selection on noise quantification. (**A**) PCC-abundance plot for a window length of 1,000 genes, ∼1/5th of the default. (**B**) PCC-abundance plot for a window length of 20 000 genes, ∼4 times the default. (**C**) Spearman correlation plotted against abundance for the default window length of ∼5500. (**D**) Inverse of the Euclidean distance plotted against abundance for the default window length of ∼5500. (**E**) Inverse of the Kulback–Leibler divergence plotted against abundance for the default window length of ∼5500. (**F**) Inverse of the Jensen-Shannon divergence plotted against abundance for the default window length of ∼5500.

Counts (28) was performed on reads aligned with STAR (27), HISAT2 (35) and Bowtie2 (34). The latter two were run both using their default parameters and with parameters set to match STAR functionality. For the count-based approach, the distribution of the Pearson Correlation Coefficients across abundance bins (Figure 6A) shows that noise levels were relatively consistent regardless of the applied alignment algorithm. Similarly, for the transcript-based approach, the correlation distributions across abundance bins (Figure 6B) illustrate little variation across aligners (additional examples in Supplementary Figure S7A and B). The estimated signal/noise thresholds were also comparable between the datasets generated by different aligners (Figure 6C), with transcripts-based noise results being less variable. Once the noise correction was applied, the substantial peak in the abundance distributions around zero (Figure 6D) was removed or significantly diminished and a second peak corresponding to the true signal was revealed around $\log_2$(abundance) of five using both counts and transcripts based approaches (Figure 6E and F respectively). The similarity of the abundance distributions across the datasets produced by the different aligners was observable both before and after the noise correction. This demonstrates that the proposed correction approaches are non-destructive and preserve the underlying biological signal. To further validate this point, the overlap between edgeR and DESeq2 analyses was investigated. The differentially expressed (DE) genes (adjusted P-value < 0.05 and $|\log_2(\mathrm{FC})| > 1$) detected by the two methods were compared for outputs produced using STAR (Figure 1J), Bowtie2 (Figure 6G) and HISAT2 (Figure 6H). In all cases, there were fewer DE genes in total after noise correction was applied, and the specific differences for each DE method were reduced. The same conclusions were reached for the processing with Bowtie2 and HISAT2 applied with their default parameters (Supplementary Figure S7C).

The effects of noise-filtering on GRN inference for single cell RNAseq data

The recent emergence of single-cell sequencing technologies enabled the simultaneous assessment of expression variation between individual cells across thousands of cell-lineages. Although conceptually powerful, sequencing depths remain constrained by cost and in comparison to bulk RNAseq experiments the total number of reads is now shared among these (hundreds-) thousands of individual cells expressing thousands of genes each. This limit on the sequencing depth per cell underlines, yet again, the technical noise, whereby the quantification of low-abundance transcripts can be the result of either low biological expression or due to stochastic effects (likelihood) of read capturing. The requirement of an adaptive noise-filtering pipeline is fulfilled by *noisyR*; the retained gene expression levels increases the robustness of quantification of single-cell data.

Analogous to the Yang et al. dataset, we used the Cuomo (24) dataset in four different setups (original, –F –N; noise-filtered but not normalized, +F –N; not filtered but normalized, –F +N; and noise-filtered and normalized, +F +N) and subsampled into three distinct biological pathways (Metabolism, 57 genes; Catalytic activity, 133 genes; Cellular metabolic process, 246 genes), we ran the Network Inference Pipeline to infer GRNs using the same three inference methods GENIE3, GRNBoost2 and PIDC (Materials and Methods) as used for bulk RNAseq data. The inferred weighted correlation networks were imported into *edgynode* and rescaled (Supplementary Figure S6C and D) analogous to the bulk RNAseq data shown in Results and Methods. The resulting pairwise Hamming distances for each gene between combinations of original, noise-filtered, and normalized datasets and for genes corresponding to various biological pathways (Figure 7A–C, Supplementary Figure S6A and B) show that total Hamming distances over all genes are larger in single-cell data. This implied that noise-filtering had a more significant/refining impact on the inference and biological interpretations drawn from single-cell data when compared with analogous bulk RNA data.

Figure 7D–F illustrates such analogous pairwise comparisons between all combinations of input datasets: (i) original –F –N; (ii) not noise-filtered but normalized –F +N; (iii) noise-filtered but not normalized +F –N and (iv) noise-filtered and normalized +F +N exemplified for 133 genes corresponding to catalytic activity pathways derived from single-cell RNAseq data (Cuomo *et al.*) and also shown for all three network inference tools (GENIE3, Figure 7A; GRNBoost2, Figure 7B; and PIDC Figure 7C). Analogous to the bulk RNAseq results, noise-filtering has smaller effects on changes in network topologies than the normalization step. Interestingly, it seems that the overall effect of noise-filtering in single-cell data has a stronger impact than in bulk RNAseq data (Figure 3F–H). Together, these conclusions hint toward a more useful effect of noise-filtering in single-cell data as is particularly expected for datasets with limited sequencing depth, but high individual cell numbers.

These highlight the positive effects of noise-filtering on magnifying meaningful biological signals in single-cell RNAseq data, with more significant effects in single-cell data due to the nature of technical noise induced by sequencing depth-constraints in combination with technical variation.

Using *noisyR*, we demonstrate that unfiltered RNAseq quantifications can cause spurious false positive effects in various standard expression-based data analysis steps. To overcome these limitations, we introduce an R package to equip life scientists with a flexible solution, applicable across different bulk and single cell datasets, for excluding inconsistent transcript quantifications that would otherwise introduce stochastic variability in processed datasets. A comprehensive selection of automatic and semi-automatic threshold detection options provided by *noisyR* allows the robust inference of noise-thresholds to exclude low-confidence transcripts from processed RNAseq data. We illustrate the importance of such a noise-filtering procedure by assessing the convergence of DE identification and by inferring and comparing gene regulatory networks from various biological pathways, across gold-standard network inference tools. As a result, we find that noise-filtering is indeed able to significantly reduce stochastic effects magni-

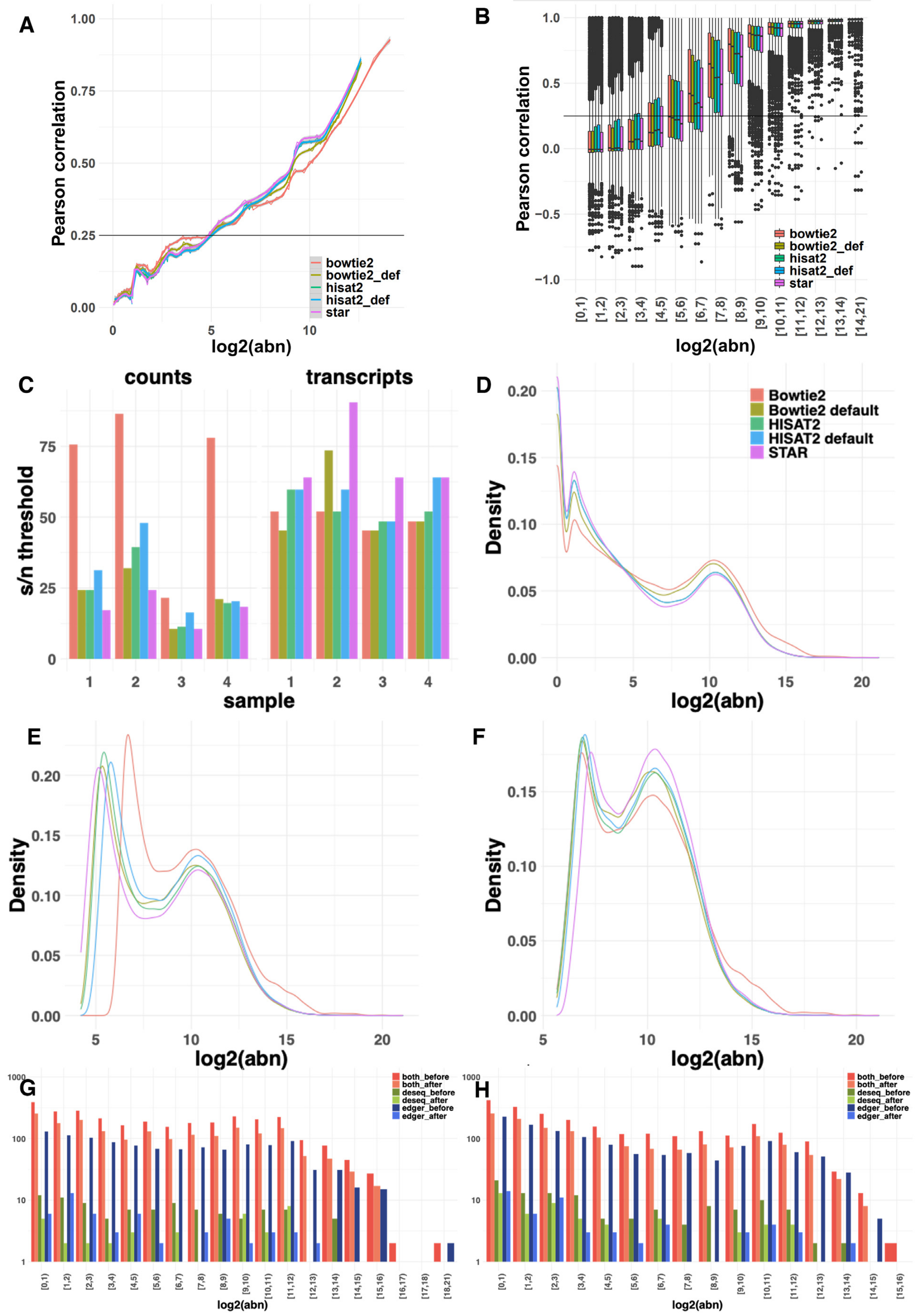

Figure 6. Assessment of aligner choice on noise quantification. (**A**) The distribution of PCC across abundance bins in datasets for a single mRNAseq sample obtained by STAR, Bowtie2 and HISAT2 alignment followed by featureCounts quantification using a counts-based noise removal approach. (**B**) The distribution of PCC across abundance bins in aligned read counts obtained by the five aligners for the same sample in the transcript-based noise correction approach. (**C**) The detected signal-to-noise thresholds in the four mRNAseq samples varied when the counts or transcripts-based noise correction methods were applied. (**D**) The distribution of abundance of reads aligned by the five algorithms and quantified by featureCounts. (**E**) The distribution of abundance of the quantified counts after counts-based noise correction (**F**) The distribution of abundance of the quantified counts after transcripts-based noise correction. (**G**) The number of the differentially expressed genes found by applying DESeq and edgeR on the original and denoised (using transcripts-based approach) count matrices obtained by Bowtie2 alignment. (**H**) The overlap between the DESeq and edgeR analyses performed on the original and denoised counts matrices obtained by HISAT2.

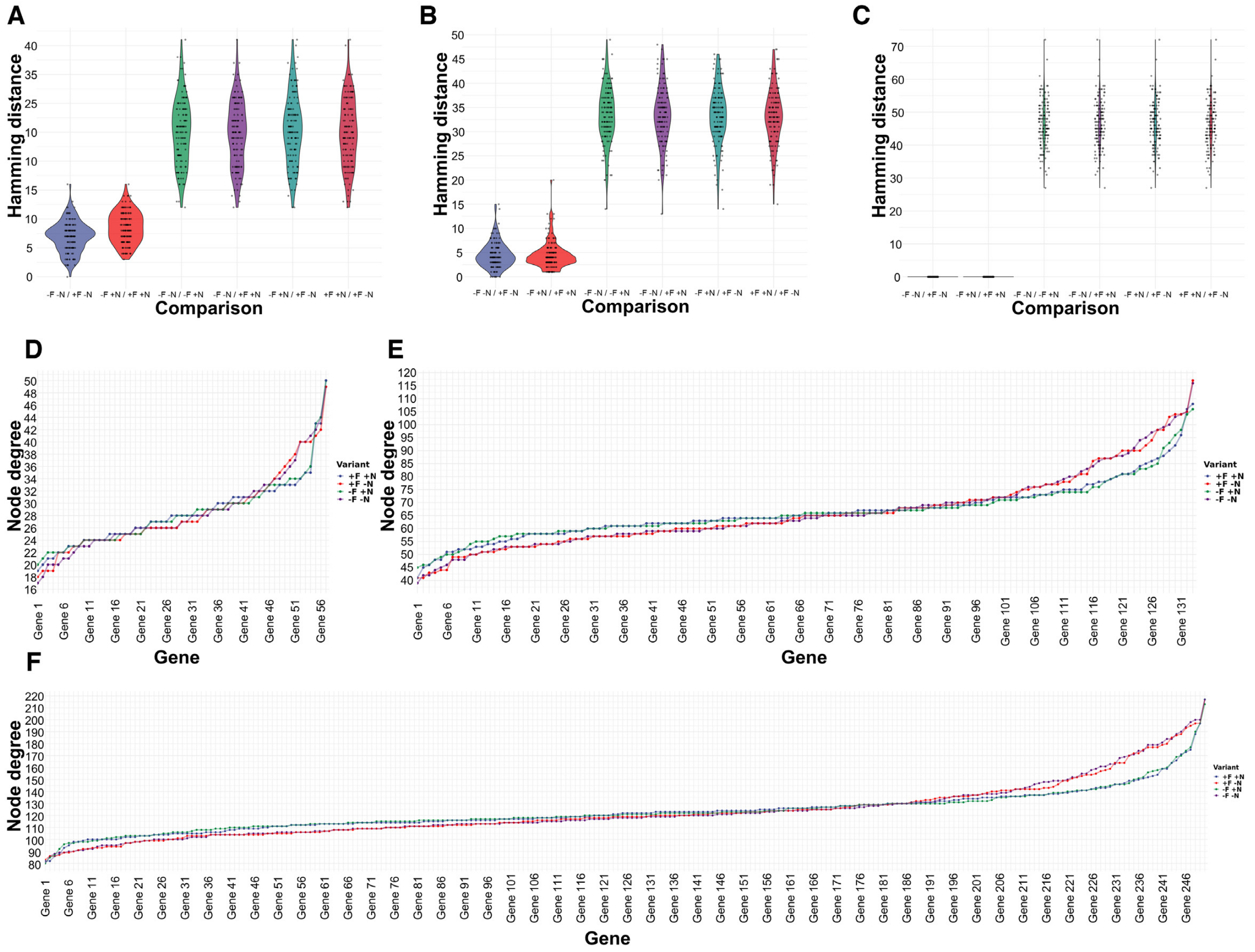

Figure 7. Node degree distributions and pairwise Hamming distance distributions between combinations of original, noise-filtered, and normalized input smartSeq datasets. (A–C) Pairwise Hamming distance comparisons for each gene between all combinations of original (–F –N), noise-filtered (+F), and normalized (+N) input datasets using 133 genes associated with catalytic activity pathways from the smartSeq (Cuomo et al.) dataset (Methods) show a comparable pattern across different gene regulatory network inference tools: (**A**) GENIE3; (**B**) GRNBoost2; (**C**) PIDC. The results consistently show that across network inference tools noise-filtering has refining effects on the inferred network topologies in original or normalized data, further illustrating the advantages of noise-filtering to magnify biological signals by reducing technical noise. (D–F). For the smartSeq (Cuomoet al.) dataset the node degree distributions (total number of edges connected to a node/gene) of three sets of genes corresponding to different pathways are shown: (**D**) 57 genes associated with metabolism pathways; (**E**) 133 genes associated with catalytic activity pathways; (**F**) 246 genes associated with the cellular metabolic process. All four input data variants are shown: original (–F –N, purple); not noise-filtered but normalized (—F +N, green); noise-filtered but not normalized (+F –N, red); and noise-filtered and normalized (+F +N, blue) sorted by increasing values using –F –N as sorting key.

fying underlying biological signals, thereby yielding more robust biological interpretations.

DATA AVAILABILITY

The ***noisyR*** package is available on CRAN (https://CRAN.R-project.org/package=noisyr) and comprises an end-to-end pipeline for quantifying and removing technical noise from HTS datasets. The package is written in the R (version 4.0.3) programming language and is actively maintained on https://github.com/Core-Bioinformatics/noisyR.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

I.M., E.L., E.W. and I.I.M. acknowledge the constructive feedback from the Core Bioinformatics group. S.A.V.U., L.M. and H.G.D. acknowledge that their work was supported by the Max Planck Society.

Author contributions: I.M. and I.I.M. designed the study and implemented the R package *noisyR*; S.A.V.U., L.M. and H.G.D. implemented the R package *edgynode*, the analyses were performed by I.M., I.I.M. (the mRNA bulk and single cell analyses), E.W., I.I.M. (sRNA analysis), L.M. and H.G.D. (GRN inference), E.L., I.M., I.I.M. (comparison across tools). I.M., H.G.D. and I.I.M. wrote the manuscript; all authors read and approved the submitted manuscript.

FUNDING

Wellcome-MRC Cambridge Stem Cells Institute. Funding for open access charge: Cambridge Stem Cell Institute and the University of Cambridge.

Conflict of interest statement. None declared.

REFERENCES

1. Stark,R., Grzelak,M. and Hadfield,J. (2019) RNA sequencing: the teenage years. *Nat. Rev. Genet.*, **20**, 631–656.
2. Oshlack,A., Robinson,M.D. and Young,M.D. (2010) From RNA-seq reads to differential expression results. *Genome Biol.*, **11**, 220.
3. Conesa,A., Madrigal,P., Tarazona,S., Gomez-Cabrero,D., Cervera,A., Mcpherson,A., Szcześniak,M.W., Gaffney,D.J., Elo,L.L., Zhang,X. *et al.* (2016) A survey of best practices for RNA-seq data analysis. *Genome Biol.*, **17**, 13.
4. Li,Mo and Belmonte,J.C.I. (2017) Ground rules of the pluripotency gene regulatory network. *Nat. Rev. Genet.*, **18**, 180–191.
5. Parekh,S., Ziegenhain,C., Vieth,B., Enard,W. and Hellmann,I. (2016) The impact of amplification on differential expression analyses by RNA-seq. *Sci. Rep.*, **6**, 25533.
6. Hansen,K.D., Brenner,S.E. and Dudoit,S. (2010) Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res.*, **38**, e131.
7. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.
8. Stuart,T. and Satija,R. (2019) Integrative single-cell analysis. *Nat. Rev. Genet.*, **20**, 257–272.
9. Rapaport,F., Khanin,R., Liang,Y., Pirun,M., Krek,A., Zumbo,P., Mason,C.E., Socci,N.D. and Betel,D. (2013) Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol.*, **14**, R95.
10. Wang,Z., Gerstein,M. and Snyder,M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, **10**, 57–63.
11. Mohorianu,I., Bretman,A., Smith,D.T., Fowler,E.K., Dalmay,T. and Chapman,T. (2017) Comparison of alternative approaches for analysing multi-level RNA-seq data. *PLoS One*, **12**, e0182694.
12. Park,G., Park,J.K., Shin,S-Ho, Jeon,H.-J., Kim,N.K.D., Kim,Y.J., Shin,H.-T., Lee,E., Lee,K.H., Son,D.-S. *et al.* (2017) Characterization of background noise in capture-based targeted sequencing data. *Genome Biol.*, **18**, 136.
13. Fischer-Hwang,I., Ochoa,I., Weissman,T. and Hernaez,M. (2019) Denoising of aligned genomic data. *Sci. Rep.*, **9**, 15067.
14. Shiroguchi,K., Jia,T.Z., Sims,P.A. and Xie,X.S. (2012) Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. *Proc. Natl. Acad. Sci.*, **109**, 1347–1352.
15. Eraslan,G., Simon,L.M., Mircea,M., Mueller,N.S. and Theis,F.J. (2019) Single-cell RNA-seq denoising using a deep count autoencoder. *Nat. Commun.*, **10**, 390.
16. Jia,C., Hu,Yu, Kelly,D., Kim,J., Li,M. and Zhang,N.R. (2017) Accounting for technical noise in differential expression analysis of single-cell RNA sequencing data. *Nucleic Acids Res.*, **45**, 10978–10988.
17. Srivastava,A., Malik,L., Sarkar,H., Zakeri,M., Almodaresi,F., Soneson,C., Love,M.I., Kingsford,C. and Patro,R. (2020) Alignment and mapping methodology influence transcript abundance estimation. *Genome Biol.*, **21**, 239.
18. Corchete,L.A., Rojas,E.A., Alonso-López,D., De Las Rivas,J., Gutiérrez,N.C. and Burguillo,F.J. (2020) Systematic comparison and assessment of RNA-seq procedures for gene expression quantitative analysis. *Sci. Rep.*, **10**, 19737.
19. Yang,P., Humphrey,S.J., Cinghu,S., Pathania,R., Oldfield,A.J., Kumar,D., Perera,D., Yang,J.Y.H., James,D.E., Mann,M. *et al.* (2019) Multi-omic profiling reveals dynamics of the phased progression of pluripotency. *Cell Syst.*, **8**, 427–445.
20. Paicu,C., Mohorianu,I., Stocks,M., Xu,P., Coince,A., Billmeier,M., Dalmay,T., Moulton,V. and Moxon,S. (2017) miRCat2: accurate prediction of plant and animal microRNAs from next-generation sequencing datasets. *Bioinformatics*, **33**, 2446–2454.
21. Wallach,T., Wetzel,M., Dembny,P., Staszewski,O., Krüger,C., Buonfiglioli,A., Prinz,M. and Lehnardt,S. (2020) Identification of CNS injury-related microRNAs as novel toll-like receptor 7/8 signaling activators by small RNA sequencing. *Cells*, **9**, 186.
22. Kozomara,A., Birgaoanu,M. and Griffiths-Jones,S. (2019) miRBase: from microRNA sequences to function. *Nucleic Acids Res.*, **47**, D155–D162.
23. Thody,J., Moulton,V. and Mohorianu,I. (2020) PAREameters: a tool for computational inference of plant miRNA–mRNA targeting rules using small RNA and degradome sequencing data. *Nucleic Acids Res.*, **48**, 2258–2270.
24. Cuomo,A.S.E., Seaton,D.D., McCarthy,D.J., Martinez,I., Bonder,M.J., Garcia-Bernardo,J., Amatya,S., Madrigal,P., Isaacson,A., Buettne,F. *et al.* (2020) Single-cell RNA-sequencing of differentiating iPS cells reveals dynamic genetic effects on gene expression. *Nat. Commun.*, **11**, 810.
25. Berardini,T.Z., Reiser,L., Li,D., Mezheritsky,Y., Muller,R., Strait,E. and Huala,E. (2015) The arabidopsis information resource: making and mining the "gold standard" annotated reference plant genome. *Genesis*, **53**, 474–485.
26. Ewels,P., Magnusson,M., Lundin,S. and Käller,M. (2016) MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, **32**, 3047–3048.
27. Dobin,A., Davis,C.A., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P., Chaisson,M. and Gingeras,T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.
28. Liao,Y., Smyth,G.K. and Shi,W. (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, **30**, 923–930.
29. Jolliffe,I.T. and Cadima,J. (2016) Principal component analysis: a review and recent developments. *Philos. Trans. R. Soc., A*, **374**, 20150202.
30. Mccarthy,D.J., Chen,Y. and Smyth,G.K. (2012) Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.*, **40**, 4288–4297.
31. Raudvere,U., Kolberg,L., Kuzmin,I., Arak,T., Adler,P., Peterson,H. and Vilo,J., (2019) g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.*, **47**, W191–W198.
32. Kanehisa,M. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, **28**, 27–30.
33. Viteri,G., Matthews,L., Varusai,T., Gillespie,M., Milacic,M., Cook,J., Weiser,J., Shorser,S., Sidiropoulos,K., Fabregat,A. *et al.* (2019) Reactome and ORCID—fine-grained credit attribution for community curation. *Database*, **2019**, baz123.
34. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
35. Kim,D., Paggi,J.M., Park,C., Bennett,C. and Salzberg,S.L. (2019) Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.*, **37**, 907–915.
36. Pratapa,A., Jalihal,A.P., Law,J.N., Bharadwaj,A. and Murali,T.M. (2020) Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nat. Methods*, **17**, 147–154.
37. Huynh-Thu,V.A., Irrthum,A., Wehenkel,L. and Geurts,P. (2010) Inferring regulatory networks from expression data using tree-based methods. *PLoS One*, **5**, e12776.
38. Moerman,T., Aibar Santos,S., Bravo González-Blas,C., Simm,J., Moreau,Y., Aerts,J. and Aerts,S. (2019) GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks. *Bioinformatics*, **35**, 2159–2161.
39. Chan,T.E., Stumpf,M.P.H. and Babtie,A.C. (2017) Gene regulatory network inference from single-cell data using multivariate information measures. *Cell Syst.*, **5**, 251–267.
40. Martin,M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, **17**, 10.
41. Stocks,M.B., Mohorianu,I., Beckers,M., Paicu,C., Moxon,S., Thody,J., Dalmay,T. and Moulton,V. (2018) The UEA sRNA Workbench (version 4.4): a comprehensive suite of tools for analyzing miRNAs and sRNAs. *Bioinformatics*, **34**, 3382–3384.
42. Zheng,G.X.Y., Terry,J.M., Belgrader,P., Ryvkin,P., Bent,Z.W., Wilson,R., Ziraldo,S.B., Wheeler,T.D., McDermott,G.P., Zhu,J. *et al.* (2017) Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.*, **8**, 14049.

43. Stuart,T., Butler,A., Hoffman,P., Hafemeister,C., Papalexi,E., Mauck III,W.M., Hao,Y., Stoeckius,M., Smibert,P. and Satija,R. (2019) Comprehensive Integration of Single-Cell Data. *Cell*, **177**, 1888–1902.
44. Becht,E., Mcinnes,L., Healy,J., Dutertre,C.-A., Kwok,I.W.H., Ng,L.G., Ginhoux,F. and Newell,E.W. (2019) Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.*, **37**, 38–44.
45. Drost,H.-.G. (2018) Philentropy: information theory and distance quantification with R. *Journal of Open Source Software*, **3**, 765.
46. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,A.D. (2002) The Human Genome Browser at UCSC. *Genome Res.*, **12**, 996–1006.
47. Mohorianu,I., Stocks,M.B., Wood,J., Dalmay,T. and Moulton,V. (2013) CoLIde: a bioinformatics tool for CO-expression-based small RNA Loci Identification using high-throughput sequencing data. *RNA Biology*, **10**, 1221–1230.
48. Kim,V.N, Han,J. and Siomi,M.C. (2009) Biogenesis of small RNAs in animals. *Nat. Rev. Mol. Cell Biol.*, **10**, 126–139.
49. Borges,F. and Martienssen,R.A. (2015) The expanding world of small RNAs in plants. *Nat. Rev. Mol. Cell Biol.*, **16**, 727–741.
50. Ha,M. and Kim,V.N (2014) Regulation of microRNA biogenesis. *Nat. Rev. Mol. Cell Biol.*, **15**, 509–524.
51. Czech,B., Munafò,M., Ciabrelli,F., Eastwood,E.L., Fabry,M.H., Kneuss,E. and Hannon,G.J. (2018) piRNA-guided genome defense: from biogenesis to silencing. *Annu. Rev. Genet.*, **52**, 131–157.
52. Papareddy,R.K., Páldi,K., Paulraj,S., Kao,P., Lutzmayer,S. and Nodine,M.D. (2020) Chromatin regulates expression of small RNAs to help maintain transposon methylome homeostasis in Arabidopsis. *Genome Biol.*, **21**, 251.
53. Kang,H.M., Subramaniam,M., Targ,S., Nguyen,M., Maliskova,L., Mccarthy,E., Wan,E., Wong,S., Byrnes,L., Lanata,C.M. *et al.* (2018) Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.*, **36**, 89–94.
54. Lun,A.T.L. and Marioni,J.C. (2017) Overcoming confounding plate effects in differential expression analyses of single-cell RNA-seq data. *Biostatistics*, **18**, 451–464.
55. Andersson,R. and Sandelin,A. (2020) Determinants of enhancer and promoter activities of regulatory elements. *Nat. Rev. Genet.*, **21**, 71–87.
56. Levo,M. and Segal,E. (2014) In pursuit of design principles of regulatory sequences. *Nat. Rev. Genet.*, **15**, 453–468.
57. Holoch,D. and Moazed,D. (2015) RNA-mediated epigenetic regulation of gene expression. *Nat. Rev. Genet.*, **16**, 71–84.
58. Thody,J., Folkes,L., Medina-Calzada,Z., Xu,P., Dalmay,T. and Moulton,V. (2018) PAREsnip2: a tool for high-throughput prediction of small RNA targets from degradome sequencing data using configurable targeting rules. *Nucleic Acids Res.*, **46**, 8730–8739.
59. Ang,C.E. and Wernig,M. (2018) Profiling DNA–transcription factor interactions. *Nat. Biotechnol.*, **36**, 501–502.
60. Lanciano,S. and Cristofari,G. (2020) Measuring and interpreting transposable element expression. *Nat. Rev. Genet.*, **21**, 721–736.
61. Mohorianu,I., Schwach,F., Jing,R., Lopez-Gomollon,S., Moxon,S., Szittya,G., Sorefan,K., Moulton,V. and Dalmay,T. (2011) Profiling of short RNAs during fleshy fruit development reveals stage-specific sRNAome expression patterns. *Plant J.*, **67**, 232–246.