



University of Dundee

Detection and Mitigation of Spurious Antisense RNA-seq Reads with RoSA

Mourao, Kira; Schurch, Nicholas; Lukoszek, Radoslaw; Froussios, Kimon; Mackinnon, Katarzyna; Duc, Celine

Published in:
F1000 Research

DOI:
[10.1101/425900](https://doi.org/10.1101/425900)
[10.12688/f1000research.18952.1](https://doi.org/10.12688/f1000research.18952.1)

Publication date:
2019

Document Version
Publisher's PDF, also known as Version of record

[Link to publication in Discovery Research Portal](#)

Citation for published version (APA):

Mourao, K., Schurch, N., Lukoszek, R., Froussios, K., Mackinnon, K., Duc, C., ... Barton, G. (2019). Detection and Mitigation of Spurious Antisense RNA-seq Reads with RoSA. *F1000 Research*, 8, [819].
<https://doi.org/10.1101/425900>, <https://doi.org/10.12688/f1000research.18952.1>

General rights

Copyright and moral rights for the publications made accessible in Discovery Research Portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from Discovery Research Portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain.
- You may freely distribute the URL identifying the publication in the public portal.


Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.



SOFTWARE TOOL ARTICLE

Detection and mitigation of spurious antisense expression with RoSA [version 1; peer review: awaiting peer review]

Kira Mourão¹, Nicholas J. Schurch², Radek Lucoszek³, Kimon Froussios⁴,
Katarzyna MacKinnon⁵, Céline Duc⁶, Gordon Simpson³, Geoffrey J. Barton ⁷

¹Synpromics Ltd, Edinburgh, Midlothian, EH25 9RG, UK

²Biomathematics and Statistics Scotland, James Hutton Institute, Aberdeen, Scotland, AB15 8QH, UK

³Centre for Gene Regulation & Expression, School of Life Sciences, University of Dundee, Dundee, Scotland, DD1 5EH, UK

⁴Research Institute of Molecular Pathology, Vienna, 1030, Austria

⁵Cell & Molecular Sciences, James Hutton Institute, Invergowrie, Dundee, Scotland, DD2 5DA, UK

⁶Équipe Épигénétique, Unité Fonctionnalité et Ingénierie des Protéines (UFIP) Faculté des Sciences et Techniques, Université de Nantes, NANTES, 92208 F44322 CEDEX 3, France

⁷Computational Biology, School of Life Sciences, University of Dundee, Dundee, Scotland, DD1 5EH, UK

V1 First published: 07 Jun 2019, 8:819 (
<https://doi.org/10.12688/f1000research.18952.1>)

Latest published: 07 Jun 2019, 8:819 (
<https://doi.org/10.12688/f1000research.18952.1>)

Abstract

Antisense transcription is known to have a range of impacts on sense gene expression, including (but not limited to) impeding transcription initiation, disrupting post-transcriptional processes, and enhancing, slowing, or even preventing transcription of the sense gene. Strand-specific RNA-Seq protocols preserve the strand information of the original RNA in the data, and so can be used to identify where antisense transcription may be implicated in regulating gene expression. However, our analysis of 199 strand-specific RNA-Seq experiments reveals that spurious antisense reads are often present in these datasets at levels greater than 1% of sense gene expression levels. Furthermore, these levels can vary substantially even between replicates in the same experiment, potentially disrupting any downstream analysis, if the incorrectly assigned antisense counts dominate the set of genes with high antisense transcription levels. Currently, no tools exist to detect or correct for this spurious antisense signal. Our tool, RoSA (Removal of Spurious Antisense), detects the presence of high levels of spurious antisense read alignments in strand-specific RNA-Seq datasets. It uses incorrectly spliced reads on the antisense strand and/or ERCC spikeins (if present in the data) to calculate both global and gene-specific antisense correction factors. We demonstrate the utility of our tool to filter out spurious antisense transcript counts in an *Arabidopsis thaliana* RNA-Seq experiment.

Availability: RoSA is open source software available under the GPL licence via the Barton Group GitHub page <https://github.com/bartongroup>.

Keywords

RNA-seq, antisense expression, gene expression, *Arabidopsis thaliana*, ENCODE

Open Peer Review

Reviewer Status AWAITING PEER REVIEW

Any reports and responses or comments on the article can be found at the end of the article.



This article is included in the **RPackage** gateway.



This article is included in the **Python Collection** collection.

Corresponding authors: Gordon Simpson (G.G.Simpson@dundee.ac.uk), Geoffrey J. Barton (g.j.barton@dundee.ac.uk)

Author roles: **Mourão K:** Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Schurch NJ:** Conceptualization, Investigation, Methodology, Project Administration, Software, Supervision, Writing – Review & Editing; **Lucoszek R:** Data Curation, Investigation, Writing – Review & Editing; **Froussios K:** Conceptualization, Methodology, Software, Writing – Review & Editing; **MacKinnon K:** Data Curation, Investigation, Writing – Review & Editing; **Duc C:** Data Curation, Writing – Review & Editing; **Simpson G:** Funding Acquisition, Writing – Review & Editing; **Barton GJ:** Funding Acquisition, Project Administration, Resources, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work has been supported by the Biotechnology and Biological Sciences Research Council [BB/M004155/1, BB/M010066/1] to G.J.B. and G.G.S.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2019 Mourão K *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution Licence](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Mourão K, Schurch NJ, Lucoszek R *et al.* **Detection and mitigation of spurious antisense expression with RoSA [version 1; peer review: awaiting peer review]** F1000Research 2019, 8:819 (<https://doi.org/10.12688/f1000research.18952.1>)

First published: 07 Jun 2019, 8:819 (<https://doi.org/10.12688/f1000research.18952.1>)

1. Introduction

Antisense RNAs are transcribed from the strand opposite to that of the sense transcript of either protein-coding or non-protein-coding genes. They appear to be widespread in all kingdoms of life and can play distinct roles in regulating gene expression or function. Typically, antisense RNAs are non-coding and expressed at lower levels than sense gene transcripts. However, they can exhibit a range of sizes, and may or may not have 5' cap or 3' poly(A) tails depending on whether they arise from either their own promoters, from divergent promoters, or from copying of sense transcripts by RNA-dependent RNA polymerases (see 1 and references therein,²⁻⁴). In *Arabidopsis thaliana*, for example, the transcription of the Flowering Locus C (FLC) gene is known to be affected by transcription of antisense ncRNAs: COOLAIR^{5,6}, a set of ncRNAs antisense to FLC, and COLDAIR⁷, antisense to COOLAIR. Both COLDAIR and COOLAIR are associated with different changes in sense strand gene expression at the FLC locus⁸. Antisense transcription is known to affect sense gene expression through multiple mechanisms¹. During transcription, RNA polymerases may physically interfere with each other if both sense and antisense transcription take place simultaneously. Interference can prevent or slow down transcription (e.g. through RNA polymerase collisions^{9,10}) or force particular isoforms to be produced preferentially¹¹. Post-transcriptionally, antisense transcripts can compete with sense transcripts for binding sites¹². For example, the transcription of the human haemoglobin gene HBA1 is affected when the LUC7L gene on the opposite strand does not terminate, due to a deletion. It produces an antisense transcript that overlaps with HBA1, and which methylates the HBA1 promoter, repressing its expression¹³. In addition, since regions of protein coding genes on opposite DNA strands can overlap, their expression effectively generates transcripts that are, to varying extents, antisense to each other. Such overlapping gene pairs are a common feature of genome organization. We and others have shown that in some eukaryotic genomes tail-tail overlap enables the use of pre-mRNA 3' processing signals in different registers for genes coded on either strand¹⁴.

Incorporating antisense RNAs into genome annotation and properly quantifying their expression patterns is thus crucial, but remains challenging. Transcriptome-wide identification of RNAs is currently dominated by RNA-Seq. In this widespread experimental approach RNA is rarely sequenced directly, but instead is fragmented and first copied into cDNA and then copied again, so that libraries of DNA are sequenced. However, the copying of RNA by viral-derived reverse transcriptases is problematic. First, these polymerases exhibit DNA dependent polymerase activity, which can result in copies of the cDNA that can be incorrectly interpreted as antisense transcription. Second, just as reverse transcriptases switch template strand in viral biology, they can similarly switch templates in RNA-Seq library preparation, resulting again, in the interpretation of non-authentic antisense RNAs¹⁵⁻²¹. Historically, in microarray and RT-PCR experiments, this step is known to assign some transcripts to the wrong strand, creating spurious antisense transcripts. Preparing samples with actinomycin D can help to reduce the number of spurious antisense transcripts¹⁷ but can have unwanted side effects²⁰. Alternative approaches to make strand-specific

RNA-Seq libraries have been developed to mitigate artefacts arising from reverse transcription, however most of these also use reverse transcription²² and so have similar problems with incorrect assignment. For example, the highly-rated^{22,23} and widely used dUTP protocol for stranded RNA-Seq²⁴ is known to generate low levels of spurious antisense reads ranging from 0.6–3% of the sense signal^{22,25,26}. Ultimately, the direct sequencing of full-length RNA molecules²⁷ will overcome many of the problems of distinguishing authentic antisense RNAs. However, currently, reverse-transcriptase based approaches dominate and the extent of spurious antisense RNAs identified in RNA-Seq datasets is rarely exposed.

In this paper, we analyse spurious antisense reads in 199 RNA-Seq experiments, across multiple organisms from both ENCODE²⁸ and our own work. Our results show that spurious antisense reads are often present in experiments, and can manifest at levels greater than 1% of sense transcript levels. Furthermore, the number of spurious antisense reads can vary substantially between replicates within the same experiment. In some cases, this variation may be sufficient to disrupt downstream analysis of antisense gene expression, by causing spurious antisense counts to dominate the set of genes with high antisense transcription levels.

To detect and correct for wrongly assigned reads we developed a tool, RoSA (Removal of Spurious Antisense), which calculates an antisense correction factor by identifying subsets of reads where all antisense reads are spurious. We evaluate the effect of using RoSA on *Arabidopsis thaliana* experimental data where varying levels of spurious antisense were present in different replicates. RoSA reduces the overall dependence of antisense counts on sense counts, a key indicator of the presence of spurious antisense. For individual genes with different real and spurious antisense characteristics, RoSA reduces spurious antisense counts while retaining the antisense signal.

2. Methods

As noted by Jiang *et al.* (2011,²⁵), spurious antisense read counts can be estimated by analysing either ERCC spike-in data or counts of sense and antisense reads around splice sites. Each approach has different advantages: using spike-ins is simpler and faster, while using spliced reads allows a gene-by-gene estimate to be made. RoSA implements both approaches, in conjunction with pre-processing scripts to generate specialised read counts required by the tool. Once RoSA has an estimate of the levels of spurious antisense, it can adjust the raw antisense counts to account for the incorrectly stranded reads.

2.1 RoSA: Removal of Spurious Antisense

Our scripts and analysis code are bundled as a tool, RoSA (Removal of Spurious Antisense), available from the Barton Group's github pages at <https://github.com/bartongroup/RoSA>. RoSA is an R package supported by two python pre-processing scripts, callable from R.

For genes with spliced transcripts which are expressed in the data, RoSA uses the subset of reads from either strand that map across the splice junctions. The antisense reads in this subset are almost certainly spurious, and so RoSA can use the read

counts to calculate a gene-specific antisense correction factor (Section 2.2). For genes without spliced transcripts, RoSA uses ERCC spike-in data, if present. Here any antisense read mappings are, by definition, spurious and the ratio of sense to antisense reads mapping to the spike-ins thus provides a global, rather than gene-specific, antisense correction factor (Section 2.3). If ERCC spike-in data is not available, RoSA instead calculates a global estimate of the spurious antisense fraction from the set of spliced reads. Counting all, or spliced-only, antisense reads is not directly supported by existing tools. RoSA's pre-processing scripts perform these functions. The *make_annotation* script creates an antisense annotation (as gtf) from a standard annotation (as gff or gtf), which can then be used to generate antisense read counts via a standard read counting tool (Section 2.4.1). RoSA doesn't specify how the sense and antisense gene expression is counted leaving users free to apply whichever tool they feel will best represent the gene expression in their experiment. However, the accuracy of the corrections calculated by RoSA will be affected by this choice in the same way as the calculation of differential gene expression. If counting methods are used that only consider regions within sense features that do not overlap any antisense feature, the gene-specific corrections calculated by RoSA may be less accurate where the overlap is large and/or the sense or antisense expression is low.

RoSA then adjusts these raw counts to produce corrected antisense counts (Section 2.4). The *count_spliced* script generates sense and antisense counts of reads at splice junctions, used when estimating spurious antisense from spliced reads. The script takes a standard annotation (as gtf/gff) and corresponding alignment (as bam) and outputs counts of spliced sense and antisense reads to a designated output file.

RoSA takes several datasets containing different read counts as its input, for each replicate:

1. Full read counts by gene
2. Antisense counts by gene (via the *make_annotation* script)
3. At least one of:
 - a. Spike-in sense and antisense counts
 - b. Spliced sense and antisense counts (via RoSA's *count_spliced* script)

RoSA calculates and returns antisense:sense ratios for the spike-in data, or spliced read data, or both, and, for each gene and replicate, outputs new read counts values corrected for spurious antisense. RoSA also plots antisense versus sense counts of the original and corrected data, by replicate.

2.2 Using spliced reads

RoSA's main approach to estimating spurious antisense is to use spliced reads within the main dataset. Reads which map antisense to a multi-exon gene, and that also show the same splicing pattern as spliced sense-mapping reads are almost certainly spurious, as the splicing motif (canonically GU-AG) will be incorrect on the opposite strand (Figure 1). An estimate

of spurious antisense can be calculated by considering only spliced reads whose splices match annotated splice sites (*splice-matched reads*), and, as with the spike-ins, calculating the ratio of antisense to sense reads.

Splice-matched reads are identified by first filtering all spliced reads in the data. In a bam file of aligned reads, spliced reads have a CIGAR string containing 'N', indicating a skipped region. SAM processing tools such as sambamba²⁹ support filtering on the CIGAR string and can extract spliced reads rapidly. A second filtering step pulls out only those reads whose splice locations match at least one intron in the annotation, by processing each read in turn, identifying the spliced positions (based on the read location and the CIGAR string) and checking the annotation for a matching intron. Finally, the strand of each spliced read can be determined from its flag field value³⁰, and compared to the strand of the matching intron(s). Reads on the same strand as the intron(s) are counted as sense reads, and remaining reads as antisense reads. Since spurious antisense reads are misallocated sense reads, the number of antisense splice-matched reads assigned to a gene is strongly positively correlated with the number of its sense splice-matched reads (see Section 3). The ratio of antisense:sense counts on the splice-matched reads thus gives a simple global estimate of the level of spurious antisense across the whole dataset. Using spliced reads has the advantage that an antisense:sense ratio can be calculated on a gene-by-gene basis, for any spliced gene. Genes without any spliced reads fall back on the global estimate, calculated either from the spike-ins (see Section 2.3) or the spliced reads.

In the case of real, unannotated, antisense expression at a gene locus, the behaviour of RoSA falls into three categories:

1. If the splicing of the true antisense transcript differs from the sense transcript (including no splicing) then RoSA's gene specific correction will remove any spurious antisense expression (identified by antisense matches to the sense splicing) and leave the true antisense expression unchanged.
2. If the splicing of the antisense expression is the same as the sense strand, then RoSA will remove this completely.
3. If the true antisense splicing is the same as the sense strand in some parts of the transcript, but not across the entire transcript, then RoSA will remove a fraction of the true antisense expression depending on how similar the splicing patterns are.

We anticipate that occurrences of 2 & 3 will be uncommon in RNA-seq datasets. Point 2 highlights a minor potential limitation of the gene-specific splicing-based corrections calculated by RoSA, namely that it cannot distinguish between spurious antisense signal and potential biological transcription from RNA-dependent RNA polymerases (RdRPs). Although RdRPs are widespread in eukaryotic genomes, only 8–30% of eukaryotic gene regions have significant length ORFs on their opposite strands³¹, providing an upper limit on the potential impact of this method of transcription on the RNA complement within a cell. Eukaryotic RdRPs evolved independently from

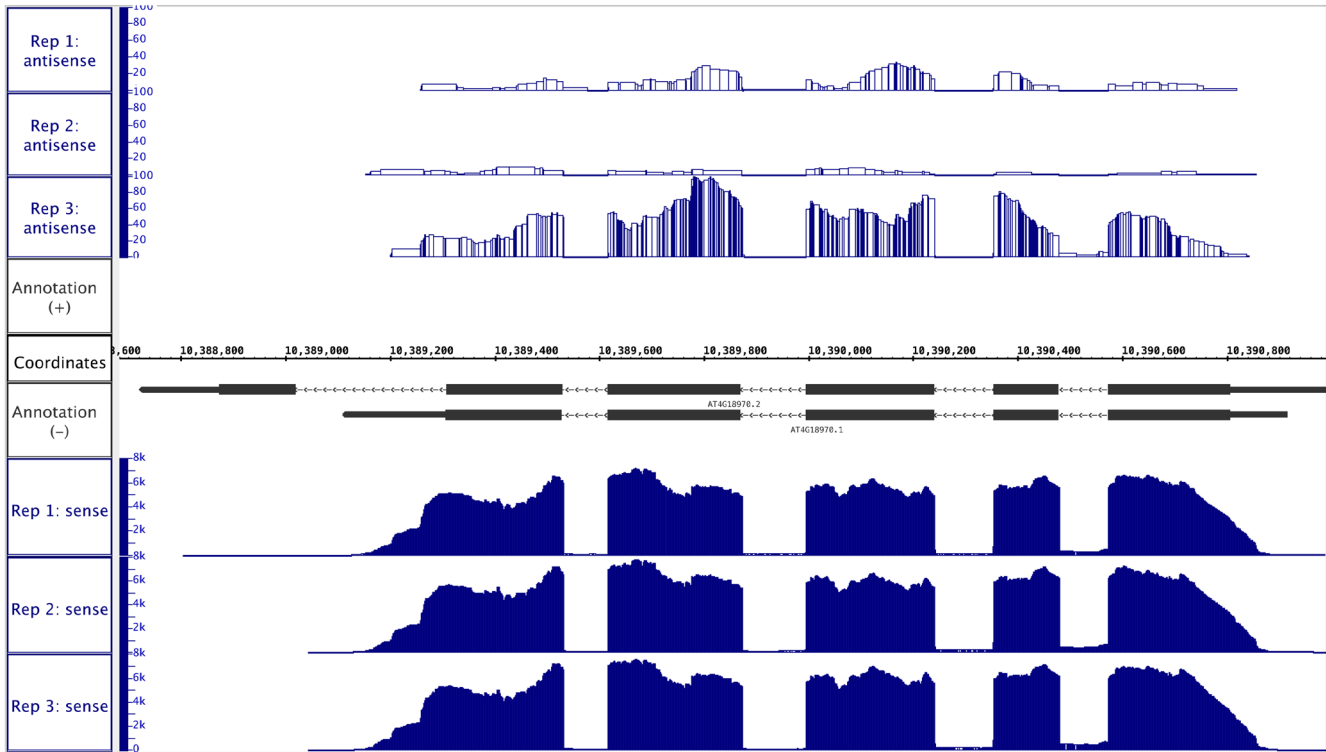


Figure 1. An example of spurious antisense reads displaying the same splicing structure as the sense strand. The reverse strand gene AT4G18970 is strongly expressed in all 3 replicates (bottom tracks). Spurious antisense can also be seen in all replicates (top tracks), with splice points in the antisense signal matching splice points in the sense signal. Furthermore, the level of spurious antisense varies noticeably between replicates. (Figure generated by IGB³²).

their viral counterparts and, in plants, are involved in siRNA transcriptional silencing³³. This is not the case in animals however (except in *C. elegans*) where their function remains elusive³⁴.

2.3 Using ERCC spike-ins

An alternative approach to estimating spurious antisense is to use ERCC spike-in data. Developed by the External RNA Control Consortium (ERCC)³⁵, the ERCC spike-in controls are synthetic RNA transcripts that are added to RNA-Seq experiments to act as controls³⁶. The 92 spike-ins are designed to mimic a range of eukaryotic mRNA characteristics, varying in length, GC-content and concentration, with a 20bp poly-A tail. They have minimal sequence similarity with known eukaryotic transcripts. Since the spike-ins are synthetic, they are unidirectional, and so any reads assigned as antisense to a spike-in can be assumed to be spurious. As the spike-ins are present at a wide range of concentrations, they are detected with a wide range of read counts, permitting an estimate of the ratio of antisense to sense read counts on the spike-ins to be calculated, which can then be used to estimate the contribution of spurious antisense transcripts across the full dataset. Obtaining sense and antisense counts for the spike-ins is straightforward. First we align the reads to the spike-ins (using the spike-in annotation file ERCC92.gtf, available at <https://www.thermofisher.com/order/catalog/product/4456739>) and then count reads, using a strand-aware read counting tool such as featureCounts³⁷, HT-SeqCount³⁸, etc. Now averaging the spurious antisense:sense ratio across all of the

spike-ins gives a global estimate of the spurious antisense, in just the same way as for the spliced reads.

2.4 Mitigating spurious antisense

Having identified high or differing levels of spurious antisense in an RNA-Seq experiment, we also want to correct for the incorrectly assigned reads so that true differential expression calling can be performed. The ratio of spurious antisense:sense read counts can be used as a simple correction factor. Defining r as the ratio of spurious antisense:sense and S and A respectively as the number of sense and antisense counts for a gene, the number of spurious antisense read counts A_s is estimated for each gene as: $A_s = r \cdot S$.

Then the antisense count can be corrected to account for the spurious antisense by taking $A - A_s$. This correction simply adjusts read counts for each gene, and does not identify specific reads as incorrectly assigned, so pile-ups cannot be adjusted. Since the spurious antisense reads are mis-assigned sense reads, RoSA then adds the spurious antisense count for each gene to its sense count.

2.4.1 Counting antisense reads. In order to apply the antisense correction factor, counts of antisense reads for each gene are required. Counting antisense reads is not directly supported by read counting tools. However, it can be performed with featureCounts³⁷ by setting parameters to indicate that reads are stranded

in the opposite direction to which they are. Unfortunately, if there are overlapping genes then reads in the overlaps will be counted twice using this tactic. As reads in regions of gene overlap are necessarily ambiguous, they cannot be considered to be antisense, spurious or otherwise. RoSA avoids this issue by building a custom antisense annotation based on the input sense annotation but excluding regions where genes on opposite strands overlap. Different gene transcripts are accounted for by merging all transcripts for a gene into a single *maximal transcript*. Whenever exons of different transcripts overlap in the annotation, the exon in the maximal transcript is the maximum extent of both exons. Given a maximal transcript, the script creates an antisense feature on the opposite strand which runs for the full extent of the maximal transcript. If the maximal transcript of another gene overlaps with the antisense feature, then the antisense feature is truncated to avoid overlapping. Once an antisense annotation is available, a read counting tool can be used to count antisense reads, by providing the antisense annotation instead of the standard annotation.

2.5 *Arabidopsis thaliana* datasets with spurious antisense

A procedure to experimentally generate RNA-Seq data with specific levels of spurious antisense is not known. Our main experimental data (Experiment 1) is therefore drawn from the study which originally motivated our investigation into incorrectly assigned antisense reads. In this study, spurious antisense occurred by chance at varying orders of magnitude across different replicates. Additionally, we perform a meta-analysis using three other *Arabidopsis thaliana* datasets (Experiments 2–4,³⁹) and data from ENCODE (see *Underlying data* for the full list of the ENA and ENCODE accessions).

2.5.1 *Arabidopsis* sample preparation and sequencing.

Briefly, the RNA-Seq data for Experiments 1 is wild-type (WT) *Arabidopsis thaliana* Colombia-0 (Col-0) biological replicates. WT *A. thaliana* Col-0 seeds were sown aseptically on MS10 plates. The seeds were stratified for 2 days at 4°C and then grown at a constant 21°C under a 16-h light/8-h dark cycle for a further 14 days, at the end of which the seedlings were harvested. Total RNA was isolated from the seedlings with the RNeasy Plant Mini Kit (Qiagen). In Experiment 1, DNase digestion was performed on column, as a part of RNA isolation, and 8 µl of ERCC spike-ins (External RNA Controls Consortium 2005) at a 1:100 dilution was added to 4 µg of total RNA. Libraries were prepared according to the TruSeq® Stranded mRNA Sample Preparation Guide Rev E. The libraries were sequenced on a HiSeq2000 at the Genomic Sequencing Unit of the University of Dundee. This preparation largely mirror the sample preparation of the datasets take from Froussios *et al.* (2017,³⁹, Experiments 2–4) In Experiments 2–4, however, the sequencing libraries were prepared using the Illumina TruSeq® Stranded Total RNA with Ribo-Zero Plant kit.

Experiment 1 has 3 replicates, processed as one batch, with a total of 4×10^8 150-bp paired-end reads. Experiments 2 and 3 have 7 biological WT replicates, while Experiment 4 has 3, for a total of 17 biological WT replicates and $\sim 1.7 \times 10^9$ 100-bp paired-end reads across the three experiments. The same lab sowed, grew and harvested the plants, and prepared the libraries. The sequencing was performed on the same machine by the

same people at the same sequencing facility and all the samples include the ERCC spike-ins which can verify the WT samples are consistent and comparable across experiments.

2.5.2 Quality control, alignment and quantification. The quality of the data was quantified using FastQC v0.11.2⁴⁰ with all the replicates performing as expected for high quality RNA-Seq data with good median per-base quality (≥ 28) across $>90\%$ of the read length. The read data for all experiments were aligned to the TAIR10⁴¹ *Arabidopsis thaliana* genome using the splice-aware aligner STAR v2.4.2a⁴² for Experiment 1 and STAR v2.5.0 for Experiments 2–4. The index was built with `--sjdbOverhang 149` (Experiment 1) or `--sjdbOverhang 99` (Experiments 2–4) and the alignment was run with parameters: `--outSJfilterIntronMaxVsReadN 5000 10000 15000 --outSAMAttributes All --outFilterMultimapNmax 2 --outFilterMismatchNmax 5 --outFilterType BySJout`.

The read data were also aligned to the ERCC spike-ins annotation, using the same parameters. Read counts per gene were then quantified from these alignments with featureCounts v1.5.0-p1 using the publicly available TAIR10 annotation with the parameters: `-s 2 -p -t exon --largestOverlap`. After running RoSA's *make_annotation* script to build an antisense annotation, antisense read counts per gene were quantified in the same way, with the parameters: `-s 2 -p -t antisense --largestOverlap`. Finally, spliced sense and antisense reads were counted using RoSA's *count_spliced* script with the TAIR10 annotation.

2.6 Operation

A full description of RoSA's environment, dependencies, installation and basic operation can be found on the RoSA GitHub repository. Briefly, RoSA is a combination of an R package and python scripts for data preprocessing. Minimal system requirements for the package are R v3.5+, python 2 v2.7+ the *LSD* R package and the python packages *scipy* (v0.16.1 - 0.17.1), *numpy*, *pandas* (not v0.20.1), *six* and, optionally, *drmaa* for cluster integration. The python scripts to find and count spliced antisense and sense reads also depends on *sambamba*. To facilitate ease-of-use, a conda environment that captures all the relevant dependencies is included as part of the RoSA codebase. RoSA's python scripts are provided as a python package and are installed via pip, while the R package can be installed directly from within R using the devtools package.

RoSA operates on the total and spliced read counts from sense and antisense bam format read alignments of stranded RNA-Seq datasets, either with or without ERCC spike-in standards. To facilitate easy generation of this read count data, RoSA includes helper pre-processing scripts to generate the antisense counterpart of the provided gtf/gff format sense-strand genome annotations (*make_annotation*), and to generate spliced-read gene count data from the bam format read alignments using both the sense- and anti-sense annotations (*count_spliced*). Both of these helper scripts can be called directly within R as part of the RoSA R package. Detailed help for the R RoSA functionality can be accessed within R with the command, *help(rosa)*.

3 Results

We used RoSA to analyze our data from Experiment 1 for spurious antisense, using both the spike-in and spliced reads

counts. RoSA calculated antisense:sense ratios for the spike-ins (Figure 2) showing that the 3 replicates have antisense:sense ratios on the spike-ins of 0.0008, 0.004 and 0.011. Although these ratios are small, if the replicates were being compared for differential expression, the differences are potentially substantial for highly expressed genes, and could lead to differential antisense expression being called erroneously.

For each replicate we calculated the spurious antisense:sense ratios for the spliced reads with RoSA, and compared them to the spike-ins. An overview of the results for all three replicates shows that the spurious antisense levels calculated from the spike-ins are in good agreement with the levels calculated from the spliced reads (Figure 3 and Figure 4, Row 1).

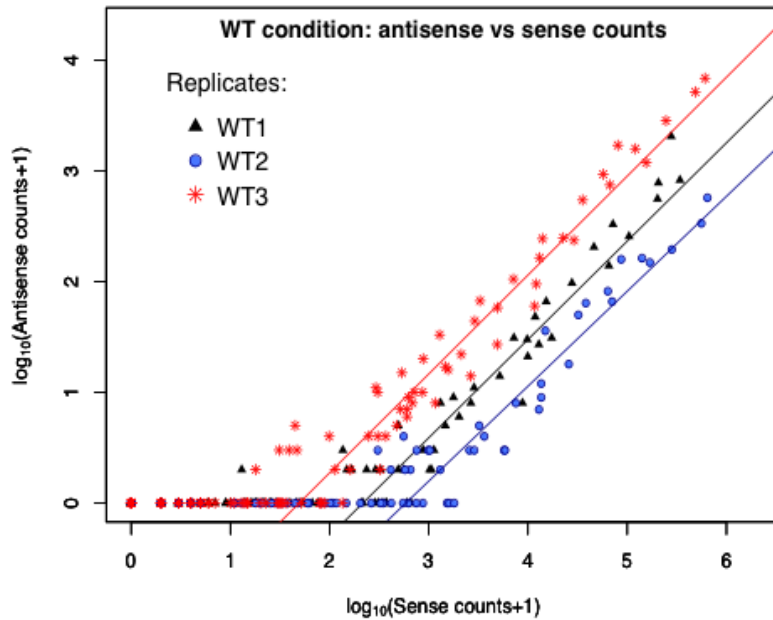


Figure 2. Antisense versus sense counts for the ERCC spike-ins for each replicate in Experiment 1. Points represent antisense and sense read counts for individual spike-ins. Each line is the average antisense:sense ratio for one replicate. Here, antisense:sense ratios vary by an order of magnitude across the 3 replicates, with values of 0.004 (WT1), 0.0008 (WT2), and 0.011 (WT3).

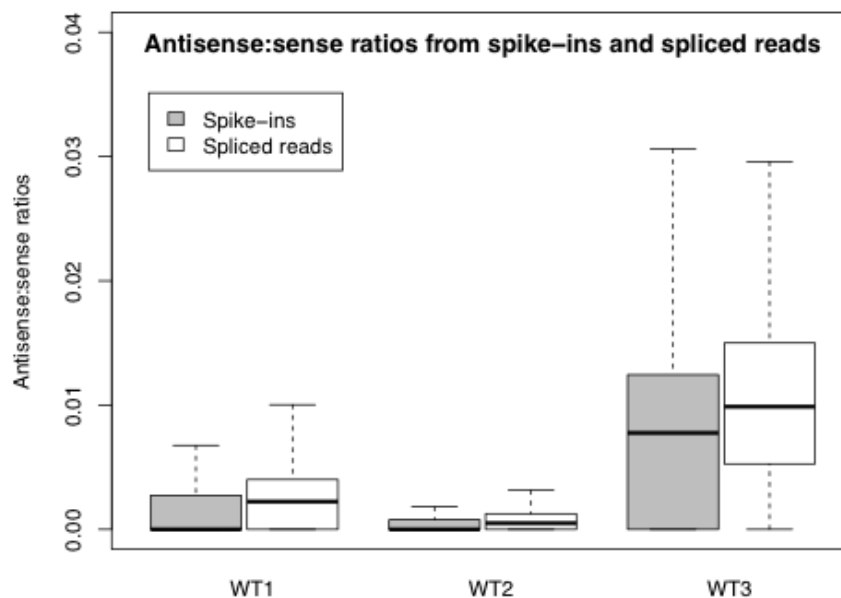


Figure 3. Comparison of antisense:sense ratios calculated from spliced reads or spike-ins, by replicate. Ratios estimated from spike-ins show good agreement with ratios estimated from spliced reads. Outliers not shown.

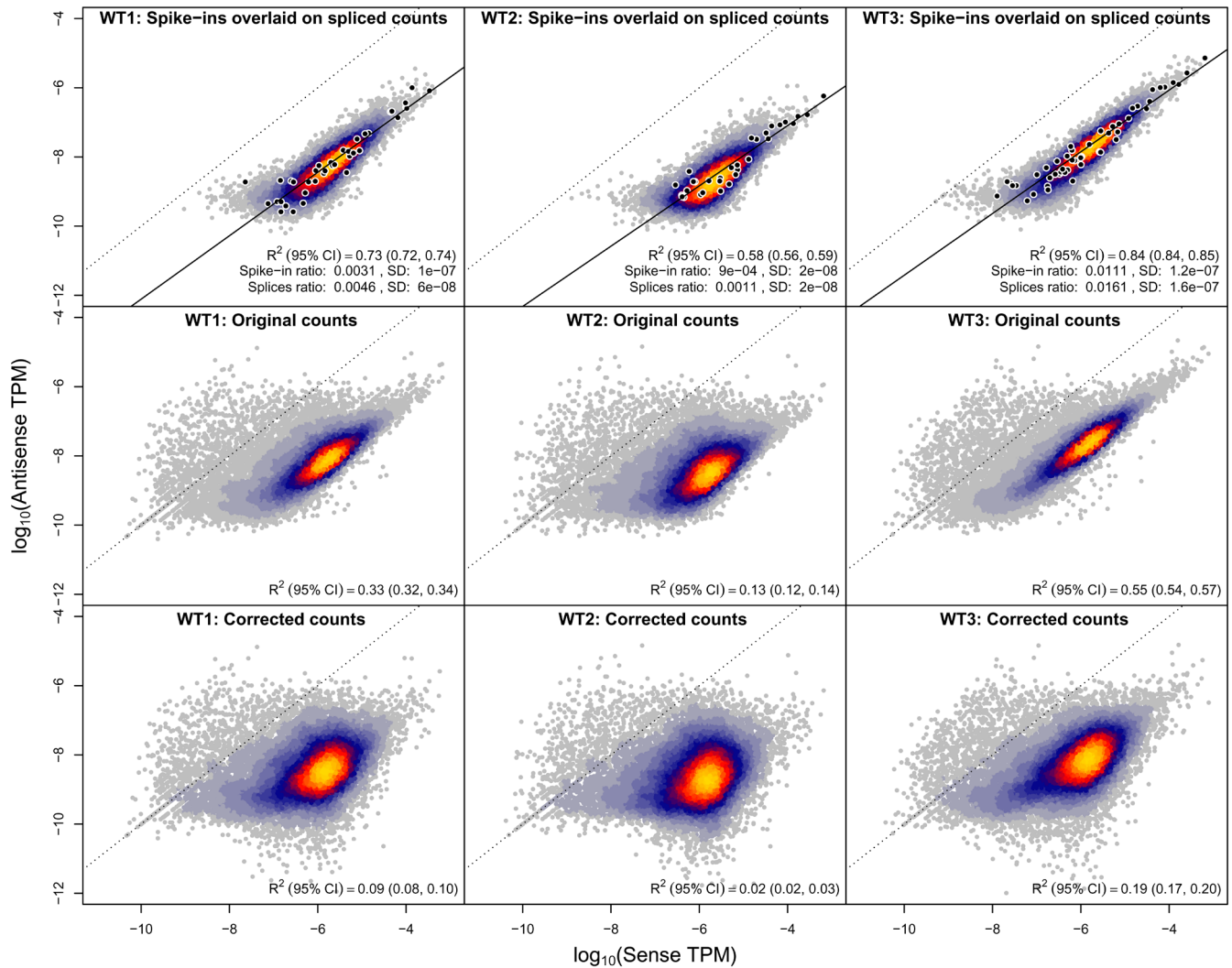


Figure 4. Normalised antisense versus sense counts by replicate. Each column presents data for one replicate. Row 1: Antisense:sense ratios calculated from spike-ins (Black points & fit line) and spliced reads for each gene (density heatmap). The antisense:sense ratios for both the spike-ins and spliced reads are in good agreement. The strong correlation between the sense and antisense spliced counts, and the constant antisense:sense ratio across all genes, indicates that the majority of the antisense expression in the data is not a sequence-, or gene-specific, phenomenon. Rather, this is what would be expected from a systematic process affecting a constant fraction of the sequenced reads. Row 2: Antisense:sense ratios calculated for the full gene counts (spliced & unspliced). The correlation between the sense and antisense expression persists, however it is weaker than the correlation using just the spliced and spike-in sense and antisense expression. This reflects the inclusion of true biological antisense expression, unspliced genes where a global correction is less accurate, and low expression genes where the splicing correction is not well measured. Row 3: Corrected antisense:sense ratios calculated for the full gene counts (spliced & unspliced). The corrected antisense counts show much weaker correlation with the corresponding gene counts reflecting the removal of the systematic spurious antisense count signal. On all plots the dashed line marks $y=x$; points above this line correspond to genes where the antisense:sense ratio is > 1 .

Finally, RoSA calculated a spurious antisense correction across the whole of each replicate. Every spliced gene was corrected with the antisense:sense ratio specific to the gene, and unspliced genes were corrected using the mean ratio calculated from the spike-ins. (RoSA also allows the unspliced correction to be calculated from the mean spliced reads ratio, for datasets without ERCC spike-ins). Overall, RoSA reduces the correlation between antisense and sense counts in the data (Figure 4, Rows 2 & 3), as would be expected with a reduction in incorrectly assigned reads. Two examples of corrections made by RoSA are shown in Figure 5, where the antisense signal appears to be

almost entirely spurious, RoSA's correction factor reduces the antisense counts substantially, but where there also appears to be some real antisense signal, RoSA's correction factor leaves a higher proportion of counts.

As well as identifying instances of antisense expression, looking at antisense counts in this way can also be useful in identifying misannotated genes. For example, in our data there are many genes where the antisense:sense ratio is more than 1 (e.g. see points lying above $x=y$ in Figure 4, Row 2), which may indicate an incorrect strand assignment in the annotation.

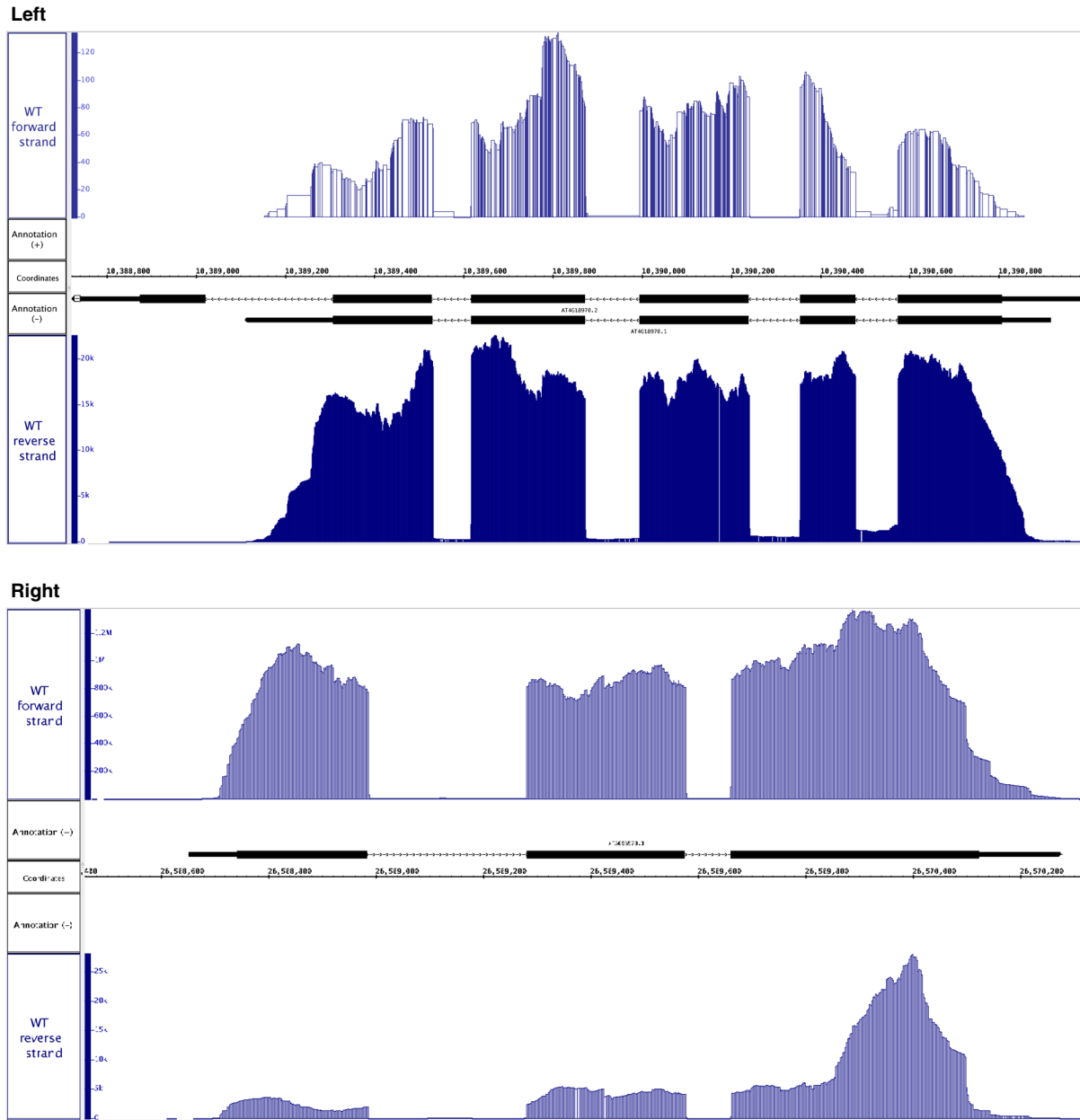


Figure 5. Two genes with differing antisense expression profiles and the read count corrections proposed by RoSA. The reverse strand gene AT4G18970 (left) has antisense expression which clearly matches the splice sites of the sense strand. RoSA eliminates almost all of the antisense reads. The forward strand gene AT5G66570 (right) has both antisense expression matching the sense strand splice sites, and a peak at the 5' end which is unlikely to have resulted from incorrect read assignment. RoSA only reduces the antisense counts by around 40%. (Figures generated by IGB³²).

3.1 Comparing antisense:sense ratios

Calculating antisense:sense ratios allows comparisons of spurious antisense to be made between replicates and between experimental condition, and can reveal whether there are systematic differences which might confound experimental comparisons. For example, [Figure 1](#) presents results from an RNA-Seq experiment where spurious antisense levels differed by an order of magnitude between replicates. In this experiment, the WT replicates had spurious antisense:sense ratios of 0.0031 (SD 0.0116), 0.0009 (SD 0.0070) and 0.0111 (SD 0.031).

To determine the extent of this problem for RNA-Seq datasets in general, we investigated the spurious antisense levels across a range of experiments and research groups. We analysed antisense reads assigned to the spike-ins from three other experiments in our lab (Experiments 2–4), as well as 195 publicly available human datasets from the ENCODE project that included the ERCC spike-ins²⁸ (see *Underlying data* for details of the sense, antisense and RoSA-corrected antisense expression for all *A. thaliana* genes in the datasets from Experiments 1–4). A separate antisense:sense ratio was calculated for each replicate in each experiment ([Figure 6](#)), showing that spurious antisense

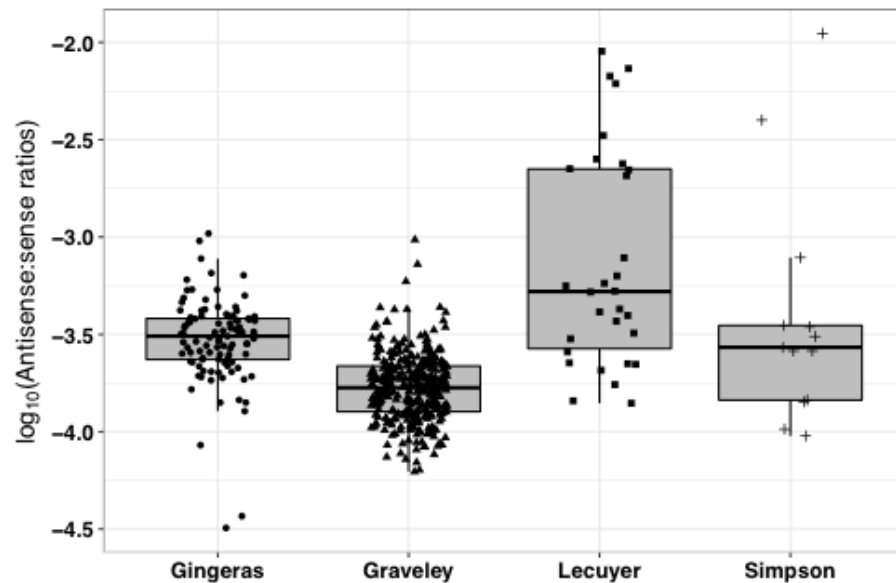


Figure 6. Spurious antisense:sense ratios for spike-ins, by research group. Data are from either from ENCODE (Gingeras, Graveley and Lecuyer) or our own group (Simpson). Each point represents the ratio for a replicate. Ratios range from 0.0111 to 0.00003.

reads are present at varying levels and can range across several orders of magnitude. This presents a serious quality control issue for anyone investigating differential antisense expression: a real difference in antisense expression could be completely masked by a difference in spurious antisense.

4 Conclusions

Spurious antisense is common in strand-specific RNA-Seq datasets and can occur at varying levels across replicates in the same experiment. Differing levels of such incorrectly assigned reads are enough to disrupt differential expression analyses of antisense gene expression.

We have developed a new tool, RoSA, which can detect, quantify and correct for spurious antisense. RoSA provides an important quality control step for researchers analyzing antisense expression in their data.

Data availability

Underlying data

Arabidopsis col-0 WT strand-specific RNA-Seq data from poly-A pulldown, Accession number E-MTAB-7990: <https://identifiers.org/arrayexpress/E-MTAB-7990>

RNA-seq data of wild type Arabidopsis seedlings, Accession number E-MTAB-5446: <https://identifiers.org/arrayexpress/E-MTAB-5446>

Extended data

Zenodo: bartongroup/RoSA: Initial, <http://doi.org/10.5281/zenodo.2661378>⁴³.

This project contains the following extended data:

- Accession numbers for ENCODE data: https://github.com/bartongroup/RoSA/tree/master/extras/F1000_manuscript/RoSA_Extended_Data.docx
- Accession details for ENCODE data: https://github.com/bartongroup/RoSA/blob/master/extras/F1000_manuscript/ENCODE_accessions.xlsx
- Arabidopsis seedlings RNA-seq read count expression counts: https://github.com/bartongroup/RoSA/tree/master/extras/F1000_manuscript/expression_data.csv

License: GNU General Public License 3.0.

Software availability

Source code available from: <https://github.com/bartongroup/RoSA>

Archived source code as at time of publication: <https://doi.org/10.5281/zenodo.2661378>⁴³.

License: GNU General Public License 3.0.

Grant information

This work has been supported by the Biotechnology and Biological Sciences Research Council [BB/M004155/1, BB/M010066/1] to G.J.B. and G.G.S.

All funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgments

A previous version of this article is available on BioRxiv: <https://doi.org/10.1101/425900>.

References

1. Pelechano V, Steinmetz LM: **Gene regulation by antisense transcription.** *Nat Rev Genet.* 2013; **14**(12): 880–893.
[PubMed Abstract](#) | [Publisher Full Text](#)
2. Matsui A, Iida K, Tanaka M, *et al.*: **Novel Stress-Inducible Antisense RNAs of Protein-Coding Loci Are Synthesized by RNA-Dependent RNA Polymerase.** *Plant Physiol.* 2017; **175**(1): 457–472.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
3. Lin S, Zhang L, Luo W, *et al.*: **Characteristics of Antisense Transcript Promoters and the Regulation of Their Activity.** *Int J Mol Sci.* 2015; **17**(1): pii: E9.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
4. Chan WY, Wu SM, Ruszczak L, *et al.*: **The complexity of antisense transcription revealed by the study of developing male germ cells.** *Genomics.* 2006; **87**(6): 681–92.
[PubMed Abstract](#) | [Publisher Full Text](#)
5. Swiezewski S, Liu F, Magusin A, *et al.*: **Cold-induced silencing by long antisense transcripts of an Arabidopsis Polycomb target.** *Nature.* 2009; **462**(7274): 799–802.
[PubMed Abstract](#) | [Publisher Full Text](#)
6. Liu F, Marquardt S, Lister C, *et al.*: **Targeted 3' processing of antisense transcripts triggers Arabidopsis FLC chromatin silencing.** *Science.* 2010; **327**(5961): 94–97.
[PubMed Abstract](#) | [Publisher Full Text](#)
7. Heo JB, Sung S: **Vernalization-mediated epigenetic silencing by a long intronic noncoding RNA.** *Science.* 2011; **331**(6013): 76–79.
[PubMed Abstract](#) | [Publisher Full Text](#)
8. Ietswaart R, Wu Z, Dean C: **Flowering time control: another window to the connection between antisense RNA and chromatin.** *Trends Genet.* 2012; **28**(9): 445–453.
[PubMed Abstract](#) | [Publisher Full Text](#)
9. Hobson DJ, Wei W, Steinmetz LM, *et al.*: **RNA polymerase II collision interrupts convergent transcription.** *Mol Cell.* 2012; **48**(3): 365–374.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
10. Crampton N, Bonass WA, Kirkham J, *et al.*: **Collision events between RNA polymerases in convergent transcription studied by atomic force microscopy.** *Nucleic Acids Res.* 2006; **34**(19): 5416–5425.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
11. Onodera CS, Underwood JG, Katzman S, *et al.*: **Gene isoform specificity through enhancer-associated antisense transcription.** *PLoS One.* 2012; **7**(8): e43511.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
12. Kawano M, Aravind L, Storz G: **An antisense RNA controls synthesis of an SOS-induced toxin evolved from an antitoxin.** *Mol Microbiol.* 2007; **64**(3): 738–754.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
13. Tufarelli C, Stanley JA, Garrick D, *et al.*: **Transcription of antisense RNA leading to gene silencing and methylation as a novel cause of human genetic disease.** *Nat Genet.* 2003; **34**(2): 157–165.
[PubMed Abstract](#) | [Publisher Full Text](#)
14. Sherstnev A, Duc C, Cole C, *et al.*: **Direct sequencing of Arabidopsis thaliana RNA reveals patterns of cleavage and polyadenylation.** *Nat Struct Mol Biol.* 2012; **19**(8): 845–52.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
15. Beiter T, Reich E, Weigert C, *et al.*: **Sense or antisense? False priming reverse transcription controls are required for determining sequence orientation by reverse transcription-PCR.** *Anal Biochem.* 2007; **369**(2): 258–261.
[PubMed Abstract](#) | [Publisher Full Text](#)
16. Haddad F, Qin AX, Giger JM, *et al.*: **Potential pitfalls in the accuracy of analysis of natural sense-antisense RNA pairs by reverse transcription-PCR.** *BMC Biotechnol.* 2007; **7**: 21.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
17. Perocchi F, Xu Z, Clauder-Münster S, *et al.*: **Antisense artifacts in transcriptome microarray experiments are resolved by actinomycin D.** *Nucleic Acids Res.* 2007; **35**(19): e128.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
18. Thomason MK, Storz G: **Bacterial antisense RNAs: how many are there, and what are they doing?** *Annu Rev Genet.* 2010; **44**: 167–88.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
19. Tzadok S, Caspin Y, Hachmo Y, *et al.*: **Directionality of noncoding human RNAs: how to avoid artifacts.** *Anal Biochem.* 2013; **439**(1): 23–29.
[PubMed Abstract](#) | [Publisher Full Text](#)
20. Yu WH, Høvik H, Olsen I, *et al.*: **Strand-specific transcriptome profiling with directly labeled RNA on genomic tiling microarrays.** *BMC Mol Biol.* 2011; **12**(1): 3.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
21. Houseley J, Tollervey D: **Apparent non-canonical trans-splicing is generated by reverse transcriptase in vitro.** *PLoS One.* 2010; **5**(8): e12271.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
22. van Dijk E, Jaszczyszyn Y, Thermes C: **Library preparation methods for next-generation sequencing: tone down the bias.** *Exp Cell Res.* 2014; **322**(1): 12–20.
[PubMed Abstract](#) | [Publisher Full Text](#)
23. Levin JZ, Yassour M, Adiconis X, *et al.*: **Comprehensive comparative analysis of strand-specific RNA sequencing methods.** *Nat Methods.* 2010; **7**(9): 709–15.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
24. Parkhomchuk D, Borodina T, Amstislavskiy V, *et al.*: **Transcriptome analysis by strand-specific sequencing of complementary DNA.** *Nucleic Acids Res.* 2009; **37**(18): e123.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
25. Jiang L, Schlesinger F, Davis CA, *et al.*: **Synthetic spike-in standards for RNA-seq experiments.** *Genome Res.* 2011; **21**(9): 1543–1551.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
26. Zeng W, Mortazavi A: **Technical considerations for functional sequencing assays.** *Nat Immunol.* 2012; **13**(9): 802–807.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
27. Garalde DR, Snell EA, Jachimowicz D, *et al.*: **Highly parallel direct RNA sequencing on an array of nanopores.** *Nat Methods.* 2018; **15**(3): 201–206.
[PubMed Abstract](#) | [Publisher Full Text](#)
28. ENCODE Project Consortium: **An integrated encyclopedia of DNA elements in the human genome.** *Nature.* 2012; **489**(7414): 57–74.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
29. Tarasov A, Vilella AJ, Cuppen E, *et al.*: **Sambamba: fast processing of NGS alignment formats.** *Bioinformatics.* 2015; **31**(12): 2032–2034.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
30. The SAM/BAM Format Specification Working Group: **Sequence Alignment/Map Format Specification.** 2017.
[Reference Source](#)
31. Winters-Hill S: **RNA-Dependent RNA Polymerase encoding Artifacts in Eukaryotic Transcriptomes.** *Int J Mol Genet Gene Ther.* 2017; **2**(1).
[Publisher Full Text](#)
32. Freese NH, Norris DC, Loraine AE: **Integrated genome browser: visual analytics platform for genomics.** *Bioinformatics.* 2016; **32**(14): 2089–2095.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
33. Iyer LM, Koonin EV, Aravind L: **Evolutionary connection between the catalytic subunits of DNA-dependent RNA polymerases and eukaryotic RNA-dependent RNA polymerases and the origin of RNA polymerases.** *BMC Struct Biol.* 2003; **3**(1): 1.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
34. Pinzón N, Bertrand S, Subirana L, *et al.*: **Functional lability of RNA-dependent RNA polymerases in animals.** *bioRxiv.* 2018.
[Publisher Full Text](#)
35. Baker SC, Bauer SR, Beyer RP, *et al.*: **The External RNA Controls Consortium: a progress report.** *Nat Methods.* 2005; **2**(10): 731–734.
[PubMed Abstract](#) | [Publisher Full Text](#)
36. ERCC: **NIST standard reference material 2374.** 2017.
[Reference Source](#)
37. Liao Y, Smyth GK, Shi W: **featureCounts: an efficient general purpose program for assigning sequence reads to genomic features.** *Bioinformatics.* 2014; **30**(7): 923–930.
[PubMed Abstract](#) | [Publisher Full Text](#)
38. Anders S, Pyl PT, Huber W: **HTSeq—a Python framework to work with high-throughput sequencing data.** *Bioinformatics.* 2015; **31**(2): 166–169.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
39. Froussios K, Schurch NJ, Mackinnon K, *et al.*: **How well do RNA-Seq differential gene expression tools perform in a eukaryote with a complex transcriptome?** *bioRxiv.* 2017.
[Publisher Full Text](#)
40. Andrews S: **FastQC: A quality control tool for high throughput sequence data.** 2010.
[Reference Source](#)
41. Arabidopsis Genome Initiative: **Analysis of the genome sequence of the flowering plant Arabidopsis thaliana.** *Nature.* 2000; **408**(6814): 796–815.
[PubMed Abstract](#) | [Publisher Full Text](#)
42. Dobin A, Davis CA, Schlesinger F, *et al.*: **STAR: ultrafast universal RNA-seq aligner.** *Bioinformatics.* 2013; **29**(1): 15–21.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
43. Schurch N: **bartongroup/RoSA: Initial (Version v1.0).** *Zenodo.* 2019.
<http://www.doi.org/10.5281/zenodo.2661378>

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research