FRONT MATTER

Title

- noisyR: Enhancing biological signal in sequencing datasets by characterising random technical noise
- Noise removal unveils biological signal from sequencing data

Authors

I. Moutsopoulos,[1] L. Maischak,[2] E. Lauzikaite,[1] S. A. Vasquez Urbina,[2] E. C. Williams,[1] H. G. Drost,[2] I. I. Mohorianu[1]*

Affiliations

[1] Wellcome-MRC Cambridge Stem Cell Institute, Jeffrey Cheah Biomedical Centre Cambridge Biomedical Campus, University of Cambridge, CB2 0AW, UK.

[2] Computational Biology Group, Department of Molecular Biology, Max Planck Institute for Developmental Biology, Max-Planck Ring 1, 72076 Tübingen, Germany.

* Corresponding author: I. I. Mohorianu, iim22@cam.ac.uk

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 unreproducible, random background noise to capture true, functionally meaningful biological signals is still a challenge. Intrinsic sequencing variability that introduces 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 can be applied to count matrices and sequencing data; it outputs sample-specific signal/noise thresholds and filtered expression matrices.

We exemplify the effects of minimising technical noise on plant and animal datasets, across various sequencing assays: coding, non-coding RNAs and their 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.

Keywords: next generation sequencing, noise, bulk sequencing, single-cell sequencing, count matrix, expression profile, differential expression, enrichment analysis, gene regulatory network.

Teaser

Noise removal from sequencing quantification improves the convergence of downstream tools and robustness of conclusions.

MAIN TEXT

Introduction

High-throughput sequencing (HTS) became a new standard in most life science studies yielding unprecedented insights into the complexity of biological processes. This 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 characterisation of differential expression (DE) at tissue and cellular levels (*3*) or the inference and characterisation 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 call or create spurious patterns biasing downstream interpretations. Statistical methods developed to date (*7-9*), focused mainly on batch/background correction, normalisation, and evaluation of DE have been developed to mitigate the impact of these biases on DE analyses (*10*). A noise filter for pre-processing 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 different technologies may exhibit different technical biases, the sequencing bias across an experiment was expected to be uniform. The initial assumption was 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 (*11*). However, in practice we observe a reproducible, yet uneven distribution of signal across transcripts (*11*); 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 (*12*). 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 characterised by such lack of coverage-uniformity as "noisy".

The presence of noise in high-throughput sequencing data has been widely acknowledged, and there have been several attempts to understand and quantify it. A recent study (*13*) presented a variety of common experimental errors that may increase sequencing noise and proposed ways to avoid them such as using a mild acoustic shearing condition to minimise the occurrence of DNA damage. Fischer-Hwang and colleagues (*14*) presented a denoising 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 (*15*).

On the computational side, several imputation and denoising algorithms have been proposed, such as a machine learning (ML) based deep count autoencoder (*16*). Other tools focus on differential expression analysis, such as TASC (*17*), which uses a hierarchical mixture model of the biological variation. However, successful methods usually rely on assumptions about the biological experiment being tailored to a specific setting or model system, thus leaving most large-scale sequencing efforts (that lack 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 (noise), some of which end up included in the DE call and thus bias the biological interpretation. In addition, the choice of tools used for pre-processing steps may influence the output (and relative quantification accuracy) of gene expression (*18*). These analytical biases mainly arise from differences in the detection and handling of isoforms or processing of unmapped and multi-mapping reads (*3*). Such variation in abundance estimation in turn can strongly affect the downstream analyses (*19*).

We developed ***noisyR***, a denoising pipeline to quantify and exclude technical noise from downstream analyses, in a robust and data-driven way. Our noise-filtering method is applicable on either the original, un-normalised count matrix, or alignment data (BAM format) for a more refined analysis. Noise is quantified based either 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.

Results

Noise quantification in bulk RNA-seq data

To exemplify the impact of denoising on the biological interpretations from bulk RNA-seq experiments, we applied ***noisyR*** on mRNA-seq and smallRNA-seq (sRNA) 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 (*20*). To assess the distributions of signal we used density plots (Fig. 1A) and summaries of Jaccard similarity indices (Fig. 1B, JSIs) across all samples. For the former, we observed a multi-modal distribution that suggests a signal to noise transition range between [3,7] on log2 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 (Fig. 1C for replicate-versus-replicate similarity and the secondary DE distribution visible in Fig. 1D), we used first the noisyR count-based pipeline, on default parameters: window length = 10% x #genes and sliding step = 5% x window length (Fig. 1, E and H). 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. 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 (Fig. 1I). The parameters that minimised 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 plot (Fig. 1F). Next, we performed a DE analysis between the 0h and 12h 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 (*12*). Moreover, when comparing the outputs of two standard DE pipelines, edgeR (*8*) 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 (Fig. 1J); 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 (Fig. 1, D vs G). This step was followed by a functional enrichment analysis focusing on the DE genes, with the genes expressed (post filtering) as background set (*21*). The number of enriched terms was lower in the denoised data, 1,108 vs 4,671 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 denoising were also found prior to noise removal). In addition, the noise-correction terms corresponded to a higher percentage of genes assigned per pathway (Fig. 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 (*22*). Also, we observed a sharper increase to high correlation entries highlighting the transition from degraded transcripts to precisely excised sRNAs (*23, 24*). 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 (*25*); this characteristic is stronger for the animal case (fig. S1C). For both animals and plants the increasing distribution was clearly detectable (fig. S1, A and C). The TE distributions also reflected the characteristics of the underlying sRNAs; for the animal example (fig. S1D) we saw a sharper increase along the abundance bins, specific for the piRNAs (*26*), whereas in plants (fig. S1B), the distribution of signal (expressed siRNAs) mirrored the biogenesis of heterochromatin siRNAs (*27*).

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 et al (*28*); 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 in particular intractable.

First, we applied ***noisyR*** using the count matrix approach on all cells with default parameters; we observed that correlation values rise to a weakly positive plateau (0.2-0.4) and remain stable for a wide range of abundances (Fig. 2A). Our interpretation of this result is that lower sequencing depths and higher resolution of smart-seq compared to bulk data induces more dissimilarity for medium abundances. To alleviate this effect, we grouped cells into a small number of "pseudosamples", both randomly and using the structure of the experiment (grouping by donor). For each pseudosample, we averaged the expression of genes across cells and applied our count-based approach on the summarised matrix. In the resulting noisyR output, we observed a clearer step in the abundance-correlation plot (fig 2B), especially when the summary was performed by donor. This indicates that an effect of the summarisation is a reduction in cell-to-cell variability which also focuses the noise identification procedure. The thresholds obtained via pseudo-sample summarisation 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 summarised 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 on individual cells were highly variable, with several instances of low or negative correlations, observed even at high abundances (fig 2A), for the transcript approach, applied on the concatenated BAM files, we observed the expected increasing trend in the distribution of correlations (Fig. 2C). The correlation distributions were high, even at low abundances, which may be a consequence of the summarisation; 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 log2-scale.

To assess the impact of noisyR on the biological interpretation of results, we performed the same 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 normalised and the cells were clustered using the Seurat R package (see Methods). The different clusterings were visualised using the UMAP (non-linear) dimensionality reduction (*29*) (Fig. 2, D and E). We observed that for the raw data the cells cluster into 3 groups of 2 donors each, while in the denoised data cells corresponding to the four donors are mixed across clusters, suggesting the part of a putative initial bath effect might have been alleviated with the noise correction. We also observe 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 (Fig. 2F). We observe a good correspondence between the original and denoised matrix; in particular, clusters 1 and 4 largely merge into cluster 0, and cluster 0 remains intact and turns into cluster 1. While the total number of clusters remains the same (under default parameters), the partitioning of cells is altered, which led us to believe there may be a qualitatively different result between the original and denoised matrix with possible consequences for downstream biological interpretations. To evaluate the changes in interpretation, we compared the pre/post filtering clusters by identifying the (positive) markers and computing the JSI between the top 50 markers of each cluster (Fig. 2G). 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 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 (Fig. 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.

The results should describe the experiments performed and the findings observed. The results section should be divided into subsections to delineate different experimental themes. Subheadings should be descriptive phrases. All data must be shown either in the main text or in the Supplementary Materials.

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 (*30, 31*), or post-transcriptional, small RNAs (*32*)) 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 (*33*). The distribution of degraded fragments across transcripts observed the same distribution of correlation vs abundance as for the bulk RNAseq data (Fig. 3A). Using a correlation threshold of 0.25, we determined a signal/noise threshold of 60 for this dataset. We then matched the highly abundant reads to known miRNAs (Methods, Fig. 3B) and illustrated that by removing the noisy reads, with abundance less than the noise threshold (Fig. 3, C-D), the prediction of interactions is simplified (*34*) i.e. for most genes only a few peaks were left. In some cases (e.g. Fig. 3C), only a very clear peak was retained after the noise removal, while for other transcripts some secondary interactions were kept.

Effect on the inference and interpretation of Gene Regulatory Networks

However, characterising direct interactions between regulatory 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). 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 regulatory interactions from high-throughput sequencing data. After the network inference step, the topology of GRNs can be used as a proxy for capturing the underlying biological 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, 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 (Methods) on both bulk and single cell RNAseq datasets using non-noise-filtered original, non-noise-filtered normalised, and noise-filtered normalised 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, this averaging across cells and tissues usually buffers the noise effect on general patterns while reducing the possibility to detect weak but biologically meaningful expression signals (e.g. transcription factor or transposable element expression).

Using the Yang et al. dataset in four different setups (original, -F(iltered) -N(normalised); noise-filtered but not normalised, +F -N; not filtered but normalised, -F +N; and noise-filtered and normalised, +F +N) and subsampled into five distinct 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 (Methods). The inferred weighted correlation networks were imported into edgynode and rescaled to [0,100] to allow comparisons across inference tools.

Next, all rescaled weight matrices (fig. S2, A and B) were converted to binary format using the median value over the entire weight matrix as threshold to assign edge weights; a zero, if their weight was below the median value, and a one, if their 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 normalised datasets (fig. S3). 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 GRN topology. Fig. 3, G-I illustrate pairwise comparisons between all combinations of input datasets: 1) original -F -N; 2) not noise-filtered but normalised -F +N; 3) noise-filtered but not normalised +F -N; and 4) noise-filtered and normalised +F +N exemplified for 102 genes corresponding to the neuron differentiation pathway and shown for all three network inference tools (GENIE3, Figure 3G; GRNBoost2, Figure 3H; and PIDC Figure 3I). For all network inference tools, a common pattern is the refining effect of noise-filtering on the overall network topologies. Interestingly, the normalisation 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, (fig. S2, A and B) shows a comparison between rescaled weight matrix distributions for an original and a noise-filtered and normalised 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 binarisation 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 normalised data, further illustrating the advantages of noise-filtering to magnify biological signals by reducing technical noise (*34*).

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 high-throughput (sequencing) datasets. 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. The package is written in the R (version 4.0.3) programming language and is actively maintained on https://github.com/Core-Bioinformatics/noisyR.

For the sample-similarity calculation, two approaches are available. The **count matrix approach** uses the original, un-normalised 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 towards a high similarity across samples. Choosing from a collection of >40 similarity metrics (*35*), users can select a measure to assess the localised consistency in expression across samples (*12*). 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 distance 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 summarising 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.

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.

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 optimise the noise filtering procedure and dampen the 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 for 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 (Fig. 5A). Even for medium-high abundances, expression or rank inconsistencies characterise 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 (Fig. 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 optimising 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 (*20*).
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 (*35*), which provides a variety of >45 distance measures. Dissimilarity measures are being inverted for comparison purposes (Fig. 5, C-F illustrates the Spearman correlation, Euclidean distance, Kulbeck-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 minimise 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 minimise 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 (*18*). To assess the effect of different alignment approaches on the quantification and observed levels of noise, mRNA quantification using featureCounts was performed on reads aligned with STAR (*36*), HISAT2 (*37*) and Bowtie2 (*38*). 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 (Fig. 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 (Fig. 6B) illustrate little variation across aligners (fig. S6, A and B). The estimated signal/noise thresholds were also comparable between the datasets generated by different aligners (Fig. 6C), with transcripts-based noise results being less variable. Once the noise correction was applied, the substantial peak in the abundance distributions around zero (Fig. 6D) was removed or significantly diminished and a second peak corresponding to the true signal was revealed around log2(abundance) of five using both counts and transcripts based approaches (Fig. 6, E 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 DE genes (adjusted p-value < 0.05 and |log2(abn)| > 1) detected by the two methods were compared for outputs produced using STAR (Fig. 1J), Bowtie2 (Fig. 6G) and HISAT2 (Fig. 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 (fig. S6C).

Materials and Methods

Materials

The bulk mRNA-seq used to illustrate noisyR was generated by Yang et al (*20*). The dataset comprises 16 samples across 8 time points [0-72 hours 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 (*39*) for the plant dataset (2 samples, a wildtype and DCL1 knockdown, with 3 biological replicates each, in A. thaliana, GSM2412286 - GSM2412291) and from Wallach et al (*40*) for the animal dataset, 6 samples generated for the identification of microRNAs as TLR-activating molecules in M. musculus (PMID: 31940779, GSE138532, GSM4110737 - GSM4110742). For both datasets, the reads were aligned to mature and hairpin miRNAs, downloaded from miRBase (*41*) 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 3 biological replicates (GSE113958) presented in Thody et al (*33*).

The single-cell mRNA-seq dataset used to illustrate noisyR was generated by Cuomo et al (study of stem cell differentiation) (*28*). 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, all four time points 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 (*42*).

Methods, bulk mRNAseq data

Data pre-processing and quality checking

Initial quality checks were performed using fastQC (version 0.11.8), summarised with multiQC (version 1.9) (*43*). Alignments to reference genomes were performed using STAR (version 2.7.0a) with default parameters (*36*); the count matrices were generated using featureCounts (version 2.0.0) (*44*) 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 (*12, 45*).

Data post-processing and biological interpretation of results

The differential expression analysis was performed after quantile normalisation of the count matrix using the standard functions from edgeR, version 3.28.0 (*8*) and DESeq2, version 1.26.0 (*7*). The thresholds for DE were |log2(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) (*21*), against the standard GO terms, and the KEGG (*46*) and reactome (*47*) pathway databases. The observed set consisted of the DE genes, the background set comprised all expressed genes, using the full or de-noised 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) (*38*) and HISAT2 (version 2.1.0) (*37*). 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 |log2(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 1)

We chose a subset of the GRN inference tools benchmarked by BEELINE (*48*): GENIE3 (*49*), GRNBoost2 (*50*), and PIDC (*51*). 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, binarised 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 visualised.

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 1) no noise filter and no quantile normalisation (original) was performed (denoted in the figures as -F -N), 2) a noise filtering but no quantile normalisation was performed (+F -N), 3) no noise filtering but a quantile normalisation was performed (-F +N), and 4) 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 (ref), 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, visualising 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 normalisation/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 visualisations, including the ones used in this study can be found at https://github.com/drostlab/network-inference-pipeline.

Methods, sRNAseq data

The 6 A thaliana sRNA samples were assessed using multiQC version 1.9 (*43*). Next, the sequencing adapters (both standard and HD) were trimmed using Cutadapt (version 3.2) (*52*) and the UEA sRNA Workbench (*53*). The larger 3 samples were subsampled without replacement to 8M reads (*12*); the smaller 3 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) (*36*)) to both microRNA hairpins (miRBase Release 22.1) (*41*) and TEs (obtained from TAIR10) (*42*).

The 6 M musculus sRNA samples were processed in a similar way as the plant samples and subsampled without replacement to 3.5M sequences (*12*). 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) (*41*) and TEs (Ensembl release 101).

Methods, PARE data

The 3 A. thaliana PARE samples (GSE113958) were QCed (multiQC version 1.9) (*43*) and the reads trimmed to 20nt; next, all samples were randomly subsampled without replacement to 25M (*12*). The subsampled reads were aligned to the reference genome (obtained from TAIR10 (*42*)) using STAR (using STAR (version 2.7.0a) (*36*)), with default parameters. The reads aligned to each position along a transcript were grouped on sequence and summarised by frequency. Each summarised fragment was matched (as reverse complement) to A thaliana miRNAs. To visualise the distribution of signal across transcripts, t-plots were created, where each point corresponds to a summarised PARE fragment; the points for which a corresponding miRNA was identified were highlighted using the miRNA label (*33*).

Methods, single cell data

For the single cell SmartSeq2 data, the cellranger software version 3.0 (*54*) 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 (*55*). The UMAP reduction method (*29*) was used for visualisation and assessment of results.

Methods, noise quantification

Two approaches were implemented for the identification of noise. (1) The “count matrix approach” is a simple, fast way to obtain a threshold utilising solely the un-normalised count matrix (m genes x n samples). (2) The “transcript approach” is more refined, as it takes into account the distribution of signal across the transcript obtained by summarising 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 (*35*). Most results were obtained using Pearson Correlation Coefficient (by default); similar results are obtained with other similarity or inverted dissimilarity measures such as Spearman Correlation, Euclidean distance, Kulbeck-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 vs the average window abundance, x-axis. A correlation threshold (as a hyper-parameter) is used to determine a corresponding abundance threshold as a cut-off - the noise threshold. The correlation threshold is inferred from the data to minimise 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 (*56*); 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 above 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.

References

1. R. Stark, M. Grzelak, J. Hadfield, RNA sequencing: the teenage years. *Nature Reviews Genetics* **20**, 631-656 (2019).

2. A. Oshlack, M. D. Robinson, M. D. Young, From RNA-seq reads to differential expression results. *Genome Biology* **11**, 220 (2010).
3. A. Conesa *et al.*, A survey of best practices for RNA-seq data analysis. *Genome Biology* **17**, 13 (2016).
4. M. Li, J. C. I. Belmonte, Ground rules of the pluripotency gene regulatory network. *Nature Reviews Genetics* **18**, 180-191 (2017).
5. S. Parekh, C. Ziegenhain, B. Vieth, W. Enard, I. Hellmann, The impact of amplification on differential expression analyses by RNA-seq. *Scientific Reports* **6**, 25533 (2016).
6. K. D. Hansen, S. E. Brenner, S. Dudoit, Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Research* **38**, e131-e131 (2010).
7. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**, (2014).
8. D. J. McCarthy, Y. Chen, G. K. Smyth, Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research* **40**, 4288-4297 (2012).
9. T. Stuart, R. Satija, Integrative single-cell analysis. *Nature Reviews Genetics* **20**, 257-272 (2019).
10. F. Rapaport *et al.*, Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biology* **14**, R95 (2013).
11. Z. Wang, M. Gerstein, M. Snyder, RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* **10**, 57-63 (2009).
12. I. Mohorianu *et al.*, Comparison of alternative approaches for analysing multi-level RNA-seq data. *PLOS ONE* **12**, e0182694 (2017).
13. G. Park *et al.*, Characterization of background noise in capture-based targeted sequencing data. *Genome Biology* **18**, (2017).
14. I. Fischer-Hwang, I. Ochoa, T. Weissman, M. Hernaez, Denoising of Aligned Genomic Data. *Scientific Reports* **9**, (2019).
15. K. Shiroguchi, T. Z. Jia, P. A. Sims, X. S. Xie, Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. *Proceedings of the National Academy of Sciences* **109**, 1347-1352 (2012).
16. G. Eraslan, L. M. Simon, M. Mircea, N. S. Mueller, F. J. Theis, Single-cell RNA-seq denoising using a deep count autoencoder. *Nature Communications* **10**, (2019).
17. C. Jia *et al.*, Accounting for technical noise in differential expression analysis of single-cell RNA sequencing data. *Nucleic Acids Research* **45**, 10978-10988 (2017).
18. A. Srivastava *et al.*, Alignment and mapping methodology influence transcript abundance estimation. *Genome Biology* **21**, (2020).
19. L. A. Corchete *et al.*, Systematic comparison and assessment of RNA-seq procedures for gene expression quantitative analysis. *Scientific Reports* **10**, (2020).
20. P. Yang *et al.*, Multi-omic Profiling Reveals Dynamics of the Phased Progression of Pluripotency. *Cell Systems* **8**, 427-445.e410 (2019).
21. U. Raudvere *et al.*, g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Research* **47**, W191-W198 (2019).
22. I. Mohorianu, M. B. Stocks, J. Wood, T. Dalmay, V. Moulton, CoLIde. *RNA Biology* **10**, 1221-1230 (2013).
23. V. N. Kim, J. Han, M. C. Siomi, Biogenesis of small RNAs in animals. *Nature Reviews Molecular Cell Biology* **10**, 126-139 (2009).
24. F. Borges, R. A. Martienssen, The expanding world of small RNAs in plants. *Nature Reviews Molecular Cell Biology* **16**, 727-741 (2015).
25. M. Ha, V. N. Kim, Regulation of microRNA biogenesis. *Nature Reviews Molecular Cell Biology* **15**, 509-524 (2014).

26. B. Czech *et al.*, piRNA-Guided Genome Defense: From Biogenesis to Silencing. *Annual Review of Genetics* **52**, 131-157 (2018).

27. R. K. Papareddy *et al.*, Chromatin regulates expression of small RNAs to help maintain transposon methylome homeostasis in Arabidopsis. *Genome Biology* **21**, (2020).

28. A. S. E. Cuomo *et al.*, Single-cell RNA-sequencing of differentiating iPS cells reveals dynamic genetic effects on gene expression. *Nature Communications* **11**, (2020).

29. E. Becht *et al.*, Dimensionality reduction for visualizing single-cell data using UMAP. *Nature Biotechnology* **37**, 38-44 (2019).

30. R. Andersson, A. Sandelin, Determinants of enhancer and promoter activities of regulatory elements. *Nature Reviews Genetics* **21**, 71-87 (2020).

31. M. Levo, E. Segal, In pursuit of design principles of regulatory sequences. *Nature Reviews Genetics* **15**, 453-468 (2014).

32. D. Holoch, D. Moazed, RNA-mediated epigenetic regulation of gene expression. *Nature Reviews Genetics* **16**, 71-84 (2015).

33. J. Thody, V. Moulton, I. Mohorianu, PAREameters: a tool for computational inference of plant miRNA–mRNA targeting rules using small RNA and degradome sequencing data. *Nucleic Acids Research* **48**, 2258-2270 (2020).

34. J. Thody *et al.*, PAREsnip2: a tool for high-throughput prediction of small RNA targets from degradome sequencing data using configurable targeting rules. *Nucleic Acids Research*, (2018).

35. H.-G. Drost, Philentropy: Information Theory and Distance Quantification with R. *Journal of Open Source Software* **3**, 765 (2018).

36. A. Dobin *et al.*, STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).

37. D. Kim, J. M. Paggi, C. Park, C. Bennett, S. L. Salzberg, Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology* **37**, 907-915 (2019).

38. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357-359 (2012).

39. C. Paicu *et al.*, miRCat2: accurate prediction of plant and animal microRNAs from next-generation sequencing datasets. *Bioinformatics* **33**, 2446-2454 (2017).

40. T. Wallach *et al.*, Identification of CNS Injury-Related microRNAs as Novel Toll-Like Receptor 7/8 Signaling Activators by Small RNA Sequencing. *Cells* **9**, 186 (2020).

41. A. Kozomara, M. Birgaoanu, S. Griffiths-Jones, miRBase: from microRNA sequences to function. *Nucleic Acids Research* **47**, D155-D162 (2019).

42. T. Z. Berardini *et al.*, The arabidopsis information resource: Making and mining the “gold standard” annotated reference plant genome. *genesis* **53**, 474-485 (2015).

43. P. Ewels, M. Magnusson, S. Lundin, M. Käller, MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047-3048 (2016).

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

45. I. T. Jolliffe, J. Cadima, Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **374**, 20150202 (2016).

46. M. Kanehisa, KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* **28**, 27-30 (2000).

47. G. Viteri *et al.*, Reactome and ORCID—fine-grained credit attribution for community curation. *Database* **2019**, (2019).

48. A. Pratapa, A. P. Jalihal, J. N. Law, A. Bharadwaj, T. M. Murali, Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature Methods* **17**, 147-154 (2020).
49. V. A. Huynh-Thu, A. Irrthum, L. Wehenkel, P. Geurts, Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. *PLoS ONE* **5**, e12776 (2010).
50. T. Moerman *et al.*, GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks. *Bioinformatics* **35**, 2159-2161 (2019).
51. T. E. Chan, M. P. H. Stumpf, A. C. Babtie, Gene Regulatory Network Inference from Single-Cell Data Using Multivariate Information Measures. *Cell Systems* **5**, 251-267.e253 (2017).
52. M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10 (2011).
53. M. B. Stocks *et al.*, The UEA sRNA workbench: a suite of tools for analysing and visualizing next generation sequencing microRNA and small RNA datasets. *Bioinformatics* **28**, 2059-2061 (2012).
54. G. X. Y. Zheng *et al.*, Massively parallel digital transcriptional profiling of single cells. *Nature Communications* **8**, 14049 (2017).
55. Y. Hao *et al.* (Cold Spring Harbor Laboratory, 2020).
56. W. J. Kent *et al.*, The Human Genome Browser at UCSC. *Genome Research* **12**, 996-1006 (2002).

Acknowledgments

IM, EL, EW and IIM acknowledge the constructive feedback from the Core Bioinformatics group and the support by the Core grant awarded to the Wellcome-MRC Cambridge Stem Cells Institute. SAVU, LM and HGD acknowledge that their work was supported by the Max Planck Society. IM and IIM designed the study and implemented the R package noisyR; SAVU, LM and HGD implemented the R package edgyNode, the analyses were performed by IM, IIM (the mRNA bulk and single cell analyses), EW, IIM (sRNA analysis), LM and HGD (GRN inference), EL, IM, IIM (comparison across tools). IM, HGD and IIM wrote the manuscript; all authors read and approved the submitted manuscript.

Figures and Tables

Figure 1 Overview of QC measures and original vs 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, normalised count matrix; the colour gradient is proportional with 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 de-noised abundances for the two 12h 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 Processed, 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.

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 “pseudosamples” 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).

Fig. 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 of the microRNA/mRNA interaction; cleavage of the mRNA transcript occurs between the 10th and 11th nucleotide of the microRNA; the 5’ fragment of the mRNA degrades, while the 3’ fragment is sequenced; sequencing outputs are compared with known microRNAs for *A. thaliana*. **(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)** For the bulk RNAseq (Yang et al.) dataset the node degree distributions (total number of edges connected to a node/gene) of 102 genes assigned to the neuron differentiation pathway are shown (Supplementary Data XY). All four input data variants are shown: original (-F -N, purple); not noise-filtered but normalised (-F +N, green); noise-filtered but not normalised (+F -N, red); and noise-filtered and normalised (+F +N, blue) sorted by increasing values using -F -N as sorting key. **(F)** Analogous node degree distribution plot to (E) for the single-cell (Cuomo et al.) dataset showing 133 genes associated with catalytic activity pathways (Supplementary Data XY). **(G-L)** Pairwise hamming distance comparisons for each gene between all combinations of original (-F -N), noise-filtered (+F), and normalised (+N) input datasets using 102 Neuron differentiation genes from the bulk RNAseq (Yang et al.) dataset and 133 genes associated with catalytic activity pathways (Methods) show a comparable pattern across different gene regulatory network inference tools: **(G)** GENIE3, 102 genes / Yang; **(H)** GRNBoost2, 102 genes / Yang; **(I)** PIDC, 102 genes / Yang; **(J)** GENIE3, 133 genes / Cuomo; **(K)** GRNBoost2, 133 genes / Cuomo; **(L)** PIDC, 133 genes / Cuomo. The results consistently show that across network inference tools and bulk vs single-cell data noise-filtering has only refining effects on the inferred network topologies in original or normalised data, further illustrating the advantages of noise-filtering to magnify biological signals by reducing technical noise.

Figure 4 Workflow diagram of the noisyR pipeline.

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 gray.

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 ~5,500 **(D)** Inverse of the Euclidean distance plotted against abundance for the default window length of ~5,500 **(E)** Inverse of the Kulbeck-Leibler divergence plotted against abundance for the default window length of ~5,500 **(F)** Inverse of the Jensen-Shannon divergence plotted against abundance for the default window length of ~5,500

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 counts-based noise removal approach **(B)** The distribution of PCC across abundance bins in aligned reads counts obtained by the five aligners for the same sample in 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 1

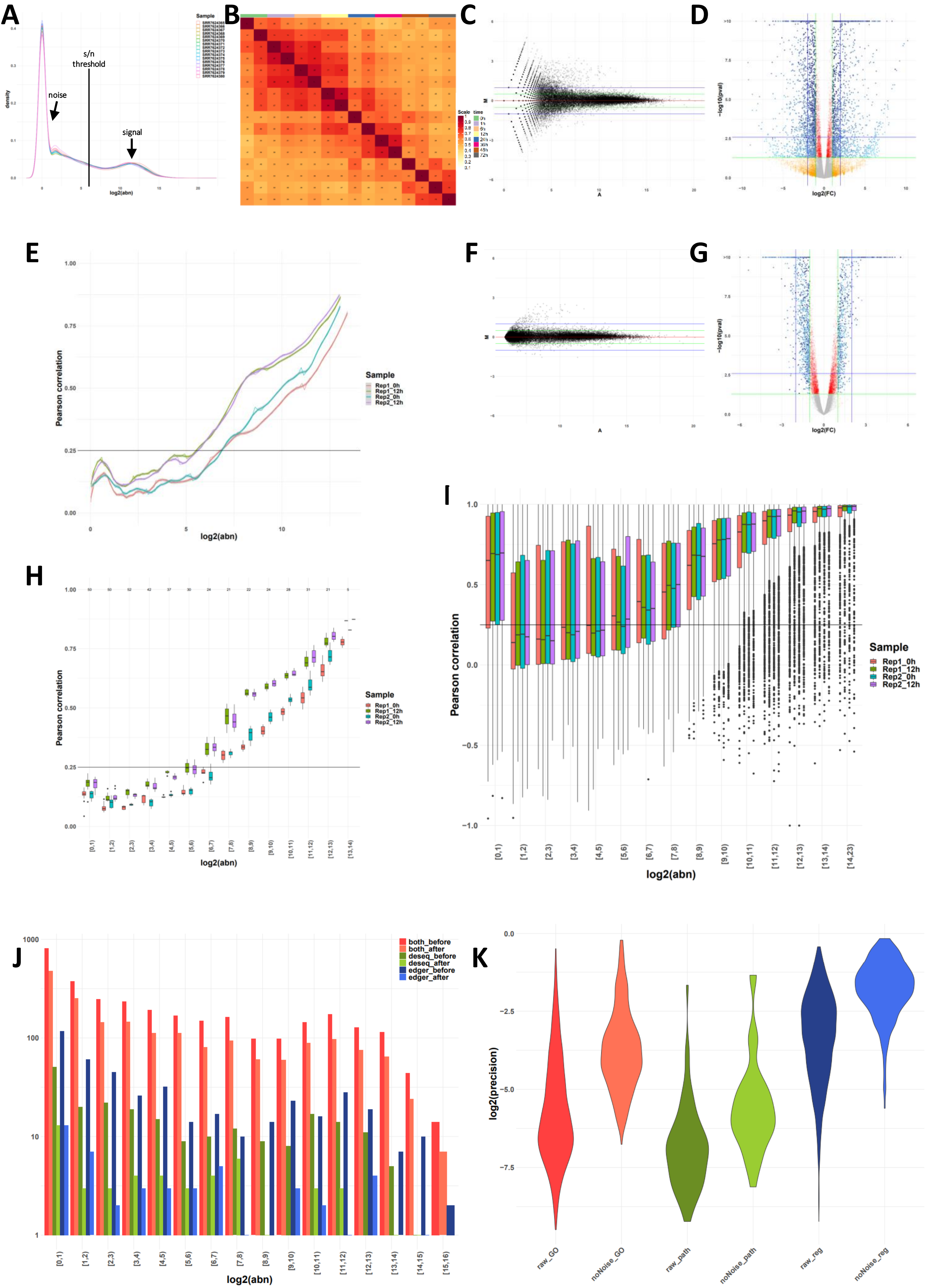

A
density
s/n
threshold
noise
signal
log2(abn)
Sample
B
Scale
time
0h
1h
6h
12h
24h
36h
48h
72h
C
M
A
D
>10
-log10(pval)
log2(FC)
E
Pearson correlation
log2(abn)
Sample
Rep1_0h
Rep1_12h
Rep2_0h
Rep2_12h
F
M
A
G
-log10(pval)
log2(FC)
H
Pearson correlation
log2(abn)
Sample
Rep1_0h
Rep1_12h
Rep2_0h
Rep2_12h
I
Pearson correlation
log2(abn)
Sample
Rep1_0h
Rep1_12h
Rep2_0h
Rep2_12h
J
both_before
both_after
deseq_before
deseq_after
edger_before
edger_after
log2(abn)
K
log2(precision)
raw_GO
noNoise_GO
raw_path
noNoise_path
raw_reg
noNoise_reg

Figure 2

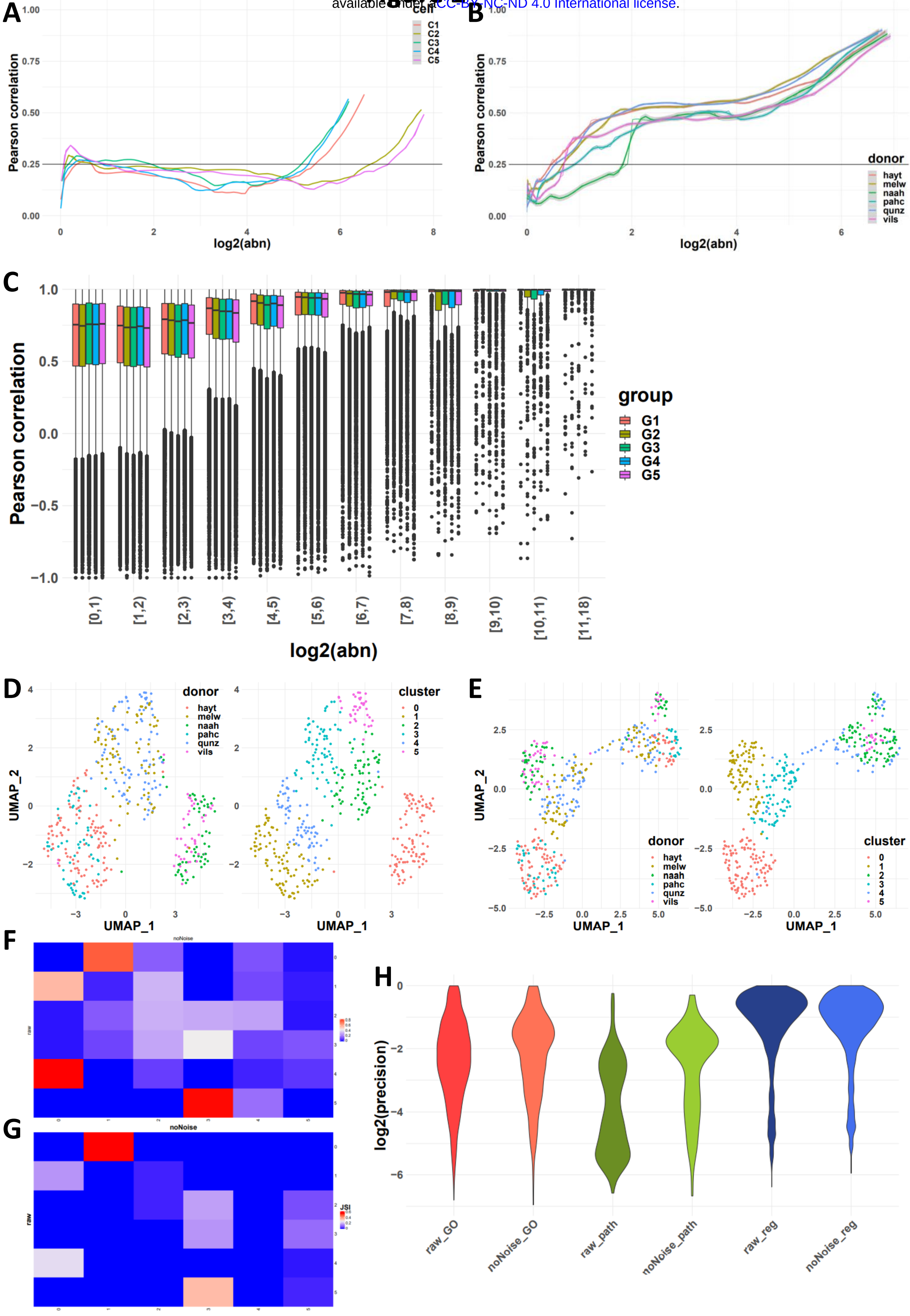

A
cell
C1
C2
C3
C4
C5
Pearson correlation
1.00
0.75
0.50
0.25
0.00
log2(abn)
0
2
4
6
8
B
donor
hayt
melw
naah
pahc
qunz
vils
Pearson correlation
1.00
0.75
0.50
0.25
0.00
log2(abn)
0
2
4
6
C
Pearson correlation
1.0
0.5
0.0
−0.5
−1.0
[0,1)
[1,2)
[2,3)
[3,4)
[4,5)
[5,6)
[6,7)
[7,8)
[8,9)
[9,10)
[10,11)
[11,18)
log2(abn)
group
G1
G2
G3
G4
G5
D
donor
hayt
melw
naah
pahc
qunz
vils
cluster
0
1
2
3
4
5
UMAP_2
UMAP_1
4
2
0
-2
-3
0
3
E
donor
hayt
melw
naah
pahc
qunz
vils
cluster
0
1
2
3
4
5
UMAP_2
UMAP_1
2.5
0.0
-2.5
-5.0
-2.5
0.0
2.5
5.0
F
noNoise
raw
0
1
2
3
4
5
G
noNoise
raw
JSI
0
1
2
3
4
5
H
log2(precision)
0
-2
-4
-6
raw_GO
noNoise_GO
raw_path
noNoise_path
raw_reg
noNoise_reg

Figure 3

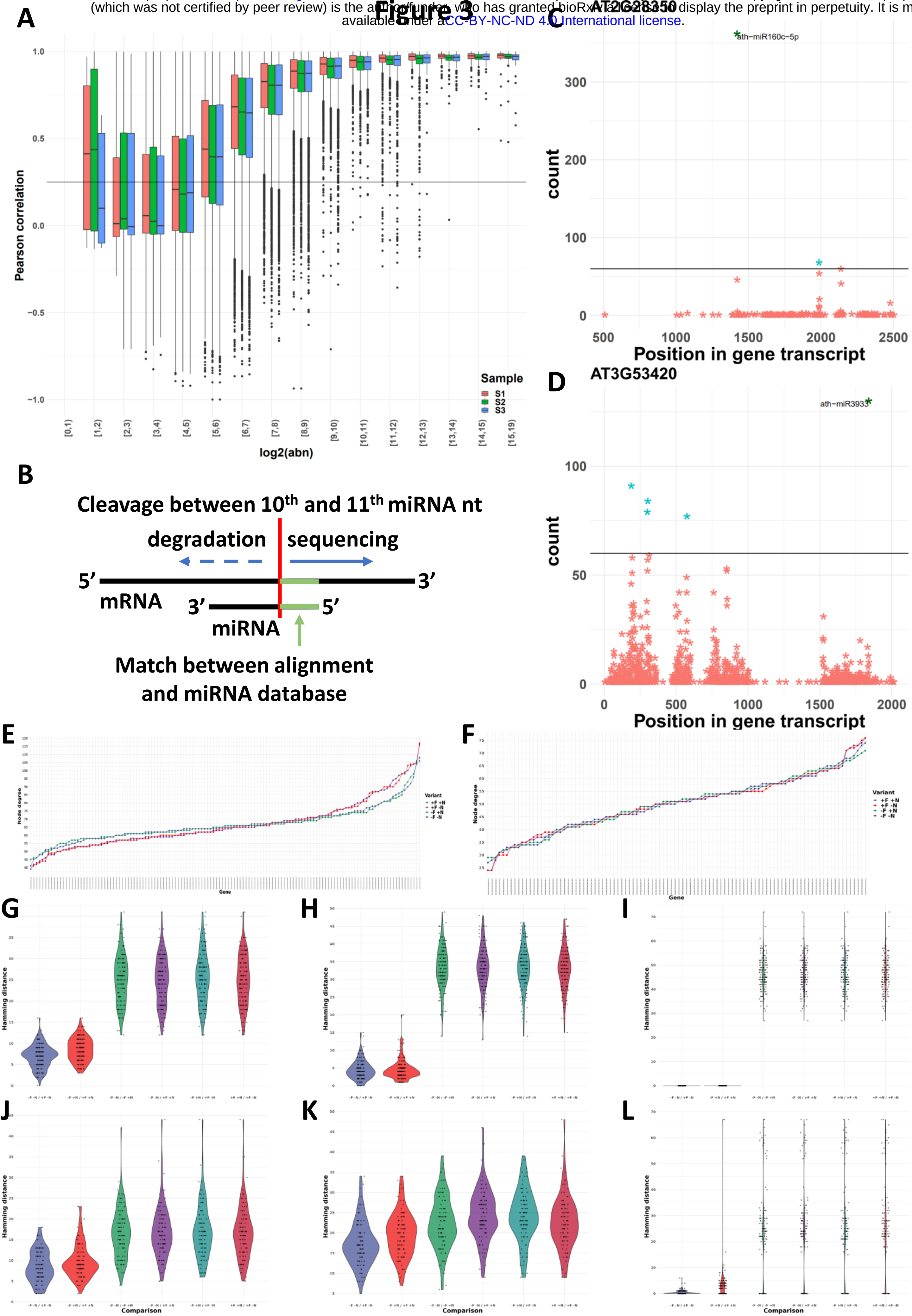

A
Pearson correlation
1.0
0.5
0.0
−0.5
−1.0
[0,1)
[1,2)
[2,3)
[3,4)
[4,5)
[5,6)
[6,7)
[7,8)
[8,9)
[9,10)
[10,11)
[11,12)
[12,13)
[13,14)
[14,15)
[15,19)
log2(abn)
Sample
S1
S2
S3
B
Cleavage between 10th and 11th miRNA nt
degradation
sequencing
5'
3'
mRNA
3'
5'
miRNA
Match between alignment and miRNA database
C
AT2G28350
ath-miR160c-5p
count
300
200
100
0
500
1000
1500
2000
2500
Position in gene transcript
D
AT3G53420
ath-miR3933
count
100
50
0
0
500
1000
1500
2000
Position in gene transcript
E
Node degree
Gene
Variant
+F +N
+F -N
-F +N
-F -N
F
Node degree
Gene
Variant
+F +N
+F -N
-F +N
-F -N
G
Hamming distance
H
Hamming distance
I
Hamming distance
J
Hamming distance
Comparison
K
Hamming distance
Comparison
L
Hamming distance
Comparison
-F -N / +F -N
-F +N / +F +N
-F -N / -F +N
-F -N / +F +N
-F +N / +F -N
+F +N / +F -N

Figure 4

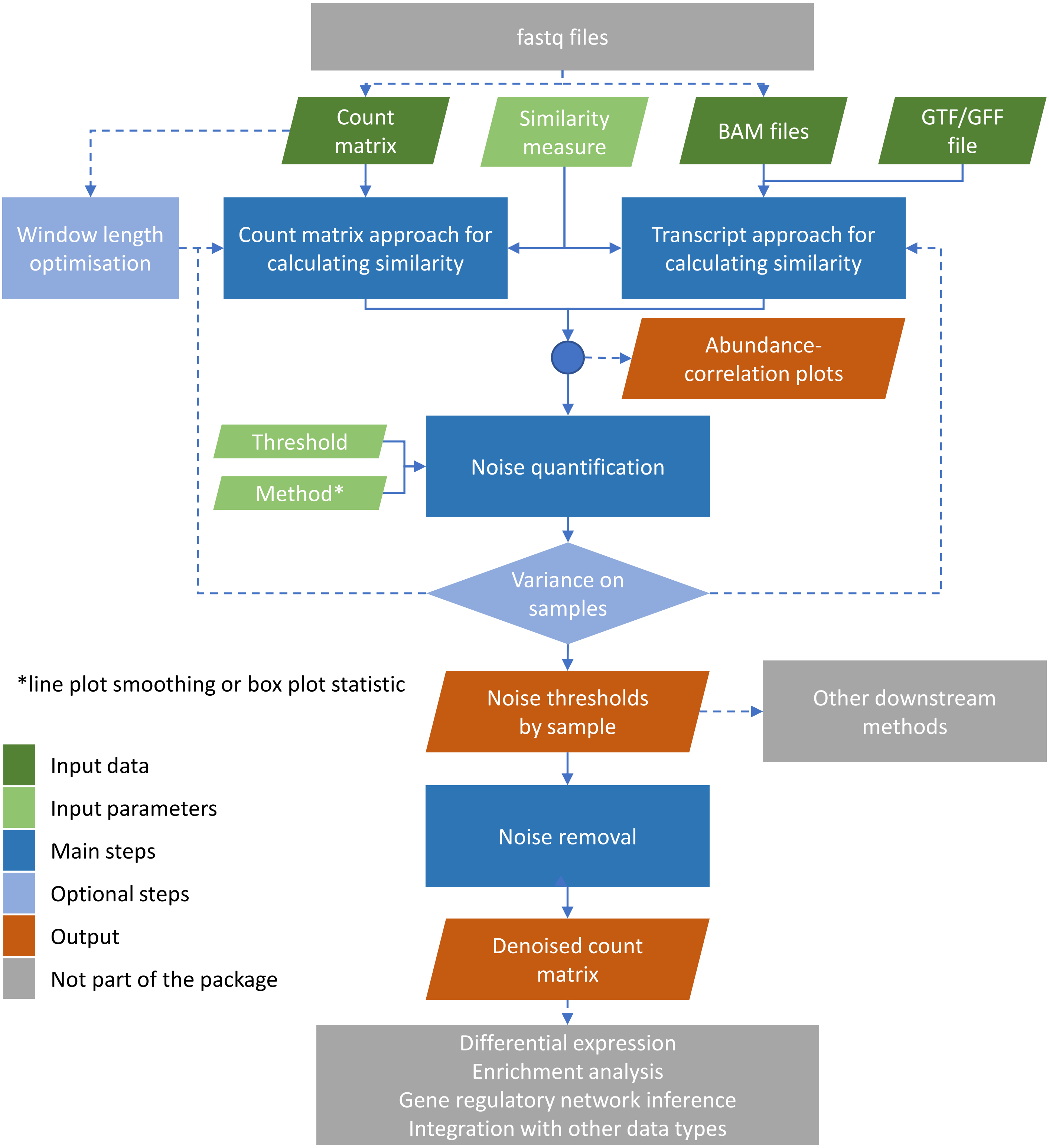

fastq files
Count matrix
Similarity measure
BAM files
GTF/GFF file
Window length optimisation
Count matrix approach for calculating similarity
Transcript approach for calculating similarity
Abundance-correlation plots
Threshold
Method*
Noise quantification
Variance on samples
*line plot smoothing or box plot statistic
Noise thresholds by sample
Other downstream methods
Noise removal
Denoised count matrix
Differential expression
Enrichment analysis
Gene regulatory network inference
Integration with other data types
Input data
Input parameters
Main steps
Optional steps
Output
Not part of the package

Figure 5

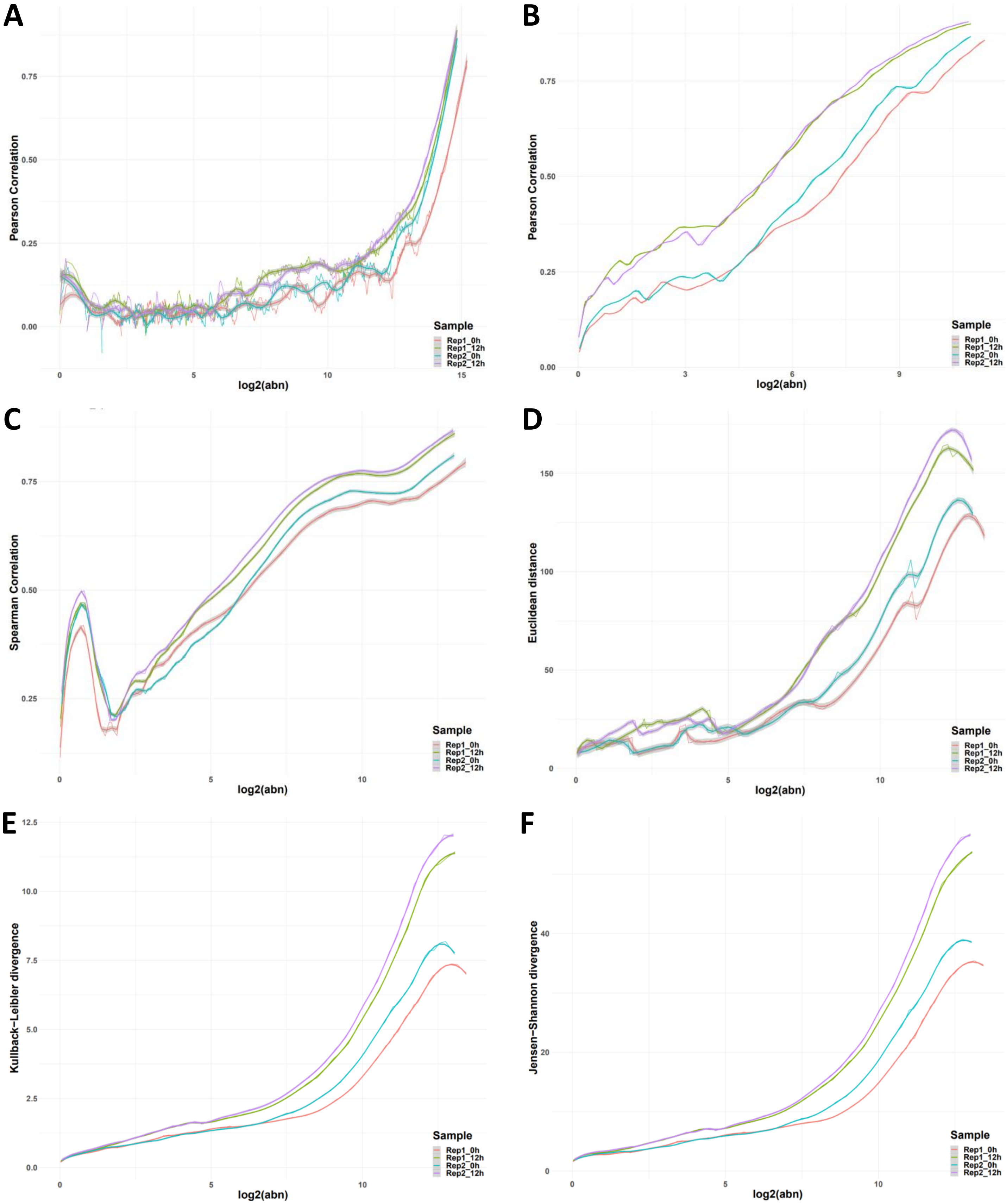
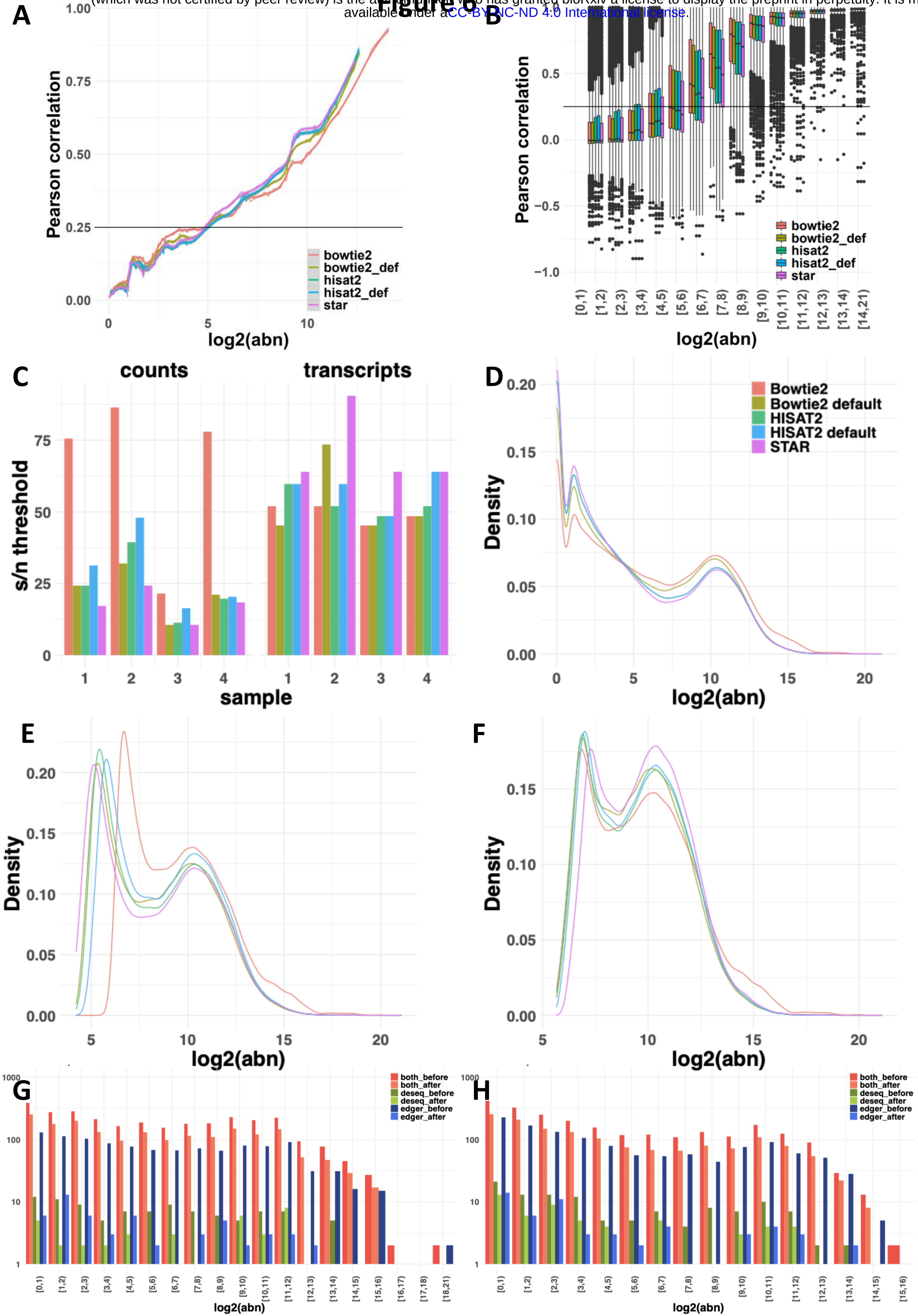

A
Pearson Correlation
0.00
0.25
0.50
0.75
log2(abn)
0
5
10
15
Sample
Rep1_0h
Rep1_12h
Rep2_0h
Rep2_12h
B
Pearson Correlation
0.00
0.25
0.50
0.75
log2(abn)
0
3
6
9
Sample
Rep1_0h
Rep1_12h
Rep2_0h
Rep2_12h
C
Spearman Correlation
0.25
0.50
0.75
log2(abn)
0
5
10
Sample
Rep1_0h
Rep1_12h
Rep2_0h
Rep2_12h
D
Euclidean distance
0
50
100
150
log2(abn)
0
5
10
Sample
Rep1_0h
Rep1_12h
Rep2_0h
Rep2_12h
E
Kullback−Leibler divergence
0.0
2.5
5.0
7.5
10.0
12.5
log2(abn)
0
5
10
Sample
Rep1_0h
Rep1_12h
Rep2_0h
Rep2_12h
F
Jensen−Shannon divergence
0
20
40
log2(abn)
0
5
10
Sample
Rep1_0h
Rep1_12h
Rep2_0h
Rep2_12h

Figure 6


A
Pearson correlation
1.00
0.75
0.50
0.25
0.00
0
5
10
log2(abn)
bowtie2
bowtie2_def
hisat2
hisat2_def
star
B
Pearson correlation
0.5
0.0
−0.5
−1.0
[0,1)
[1,2)
[2,3)
[3,4)
[4,5)
[5,6)
[6,7)
[7,8)
[8,9)
[9,10)
[10,11)
[11,12)
[12,13)
[13,14)
[14,21)
log2(abn)
bowtie2
bowtie2_def
hisat2
hisat2_def
star
C
counts
transcripts
s/n threshold
75
50
25
0
1
2
3
4
sample
D
Density
0.20
0.15
0.10
0.05
0.00
0
5
10
15
20
log2(abn)
Bowtie2
Bowtie2 default
HISAT2
HISAT2 default
STAR
E
Density
0.20
0.15
0.10
0.05
0.00
5
10
15
20
log2(abn)
F
Density
0.15
0.10
0.05
0.00
5
10
15
20
log2(abn)
G
1000
100
10
1
both_before
both_after
deseq_before
deseq_after
edger_before
edger_after
[0,1)
[1,2)
[2,3)
[3,4)
[4,5)
[5,6)
[6,7)
[7,8)
[8,9)
[9,10)
[10,11)
[11,12)
[12,13)
[13,14)
[14,15)
[15,16)
[16,17)
[17,18)
[18,21)
log2(abn)
H
1000
100
10
1
both_before
both_after
deseq_before
deseq_after
edger_before
edger_after
[0,1)
[1,2)
[2,3)
[3,4)
[4,5)
[5,6)
[6,7)
[7,8)
[8,9)
[9,10)
[10,11)
[11,12)
[12,13)
[13,14)
[14,15)
[15,16)
log2(abn)