Incomplete removal of ribosomal RNA can affect chromatin RNA-seq data analysis.

Michael Tellier* and Shona Murphy

Sir William Dunn School of Pathology, University of Oxford, Oxford, UK, OX1 3RE

* Contact: michael.tellier@path.ox.ac.uk

Keywords: Chromatin RNA-seq; ribosomal RNA; rRNA; transcription

MAIN TEXT

Next generation sequencing has become one of the major approaches to investigate transcription regulation. RNA-seq, which sequences the RNA complement, can provide a snapshot of the steady-state level of RNA. In addition to whole cell analysis, RNA-seq can be performed on different cellular fractions, such as chromatin, nucleoplasm, and cytoplasm, to investigate post-transcriptional regulation, for example. Chromatin RNA-seq provides a picture of the RNAs transcribed by RNA polymerase (pol) I, pol II, and pol III, associated with chromatin, which is a combination of nascent and processed transcripts. As chromatin RNA-seq cannot rely on a poly(A) tail enrichment method, a ribosomal (r)RNA-depletion step is performed during library preparation to avoid an over-representation of rRNAs in the sequencing data. The variety of commercial kits and protocols that are available each has its own strengths and weaknesses (1). Contrary to analysis of the data obtained from standard whole-cell RNA-seq or techniques investigating simultaneously nascent transcription of the three polymerases, such as precision nuclear run-on (PRO-seq), no bioinformatics step to remove the remaining rRNA reads is generally performed in chromatin RNA-seq (2-4).

We have reanalysed 16 chromatin RNA-seq datasets from four different studies investigating the roles in transcription of the elongation factor SPT6 (5), the termination factors CPSF73 (6) and PCF11 (7), and the RNaseH1 enzyme, which degrades the RNA of RNA-DNA hybrids (8). We found that after rRNA-depletion with the Illumina Ribo-Zero gold rRNA-removal kit, rRNAs reads are still present in the sequencing data and represent between 0.8% and 92.1% of the reads mapping to the GRCh38.p13 version of the human genome (Table 1). The proportion of reads mapping to rRNA genes was calculated by comparing the total number of reads mapping to the human genome with and without an intermediate step, where the reads are first mapped to an rRNA repeat to remove them (see Methods section). Importantly, differences across samples in the number of remaining rRNA reads within each

study indicate uneven rRNA depletion, which could affect downstream data analysis. The two samples with more than 90% of rRNA reads are likely due to a failure of the rRNA depletion step. However, their high sequencing depth is sufficient to still obtain more than 100 million paired-end reads mapping to non-rRNA regions.

We also investigated whether the results were the same using the GRCh37.p13 version of the human genome, which is still widely used in the literature and was the version of the genome used for the original analysis of these16 chromatin RNA-seq datasets (5-8). For this comparison, we mapped four of the 16 samples to GRCh37.p13 (Table 2). We found a similar result using both the GRCh38.p13 and GRCh37.p13 version of the human genome, indicating that rRNA genes are annotated in both versions of the human genome and therefore contaminating rRNA reads may also be an issue with older versions of the human genome. As GRCh38.p13 is a corrected and improved version of the genome, mapping to GRCh37.p13 generally gave a lower number of mapped reads.

An indication of incomplete rRNA depletion can be easily monitored in the distribution of reads between the forward and the reverse strand, as most of the rRNA genes are located on the forward strand. Human genes, either all genes or only protein-coding genes, are distributed evenly between both strands (Figure 1A). Similarly, analysis of nascent transcription by mNET-seq with a total pol II antibody shows an even distribution of the reads on the forward and reverse strands (Figure 1B). In the case of chromatin RNA-seq, a bias of the reads towards the forward strand is observed when the rRNA reads are still present, whereas the uneven strand distribution disappears following bioinformatics removal of the rRNA reads (Figure 1C). We then investigated whether the presence of contaminating rRNA reads could affect chromatin RNA-seq data analysis. To test this, we produced metagene profiles on a set of highly-expressed genes and compared the ratio of reads with and without the noted treatment, either without rRNA depletion (All reads) or with a bioinformatics rRNA depletion step (rRNA depletion) (Figure 2A-D). For some of the comparisons, such as siPCF11/siLuc repeat 2 and RNaseH1 overexpression/WT repeat 2, a clear change in the ratios is observed, indicating that the presence of different proportions of rRNA reads amongst the samples affects the outcome of the comparison. To determine whether the differences between presence or absence of rRNA reads are statistically significant, we quantified the chromatin RNA-seq data between the transcription start sites (TSSs) and the poly(A) site of the highly expressed genes followed by the calculation of the ratio of treatment to control (see Methods section, Figure 2E-H). We found statistically significant differences when rRNA

reads are effectively depleted or not effectively depleted in the eight comparisons, even where there is no visible change in the metagene profiles.

To determine whether the issue with contaminating rRNA reads can be avoided by mapping the reads to the genome without the genomic contigs, we investigated the location in the genome where the rRNA reads map (Table 3). We found that 99.9% of the rRNA reads map to three locations: two genomic contigs, GL000220.1 and KI270733.1 (26% each), which are known to contain rRNA genes, and chromosome 21 (48%), which also has rRNA genes (9). The mapping of ~ 50% of the rRNA reads to chromosome 21 indicates that even if the two genomic contigs containing rRNA genes are removed, high levels of contaminating rRNA reads may affect data analysis. We also note that reads mapping to rRNA genes are highly specific as only a tiny fraction of the mapped reads (< 0.03%) are lost from chromosomes without rRNA genes from the bioinformatics rRNA removal step.

We show here that even with a ribodepletion step, rRNA reads can still represent a significant proportion of the total reads of a chromatin RNA-seq library (Table 1). The non-removal of the rRNA reads can affect the conclusion when comparing treatment and control conditions, due to uneven rRNA depletion between samples as shown at the 5' end of two single genes, *NEK9* and *PGP*, in Figure 3. Although most of the samples analysed here did not show a major change following rRNA depletion, any comparison between two samples with an unaccounted significant difference in the proportion of rRNA reads can lead to an incorrect conclusion. We therefore recommend that any remaining rRNA reads are routinely removed from chromatin RNA-seq data with an additional bioinformatics step before downstream analysis.

METHODS

Transcription units' annotation

Gencode V31 annotation, based on the hg38 version of the human genome, was used to extract the location of the transcription units and the protein-coding genes. The set of 931 highly expressed genes in HeLa cells used for the metagene analyses are from (10).

mNET-seq data processing

Total pol II mNET-seq data from the GEO submission GSE110028 (5). mNET-seq data were processed as follows: adapters were trimmed with Cutadapt version 1.13 (11) in paired-end mode with the following parameters: -q 15, 10 **--**minimum-length 10 –A GATCGTCGGACTGTAGAACTCTGAAC –a

AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC. Trimmed reads were mapped to the human GRCh38.p13 reference sequence with with STAR version 2.6.1d (12) and the parameters --runThreadN 16 --readFilesCommand gunzip -c –k --limitBAMsortRAM 20000000000 --outSAMtype BAM SortedByCoordinate. SAMtools version 1.3.1 (13) was used to retain only properly paired and mapped reads (-f 3). A custom python script (14) was used to obtain the 3' nucleotide of the second read and the strandedness of the first read. Strand-specific bam files were generated with SAMtools.

Chromatin RNA-seq data processing

Chromatin RNA-seq data were obtained from the following GEO submissions: GSE137727: CPSF73 degradation (6), GSE110028: SPT6 knockdown (5), GSE123105: PCF11 knockdown (7), and GSE87607: RNaseH1 overexpression (8). Chromatin RNA-seq data were processed as follows: adapters were trimmed with Cutadapt version 1.13 in paired-end mode with the following parameters: -q 15, 10 –minimum-length 10 -A AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT -a AGATCGGAAGAGCACACGTCTGAACTCCAGTCA. The remaining rRNA reads were removed by mapping the trimmed reads to the rRNA genes defined in the Human ribosomal DNA complete repeating unit (GenBank: U13369.1) (Supplementary File 1) with STAR version 2.6.1d and the parameters --runThreadN 16 --readFilesCommand gunzip -c –k --outReadsUnmapped Fastx --limitBAMsortRAM 20000000000 --outSAMtype BAM SortedByCoordinate.

The unmapped reads were mapped to the human GRCh38.p13 reference genome with STAR and the parameters --runThreadN 16 --readFilesCommand gunzip -c –k --limitBAMsortRAM 20000000000 --outSAMtype BAM SortedByCoordinate. SAMtools version 1.3.1 was used to retain only properly paired and mapped reads (-f 3). Strand-specific bam files were generated with SAMtools. FPKM-normalized bigwig files were created with deepTools version 2.5.0.1 (15) bamCoverage tool with the parameters –bs 10 –p max –-normalizeUsing RPKM.

Metagene profiles

Metagene profiles were generated from FPKM-normalized bigwig files with Deeptools2 computeMatrix tool with a bin size of 10 bp and the plotting data obtained with plotProfile –outFileNameData tool. Graphs were then created with GraphPad Prism 8.4.2.

Reads quantification

Total read base count for chromatin RNA-seq data were computed with samtools bedcov tool using strand-specific bam files and normalized to 100 million paired-end reads and to the region's length. The

quantification is thus defined: log2(([region] * normalization factor) / length region). The quantification region was defined as the TSS to the poly(A) site. The ratio of treatment over control for each gene was then calculated. Violin plots, which were plotted with the minimal, first quartile, median, third quartile, and maximal values, were created with GraphPad Prism 8.4.2

P values and significance tests

Wilcoxon matched-pairs signed rank test was performed in GraphPad Prism 8.4.2.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Funding

This work was supported by Wellcome Trust Investigator Awards [WT106134AIA and WT210641/Z/18/Z to SM].

REFERENCES

1. Herbert, Z.T., Kershner, J.P., Butty, V.L., Thimmapuram, J., Choudhari, S., Alekseyev, Y.O., Fan, J., Podnar, J.W., Wilcox, E., Gipson, J. *et al.* (2018) Cross-site comparison of ribosomal depletion kits for Illumina RNAseq library construction. *BMC Genomics*, **19**, 199.
2. Lahens, N.F., Kavakli, I.H., Zhang, R., Hayer, K., Black, M.B., Dueck, H., Pizarro, A., Kim, J., Irizarry, R., Thomas, R.S. *et al.* (2014) IVT-seq reveals extreme bias in RNA sequencing. *Genome Biol*, **15**, R86.
3. Hebert, P.D.N., Braukmann, T.W.A., Prosser, S.W.J., Ratnasingham, S., deWaard, J.R., Ivanova, N.V., Janzen, D.H., Hallwachs, W., Naik, S., Sones, J.E. *et al.* (2018) A Sequel to Sanger: amplicon sequencing that scales. *BMC Genomics*, **19**, 219.
4. Mahat, D.B., Kwak, H., Booth, G.T., Jonkers, I.H., Danko, C.G., Patel, R.K., Waters, C.T., Munson, K., Core, L.J. and Lis, J.T. (2016) Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). *Nat Protoc*, **11**, 1455-1476.
5. Nojima, T., Tellier, M., Foxwell, J., Ribeiro de Almeida, C., Tan-Wong, S.M., Dhir, S., Dujardin, G., Dhir, A., Murphy, S. and Proudfoot, N.J. (2018) Deregulated Expression of Mammalian lncRNA through Loss of SPT6 Induces R-Loop Formation, Replication Stress, and Cellular Senescence. *Mol Cell*, **72**, 970-984 e977.
6. Eaton, J.D., Francis, L., Davidson, L. and West, S. (2020) A unified allosteric/torpedo mechanism for transcriptional termination on human protein-coding genes. *Genes Dev*, **34**, 132-145.
7. Kamieniarz-Gdula, K., Gdula, M.R., Panser, K., Nojima, T., Monks, J., Wisniewski, J.R., Riepsaame, J., Brockdorff, N., Pauli, A. and Proudfoot, N.J. (2019) Selective Roles of Vertebrate PCF11 in Premature and Full-Length Transcript Termination. *Mol Cell*, **74**, 158-172 e159.

8. Tan-Wong, S.M., Dhir, S. and Proudfoot, N.J. (2019) R-Loops Promote Antisense Transcription across the Mammalian Genome. *Mol Cell*, **76**, 600-616 e606.
9. Kim, J.H., Dilthey, A.T., Nagaraja, R., Lee, H.S., Koren, S., Dudekula, D., Wood Iii, W.H., Piao, Y., Ogurtsov, A.Y., Utani, K. *et al.* (2018) Variation in human chromosome 21 ribosomal RNA genes characterized by TAR cloning and long-read sequencing. *Nucleic Acids Res*, **46**, 6712-6725.
10. Chen, Y., Pai, A.A., Herudek, J., Lubas, M., Meola, N., Jarvelin, A.I., Andersson, R., Pelechano, V., Steinmetz, L.M., Jensen, T.H. *et al.* (2016) Principles for RNA metabolism and alternative transcription initiation within closely spaced promoters. *Nat Genet*, **48**, 984-994.
11. Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *2011*, **17**, 3.
12. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. and Gingeras, T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15-21.
13. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. and Genome Project Data Processing, S. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078-2079.
14. Nojima, T., Gomes, T., Grosso, A.R.F., Kimura, H., Dye, M.J., Dhir, S., Carmo-Fonseca, M. and Proudfoot, N.J. (2015) Mammalian NET-Seq Reveals Genome-wide Nascent Transcription Coupled to RNA Processing. *Cell*, **161**, 526-540.
15. Ramirez, F., Ryan, D.P., Gruning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dundar, F. and Manke, T. (2016) deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res*, **44**, W160-165.

GEO submission	Sample Name	Number of total mapped single/paired-end reads	Number of mapped reads after rRNA depletion	Number of reads mapping to rRNA genes (% of total mapped reads)
GSE137727; Single-end sequencing data (6)	Hela CPSF73-AID, no Auxin, Repeat 1	40,330,665	39,599,872	730,793 (1.8%)
	Hela CPSF73-AID, no Auxin, Repeat 2	51,271,377	49,895,133	1,376,244 (2.7%)
	Hela CPSF73-AID, with Auxin, Repeat 1	49,101,660	48,688,254	413,406 (0.8%)
	Hela CPSF73-AID, with Auxin, Repeat 2	51,551,588	48,317,460	3,234,128 (6.3%)
GSE110028; Paired-end sequencing data (5)	HeLa siLuc, Repeat 1	188,484,790	174,108,812	14,375,978 (7.6%)
	HeLa siLuc, Repeat 2	79,862,134	71,548,900	8,313,234 (10.4%)
	HeLa siSPT6, Repeat 1	184,121,714	177,454,196	6,667,518 (3.6%)
	HeLa siSPT6, Repeat 2	75,109,180	69,174,938	5,934,242 (7.9%)
GSE123105; paired-end sequencing data (7)	HeLa siLuc, Repeat 1	102,235,104	64,429,148	37,805,956 (37.0%)
	HeLa siLuc, Repeat 2	87,992,826	57,666,978	30,325,848 (34.5%)
	HeLa siPCF11, Repeat 1	105,287,068	64,606,322	40,680,746 (38.6%)

	HeLa siPCF11, Repeat 2	111,033,410	63,653,258	47,380,152 (42.7%)
GSE87607; paired-end sequencing data (8)	HeLa WT, Repeat 1	376,740,560	297,520,806	79,219,754 (21.0%)
	HeLa WT, Repeat 2	1,536,599,080	122,125,570	1,414,473,510 (92.1%)
	HeLa RNaseH1 overexpression, Repeat 1	382,709,514	312,362,380	70,347,134 (18.4%)
	HeLa RNaseH1 overexpression, Repeat 2	1,349,159,646	134,230,518	1,214,929,128 (90.1%)

Table 1: Summary of the chromatin RNA-seq reads mapped to the GRCh38.p13 version of the human genome with and without removal of rRNA reads.

GEO submission	Sample Name	Number of total mapped single/paired-end reads	Number of mapped reads after rRNA depletion	Number of reads mapping to rRNA genes (% of total mapped reads)
GSE87607; paired-end sequencing data (8)	HeLa WT, Repeat 1	276,453,712	251,959,340	24,494,372 (8.9%)
	HeLa WT, Repeat 2	582,483,250	102,627,365	479,855,885 (82.4%)
	HeLa RNaseH1 overexpression, Repeat 1	334,113,048	312,321,083	21,791,965 (6.5%)
	HeLa RNaseH1 overexpression, Repeat 2	537,888,953	120,001,805	417,887,148 (77.7%)

Table 2: Summary of the chromatin RNA-seq reads mapped to the GRCh37.p13 version of the human genome with and without removal of rRNA reads.

Chromosome or genomic contig	% of rRNA reads mapping in chromatin RNA-seq
Chr21	48% (± 0.8)
GL000220.1	26% (± 0.5)
KI270733.1	26% (± 0.3)

Table 3: Summary of the proportion of rRNA reads mapping to the GRCh38.p13 version of the human genome and contigs (n = 16 biological replicates).

FIGURES LEGENDS

Figure 1. Incomplete rRNA reads depletion affects the distribution of chromatin RNA-seq reads on forward and reverse strands.

A/ Proportion of transcription units (all genes) or protein-coding genes on the forward or reverse strand in the Gencode V31 annotation. **B/** Distribution of total pol II mNET-seq reads across the forward and reverse strand. **C/** Proportion of mapped chromatin RNA-seq reads for the CPSF73 degradation, SPT6 depletion, PCF11 depletion, or RNaseH1 overexpression datasets on the forward or reverse strand without (all reads) or with bioinformatics rRNA reads depletion (rRNA depletion).

Figure 2. Incomplete rRNA reads depletion affects chromatin RNA-seq data analysis for CPSF73 auxin-mediated depletion, SPT6 knockdown, PCF11 knockdown, and RNaseH1 overexpression.

A/ Metagene analyses of the chromatin RNA-seq across highly expressed protein-coding genes without (all reads, blue and light blue) or with rRNA reads depletion (red and orange). The metagene analysis are shown as the ratio of Auxin+ (CPSF73 degradation) over Auxin- (no CPSF73 degradation). **B/** Metagene analyses of the chromatin RNA-seq across highly expressed protein-coding genes without (all reads, blue and light blue) or with rRNA reads depletion (red and orange). The metagene analysis are shown as the ratio of siSPT6 over siLuc. **C/** Metagene analyses of the chromatin RNA-seq across highly expressed protein-coding genes without (all reads, blue and light blue) or with rRNA reads depletion (red and orange). The metagene analysis are shown as the ratio of siPCF11 over siLuc. **D/** Metagene analyses of the chromatin RNA-seq across highly expressed protein-coding genes without (all reads, blue and light blue) or with rRNA reads depletion (red and orange). The metagene analysis are shown as the ratio of RNaseH1 overexpression (OE) over wild-type (WT). **E/** Quantification of the chromatin RNA-seq ratio Auxin+ over Auxin- across the gene body of the highly expressed protein-coding genes, defined as TSS to poly(A) site. All reads: blue and light blue, rRNA depletion: red and orange. Statistical test: Wilcoxon matched-pairs signed rank test. **F/** Quantification of the chromatin RNA-seq ratio siSPT6 over siLuc across the gene body of the highly expressed protein-coding genes, defined as TSS to poly(A) site. All reads: blue and light blue, rRNA depletion: red and orange. Statistical test: Wilcoxon matched-pairs signed rank test. **G/** Quantification of the chromatin RNA-seq ratio siPCF11 over siLuc across the gene body of the highly expressed protein-coding genes, defined as TSS to poly(A) site. All reads: blue and light blue, rRNA depletion: red and orange. Statistical test:

Wilcoxon matched-pairs signed rank test. **H/** Quantification of the chromatin RNA-seq ratio RNaseH1 overexpression (OE) over WT across the gene body of the highly expressed protein-coding genes, defined as TSS to poly(A) site. All reads: blue and light blue, rRNA depletion: red and orange. Statistical test: Wilcoxon matched-pairs signed rank test.

Figure 3. Examples of incomplete rRNA reads depletion affecting the comparison between treatment and control.

Chromatin RNA-seq profiles of WT or RNAseH1 OE repeat 2 across *NEK9* and *PGP* without (all reads, light blue) or with bioinformatics rRNA reads depletion (orange). The sense of transcription is shown by the arrow above the gene.

Figure 1

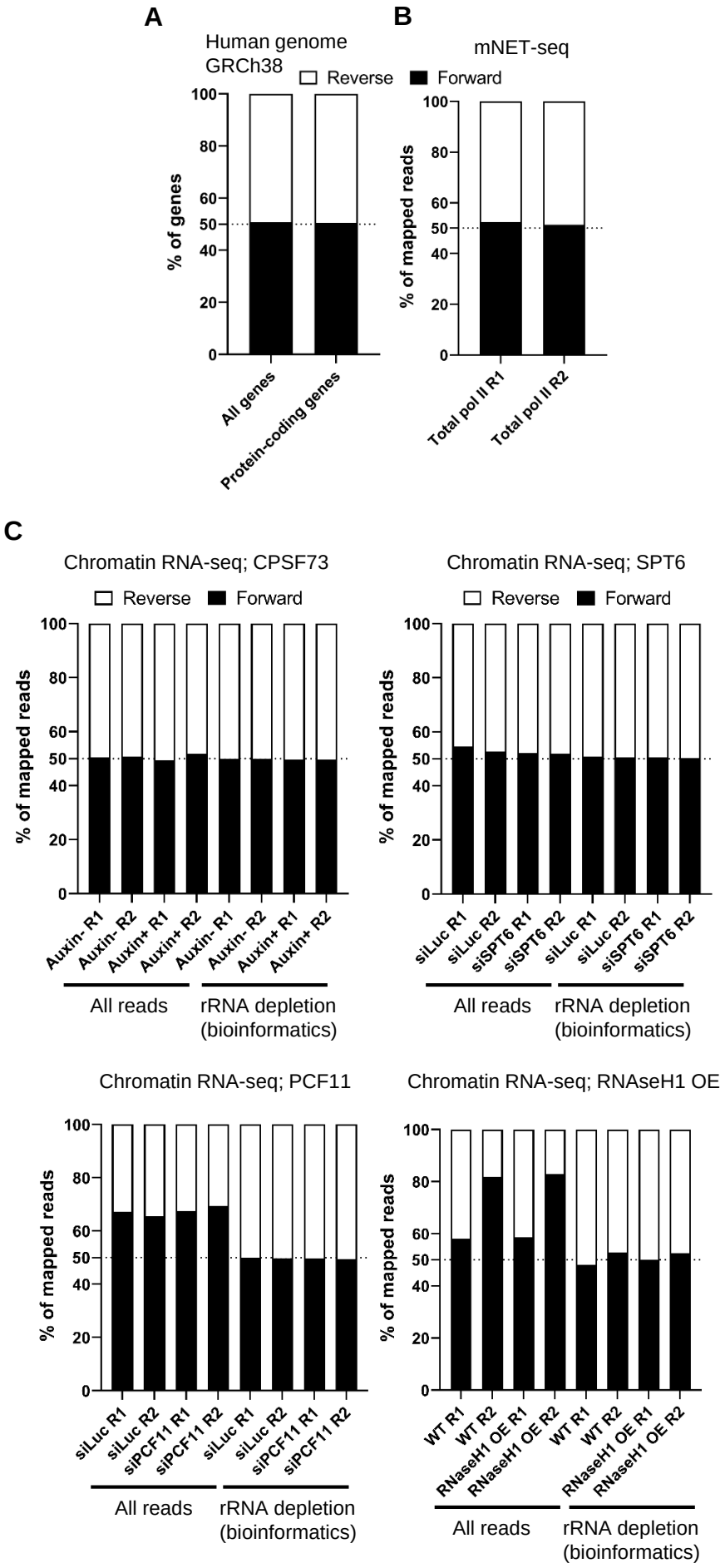

A
Human genome
GRCh38
Reverse
Forward
% of genes
All genes
Protein-coding genes
B
mNET-seq
% of mapped reads
Total pol II R1
Total pol II R2
C
Chromatin RNA-seq; CPSF73
% of mapped reads
Auxin- R1
Auxin- R2
Auxin+ R1
Auxin+ R2
All reads
rRNA depletion
(bioinformatics)
Chromatin RNA-seq; SPT6
siLuc R1
siLuc R2
siSPT6 R1
siSPT6 R2
Chromatin RNA-seq; PCF11
siPCF11 R1
siPCF11 R2
Chromatin RNA-seq; RNAseH1 OE
WT R1
WT R2
RNaseH1 OE R1
RNaseH1 OE R2

Figure 2

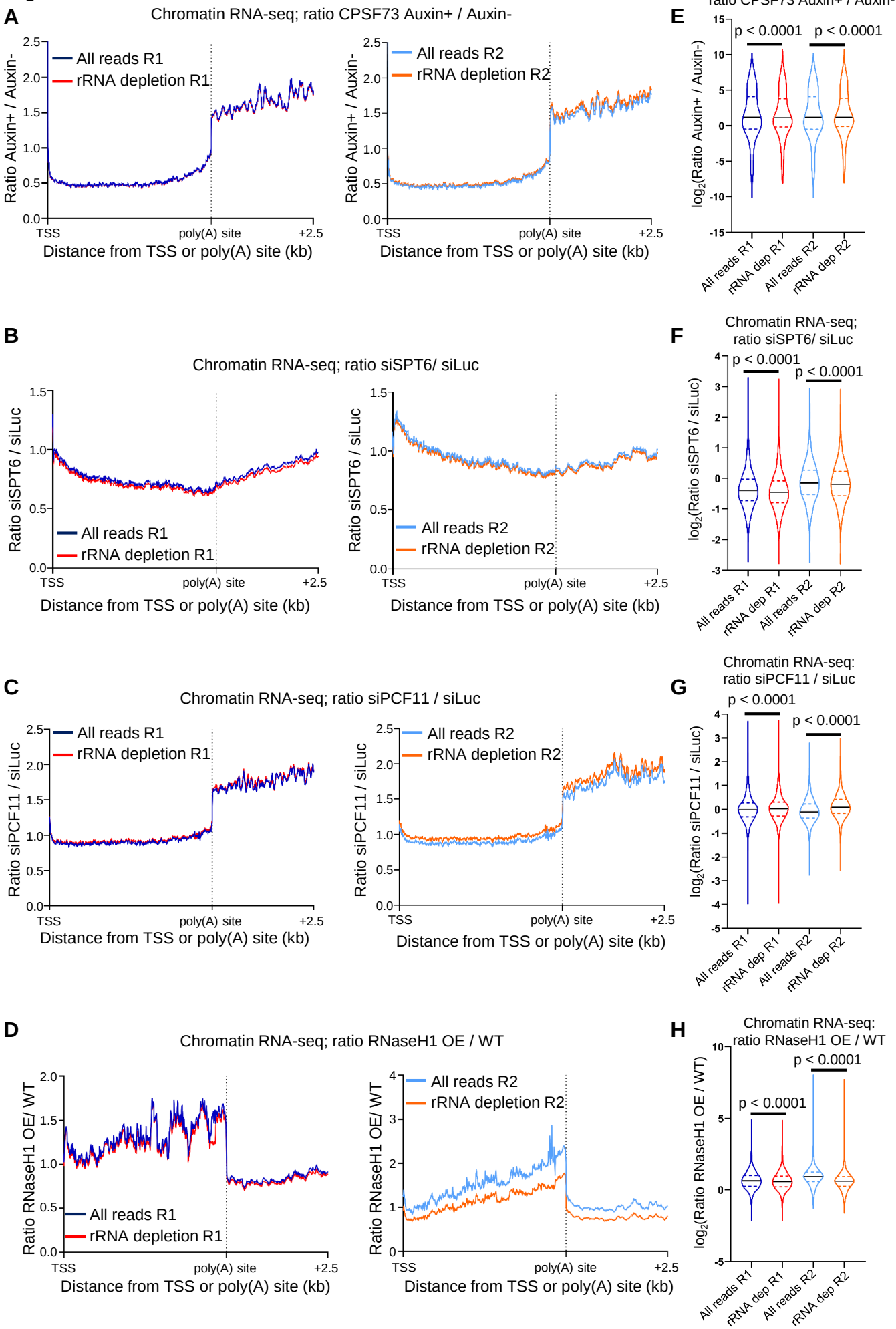
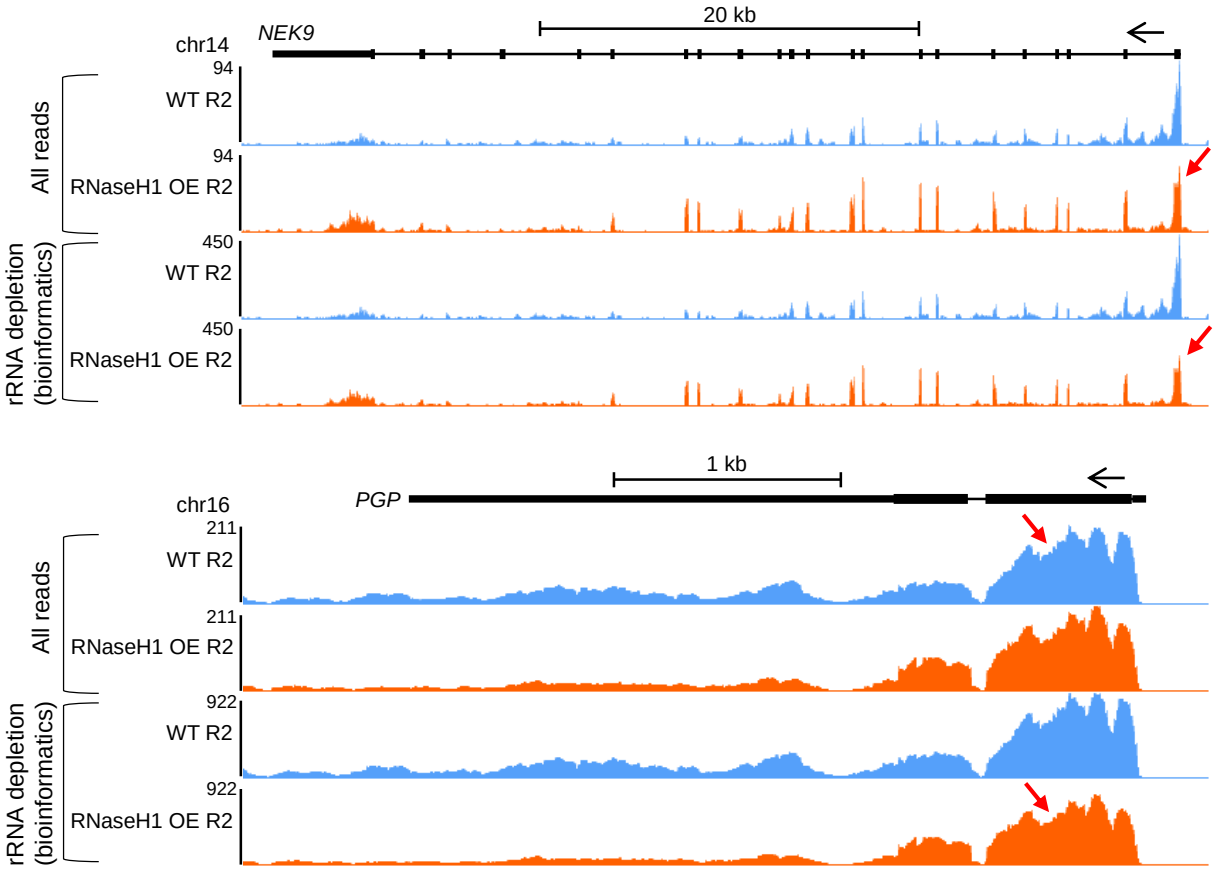

A
Chromatin RNA-seq; ratio CPSF73 Auxin+ / Auxin-
All reads R1
rRNA depletion R1
Ratio Auxin+ / Auxin-
TSS
poly(A) site
+2.5
Distance from TSS or poly(A) site (kb)
All reads R2
rRNA depletion R2
E
Chromatin RNA-seq;
ratio CPSF73 Auxin+ / Auxin-
p < 0.0001
p < 0.0001
log2(Ratio Auxin+ / Auxin-)
All reads R1
rRNA dep R1
All reads R2
rRNA dep R2
B
Chromatin RNA-seq; ratio siSPT6/ siLuc
Ratio siSPT6 / siLuc
F
Chromatin RNA-seq;
ratio siSPT6/ siLuc
log2(Ratio siSPT6 / siLuc)
C
Chromatin RNA-seq; ratio siPCF11 / siLuc
Ratio siPCF11 / siLuc
G
Chromatin RNA-seq:
ratio siPCF11 / siLuc
log2(Ratio siPCF11 / siLuc)
D
Chromatin RNA-seq; ratio RNaseH1 OE / WT
Ratio RNaseH1 OE/ WT
H
Chromatin RNA-seq:
ratio RNaseH1 OE / WT
log2(Ratio RNaseH1 OE / WT)

Figure 3


20 kb
chr14
NEK9
All reads
94
WT R2
94
RNaseH1 OE R2
rRNA depletion (bioinformatics)
450
WT R2
450
RNaseH1 OE R2
1 kb
chr16
PGP
All reads
211
WT R2
211
RNaseH1 OE R2
rRNA depletion (bioinformatics)
922
WT R2
922
RNaseH1 OE R2