Reports

BioTechniques

ncRNAseq: simple modifications to RNA-seq library preparation allow recovery and analysis of mid-sized non-coding RNAs

Nicola Minshall[‡,2], **Igor Chernukhin**[‡,1], **Jason S Carroll**[1] & **Anna Git***[,2]

[1]*Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, UK;* [2]*Department of Biochemistry, University of Cambridge, Cambridge, UK; *Author for correspondence: ag229@cam.ac.uk; ‡Authors contributed equally*

BioTechniques 72: 00–00 (January 2022) 10.2144/btn-2021-0035

First draft submitted: 10 October 2021; Accepted for publication: 9 November 2021; Published online: 29 November 2021

ABSTRACT

Despite their abundance, mid-sized RNAs (30–300 nt) have not been extensively studied by high-throughput sequencing, mostly due to selective loss in library preparation. The authors propose simple and inexpensive modifications to the Illumina TruSeq protocol (ncRNAseq), allowing the capture and sequencing of mid-sized non-coding RNAs without detriment to the coverage of coding mRNAs. This protocol is coupled with a two-step alignment: a pre-alignment to a curated non-coding genome, passing only the non-mapping reads to a standard genomic alignment. ncRNAseq correctly assigns the highest read-numbers to established abundant non-coding RNAs and correctly identifies cytosolic and nuclear enrichment of known non-coding RNAs in two cell lines.

METHOD SUMMARY

In order to retain RNAs shorter than 300 nt and the cDNAs they give rise to, the authors reduce the fractionation time of the RNA and increase the efficiency of four precipitation steps in the Illumina TruSeq protocol (ncRNAseq): twice pre-ligation by adding isopropanol to AMPure beads and twice post-ligation by using a higher AMPure-to-sample ratio. The resulting sequencing reads are aligned first to a custom non-redundant curated 'non-coding genome', which allows the user to exert control, for example, on amalgamation of multi-copy genes or separation of non-coding intronic gene counts from coding host gene counts. The non-mapping reads are then passed to a standard genomic alignment in a lossless manner.

KEYWORDS:

5′ coverage ● high-throughput sequencing ● next-generation sequencing ● non-coding RNA ● size-selection

The transcriptome sequencing (RNAseq) efforts of the last decade greatly advanced our knowledge of the expression levels of coding mRNAs, and to a lesser extent that of long non-coding RNAs (lncRNAs) and short RNAs [1], such as microRNAs, piRNAs or tRNA fragments. Mid-sized RNAs (30–300 nt) include snRNAs, snoRNAs, tRNAs, Y RNAs, vault RNAs and numerous other classes, as well as potentially active RNA fragments [2]. Cumulatively, they can constitute up to 15% of mammalian cellular total RNA by mass (Figure 1A, agarose gel) and even more by transcript number but have been largely passed over by the RNAseq efforts.

Three types of reasoning led to this lack of interest, especially in mammalian cell research. First, the 30–300 nt RNA pool is often regarded as rRNA/mRNA degradation products. In fact, some RNA purification kits (e.g., Qiagen RNeasy) intentionally remove most of this fraction, while others (e.g., Zymo Direct-Zol or NEB Monarch) offer purification protocols to either include or exclude these sizes. Poor precipitation of short RNA using solid phase reversible immobilization (SPRI) beads adds further undeclared size-selective steps during common RNAseq library preparation methods. In reality, in good quality total RNA, the vast majority of the <300 nt human RNAs are discrete transcripts with little background of non-specific degradation products (Figure 1A, PAGE gel). While some of the bands could arise from preferential degradation products of abundant RNAs (most likely rRNA), most can be matched by length to abundant sequenced non-coding transcripts.

Second, many classes of mammalian mid-sized RNAs, such as tRNAs and snRNAs, were traditionally deemed 'housekeeping', and thus of no research interest. The current, more rounded understanding of cell biology largely rejects the concept of 'housekeeping' genes. For example, not only does the oncogenic potential of RNA Pol III [3,4] suggest the potential physiological importance of the mid-sized RNA classes it transcribes, but there is a growing body of direct evidence implicating 'classical' non-coding RNA in development and disease [5,6].

Finally, investigating mid-sized RNAs is both technically and computationally challenging. At the bench, care is required to purify them and to maintain them during library preparation; their reverse transcription (RT) is often hindered by secondary structures or base modifications [7]; and stringent selection of sequencing library insert size (to avoid primer dimers) further biases the library against RNAs that cannot yield long inserts. Computationally, many mid-sized RNAs align to multiple (near-)identical genomic loci, exacerbated by possible

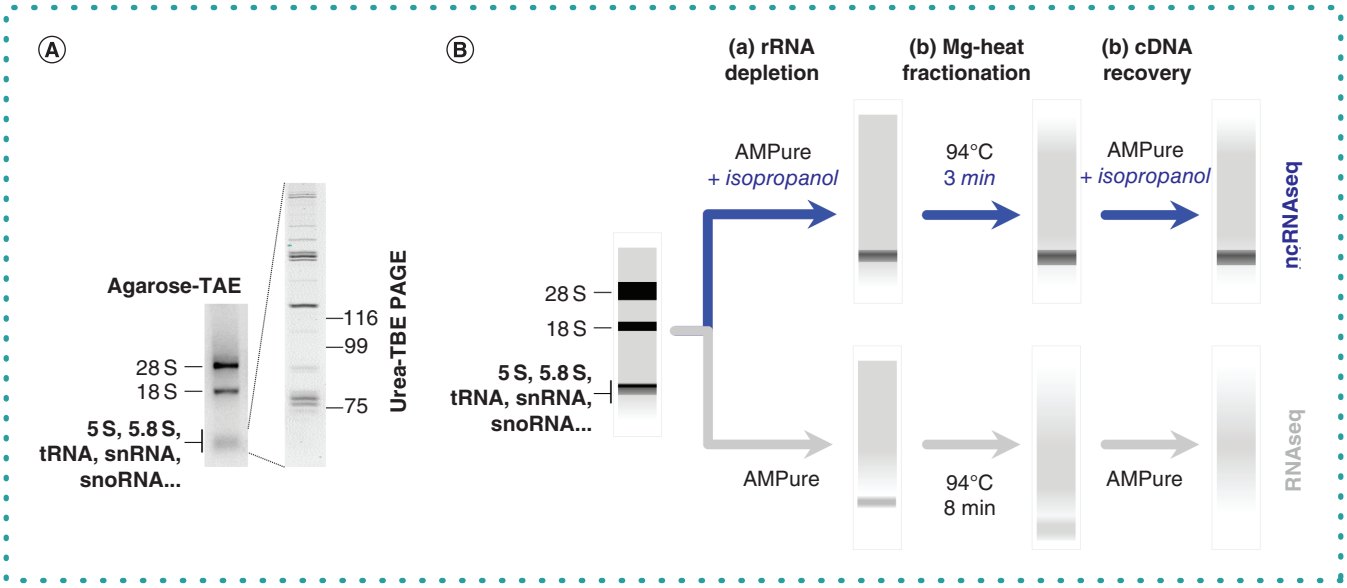

Figure 1. RNA size distribution in cellular total RNA and during sequencing library preparation. (A) Up to 15% of cellular RNA is shorter than 300 nt, and that fraction contains predominantly discrete transcripts. The migration of the mid-sized (~30–300 nt) RNAs is indicated on a typical image obtained after resolving total human cell line RNA on a non-denaturing 1% agarose-TAE gel impregnated with SYBR Safe (left). SYBR Gold staining of total RNA resolved on a denaturing 10% urea-TBE polyacrylamide gel (PAGE; right) reveals mostly discrete transcripts. Migration of rRNA and other types of non-coding RNA is indicated on the left; single-stranded DNA markers (nt) are indicated on the right. **(B)** A schematic representation of the RNA populations retained during key stages of library prep pre-ligation. The migration of rRNA and other types of non-coding RNA in the input RNA sample is indicated on the left. Three key stages **(a–c)** and the proposed changes to protocol (blue) are indicated; see the text for details.

misincorporation of nucleotides due to chemical modification of the RNA [8]. Multi-mapping of reads challenges the enumeration steps of the analytical pipeline [9], further confounded by incomplete gene annotation due to a historical shortage of data.

Materials & methods

Total RNA extraction

In order to minimize the effect of possible polymorphisms, the ncRNAseq method was optimized using a pool of RNAs from 12 breast cancer cell lines (HB4a, SkBr-7, MCF-10a, PMC42, MFM223, MDA-MB-231, MDA-MB-436, MDA-MB-468, BT-549, Hs578T, Hs578Bst and HMT-3522). For simplicity, all lines were propagated in antibiotic-free Dulbecco's Modified Eagle Medium (DMEM) supplemented with 10% fetal calf serum and glutamine. Subconfluent cultures were harvested directly into Qiazol (QIAGEN), and total RNA was extracted using miRNeasy columns (QIAGEN) following the manufacturer's protocol. A sample pool was generated by mixing equal amounts of total RNA (by mass).

Cytosolic-nuclear RNA extraction

20–25 × 10^6 monolayer cells were harvested by trypsinization, washed in phosphate buffered saline (PBS) and collected by centrifugation. The pellet (dry volume ~50 µl) was resuspended in 750 µl of hypotonic lysis buffer (HLB: 10 mM Hepes pH 7.5, 10 mM NaCl, 3 mM $MgCl_2$ and 0.35 M Sucrose), and NP-40 was added to a final 0.25% and mixed by gentle inversion. Following a 5 min incubation on ice, nuclei were pelleted by a 5 min spin at 1300 g. After removal of the supernatant (the cytosolic fraction), the pellet was washed twice as above in HLB and finally resuspended in 750 µl HLB with 0.25% NP-40. 50 µl cytosolic and nuclear fractions were denatured in 650 µl TRIzol (Sigma), and total RNA was extracted on DirectZol columns (Zymo) following the manufacturer's protocol, including an on-column DNaseI digest. The remaining ~700 µl lysates were used for protein analysis to verify negligible cross-contamination between the fractions (Coomassie Blue staining to visualize histones in the nuclear fraction and western blot analysis to confirm Major Vault Protein in the cytosolic fraction; data not shown).

ncRNAseq

500 ng of pooled total RNA was sequenced using TruSeq Stranded Total RNA protocol (part no. 15031048 Rev. E October 2013) either as recommended (RNAseq) or with the changes outlined in the main text (ncRNAseq). The resulting libraries had an average size of 424 bp (RNAseq) or 291 bp (ncRNAseq) and were sequenced using a Single End 50 cycling, yielding 31 M and 35 M reads, respectively. Bioanalyzer traces of the libraries are available in Supplementary Figure 1 (in Supplementary File 1). Raw fastq files can be found in the Apollo repository [10].

The typical reproducibility of ncRNAseq between technical replicates is 0.98–0.99 (data not shown). Supplementary Figure 2 (in Supplementary File 1) demonstrates a correlation of $R^2 \geq 0.97$ between independently processed and sequenced biological replicates of two cell lines for both ncGenome- and nuclear genome-aligned reads. Please note that the data points 'off the correlation diagonal' map to 28S, 18S and mitochondrial rRNA transcripts, which result from slight differences in efficiency of rRNA depletion in independent library preparations.

300 ng of cytosolic/nuclear RNA from individual cell lines was only sequenced using ncRNAseq.

Test ncGenome construction

The base ncRNA annotation Homo_sapiens.GRCh38.ncrna.fa was obtained from Ensembl [11]. Of the 40,001 entries, those corresponding to lncRNAs shorter than 300 bp and all entries with a BIOTYPE of "miRNA", "misc_RNA", "Mt_rRNA", "Mt_tRNA", "non_coding", "ribozyme", "rRNA", "scaRNA", "scRNA", "snoRNA", "snRNA", "sRNA", "Nc_tRNA", "RNY", "7SK", "7SL", "ALU" or "vaultRNA" were retained. tRNA annotation was added from GtRNAdb [12].

Entries corresponding to 2904 pseudogenes of key ncRNA families (7SL, 7SK, RNY, RNU, RNA5S) and 2118 genes of unknown function starting with RF0 were omitted. Seventy-eight individual highly homologous members of the SNORD115/116 family were replaced with a consensus calculated using the 'cons' option of EMBOSS [13]. Finally, differently named entries spanning identical sequences were collapsed, keeping the longest entry and collating all names, resulting in 3111 non-redundant entries (Supplementary File 2). A calculated database of shared 50-mer sequence stretches was kept as a reference for manually scrutinizing specific gene enumerations.

Read alignment

Pre-alignment to the test ncGenome was carried out on raw Illumina read files without pre-processing [14] using *Rsubread/align* [15]. Alignment to the standard genome and computation of exon-exon junctions was carried out using *Rsubread/subjunc* with the reference GRCh38 human genome gap-indexed, omitting TH_subreads that occur >1000-times in the genome. Input fastq files were raw Illumina fastq files or the same files with ncGenome-aligning reads removed using *samtools view –f 4* and then converted using *samtools fastq* [16]. Key alignment statistics are summarized in Figure 3C. Genes with fewer than 10 reads in all samples were omitted for downstream analysis. RNAseq/ncRNAseq reads were analyzed with no further normalization. Cytosolic/nuclear ncGenome read counts were linearly scaled to 20 M reads after elimination of reads mapping to rRNA and the highly abundant 7SL and 7SK RNAs (Supplementary File 3). For visualization, bam files were loaded into an Integrative Genomics Viewer (IGV) genome browser [17] and example loci were graphically captured.

Results & discussion

In order to minimize selective loss of mid-sized RNAs and their cDNAs during library preparation, five simple changes were introduced into the Illumina TruSeq Stranded Total RNA protocol (part no. 15031048 Rev. E October 2013), as listed below. Unlike previously suggested *de novo* protocols [18], which many labs are reluctant to risk, these minor biochemical adjustments remain within the robust and well-studied TruSeq pipeline. Similar changes can be rationally introduced into other kit protocols. A schematic effect of changes (a–c) is offered in Figure 1B.

(a) *'Clean Up RCP' step 1.* To avoid selective loss of shorter RNA from post-rRNA depletion supernatant: in addition to 99 μl AMPure beads, add 250 μl isopropanol to the ~40 μl supernatant. Following mixing and the recommended 15 min incubation, magnetic bead capture needs to be extended to 15 min.

(b) *'Incubate 1 DFP' step 1b.* To minimize critical damage to short RNAs while allowing the necessary fragmentation of longer transcripts: fragment the RNA @ 94°C for 3 min instead of the default 8 min.

(c) *'Clean Up DFP' step 2.* To avoid selective loss of short cDNAs from the 2nd strand synthesis reaction: in addition to 90 μl AMPure beads, add 250 μl isopropanol to the 50 μl reaction. Following mixing and the recommended 15 min incubation, magnetic bead capture needs to be extended to 15 min.

(d) *'Clean Up ALP' step 2.* To recover ligated cDNA including shorter inserts: use 1.25x volumes of AMPure beads instead of the recommended 1x volume.

(e) *'Clean Up PCR' step 3.* To recover PCR-amplified library spanning shorter inserts: use 1.2x volumes of AMPure beads instead of the recommended 1x volume.

The same pool of human cell line RNA was sequenced using the recommended (RNAseq) and modified (ncRNAseq) methods side-by-side. Direct comparison of the read counts after a standard genomic alignment revealed gains of counts up to >1000-fold, predominantly for genes shorter than 300 nt (Figure 2, left), across the entire abundance range (read counts used as a proxy; Figure 2, right). In particular, tRNAs, snoRNAs and snRNAs are highlighted. In contrast, there are no observable systematic losses of reads across all gene lengths or expression abundance. Indeed, the median enhancement (solid blue line) is essentially zero (-0.13), demonstrating no crippling loss of sequencing depth for longer genes despite a substantial gain in information regarding the mid-sized transcriptome. Figure 3A visualizes the ncRNAseq enhancement in two example genomic loci with snoRNA genes resident in introns of coding genes (solid green boxes).

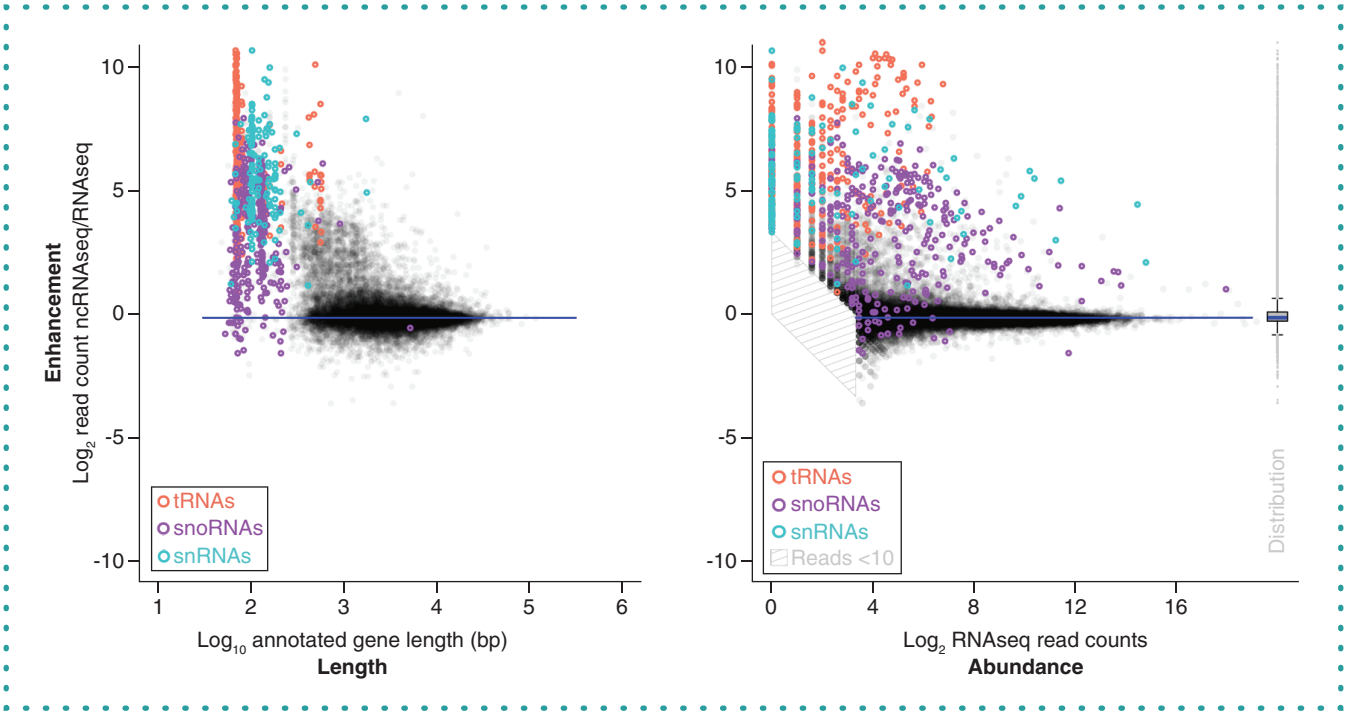

Figure 2. ncRNAseq selectively enhances the representation of mid-sized RNAs. RNAseq and ncRNAseq reads of the same RNA sample were aligned to the human genome, and the enhancement in number of reads was plotted against the $\log_{10}$ of the annotated gene lengths (left) or against the $\log_2$ read count in RNAseq as a surrogate measure of transcript abundance (right). Genes with fewer than 10 reads in both libraries were excluded. The distribution of enhancement across all genes is represented as a boxplot on the right with the median enhancement in blue. The effect of ncRNAseq on reads which map to tRNAs, snoRNAs and snRNAs (identified as gene names structured 'TR_–', starting with 'SNOR' or starting with 'RNU', respectively) is highlighted in orange, purple and aqua, respectively.

Genes longer than 300 bp with increased read number in ncRNAseq (Figure 2, left) are by and large poorly transcribed pseudogenes (median read count of 34, compared with 212 in the entire dataset), often with misannotated gene boundaries that span non-coding genes. Almost indistinguishable observations were obtained using a different cell line RNA pool and in technical repeats of both datasets (data not shown).

To evaluate the representation of ncRNAs in the two methods and to aid subsequent enumeration of multi-copy genes, the non-coding RNA annotation of the Human Genome (GRCh38) was curated to generate an artificial reference ncGenome of well-studied ncRNA biotypes. The test ncGenome was incorporated as a first-step 'pre-alignment' in a two-step approach (schematically illustrated in Figure 3B), which removes reads mapping to known ncRNAs from the standard alignment. In particular, reads corresponding to multi-copy ncRNA genes are recovered and their total read-count enumerated by the non-redundant test ncGenome instead of becoming flagged as multi-mapping (and thus commonly filtered) in a standard genomic alignment. In one practical example, presented in Figure 3C, this pre-alignment (performed allowing 0–3 mismatches) identifies up to 45% of multi-mapping genomic RNAseq reads, and up to 58% of ncRNAseq. This proportion can undoubtedly be further increased by relaxing the curation of the ncGenome (e.g., to include transcripts of unknown biological function). Relaxation of the alignment parameters to allow >2 mismatches had little effect on the pre-alignment. Whether the remaining multi-mapping reads result from modified transcripts, which cause misincorporation of nucleotides during RT, for example [19,20]), or from insufficient annotation of the non-coding transcriptome is at present unknown.

Consistent with their estimated proportion of the total RNA, up to 30% of the total reads of ncRNAseq libraries align to the test ncGenome, depending on the number of allowed mismatches (compared with ~18% in standard RNAseq; Figure 3C). Among these, the highest read counts (when normalized to transcript length) belong to well-established, abundant transcripts such as the signal recognition particle 7SL RNA, the transcription regulator 7SK RNA, all snRNA components of the spliceosome, Y RNAs, the mitochondrial RNA processing component RMRP and the ribonuclease component RPPH1. Of the snoRNAs, the highest read counts belong to U3 (C/D class SNORD3 gene family) and U17 (H/ACA class SNORA73 gene family). These snoRNAs are needed in large quantities to assist in the processing of rRNA [21,22] and are thus transcribed from independent promoters [23,24]. The correct identification of known abundant ncRNAs supports the physiological relevance of the ncRNAseq-ncGenome pipeline. Please note that rRNA transcripts are not included in this analysis, as they only result from a minor non-depleted carryover and are thus neither reliable nor informative.

In addition to benefiting studies of the non-coding transcriptome, the ncRNAseq protocol also retains short coding cDNAs such as those arising from a reverse transcription (RT) priming site close to mRNA 5′ ends. The added reads allow better coverage of the often

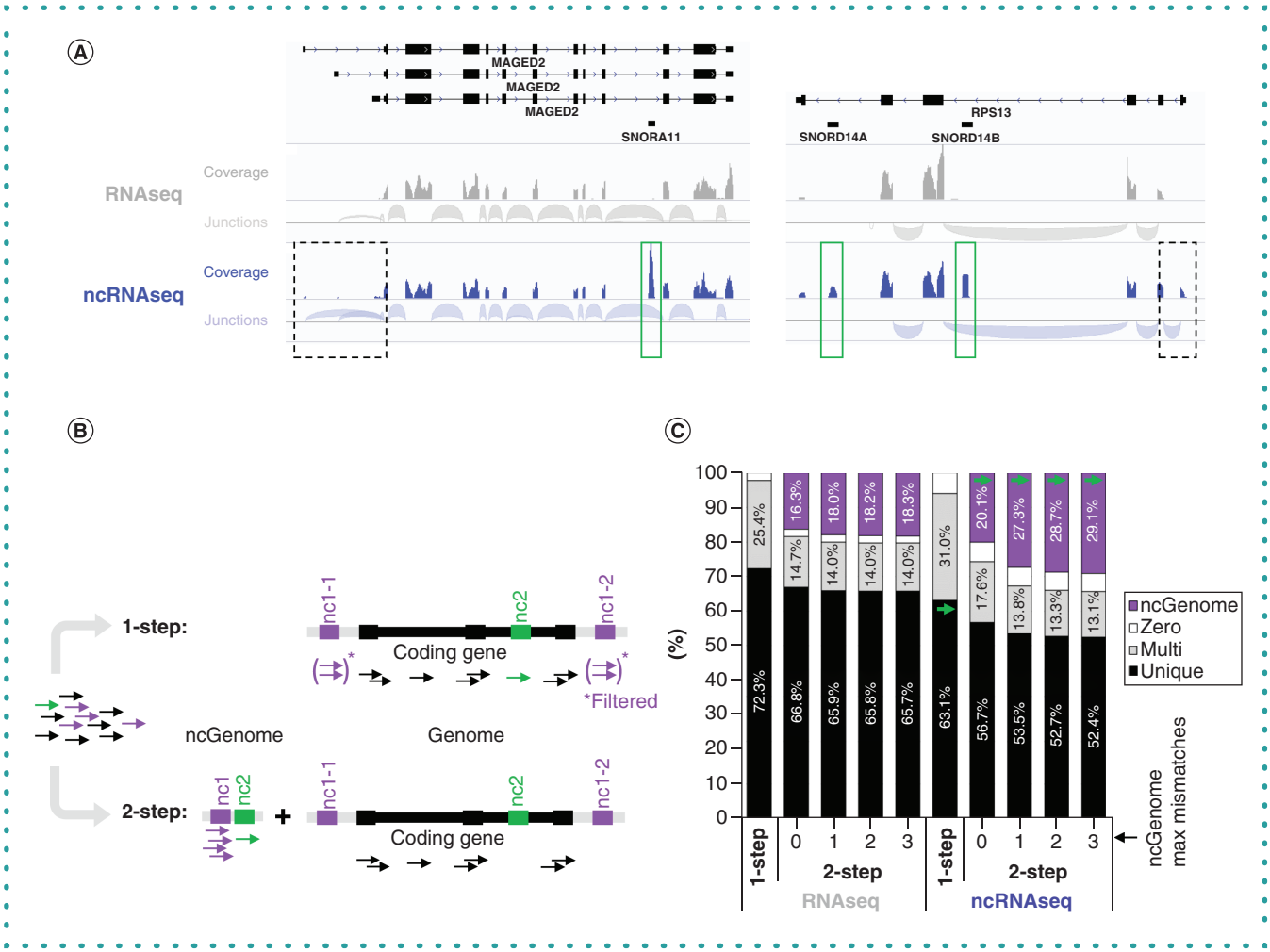

Figure 3. A two-step alignment improves analysis of non-coding RNAs. (A) Integrative Genomics Viewer (IGV) genome browser screenshots of two loci following a *Rsubread*/*subjunc* alignment of RNAseq and ncRNAseq of the same sample. Tracks from both methods are scaled identically. Arches designate exon-exon junctions. Boxes frame read coverage found in ncRNAseq but not in RNAseq (solid green: reads mapping to snoRNAs; dashed black: reads mapping to the 5′ end of the coding transcript). **(B & C)** RNAseq and ncRNAseq reads of the same RNA sample were either aligned directly to the human genome or first pre-aligned to a custom non-redundant non-coding genome (ncGenome), passing only the non-aligning reads through to the standard genomic alignment. The fate of reads belonging to a single-copy ncRNA (green), multi-copy ncRNA (purple: split between gene copies or filtered) and a coding gene (black) is schematically represented in **(B)**. **(C)** summarizes the proportion of reads mapping to the ncGenome or mapping once (unique), more than once (multi) or remain unmapped (zero) to the standard genome in each analysis, when ncGenome alignment was carried out allowing 0–3 mismatches. In the two-step pipeline, reads belonging to a single-copy ncRNA (green) relocate from unique reads to ncGenome, which also recovers the non-uniquely mapping reads belonging to a multi-copy ncRNA (purple), which are otherwise potentially lost or arbitrarily distributed.

undersequenced first exon. Figure 3A exemplifies how in genes with short exons, such as *RPS13* and *MAGED2*, improved 5′ coverage is even sufficient to aid annotation of alternative splicing and transcription start sites (dashed black boxes) without requiring additional sequencing depth. This biochemically driven benefit is unaffected by read pre-alignment, which also does not affect the overall number of identified exon-exon junctions (<0.1% difference). Coding genes might also be affected by ncRNAseq if they host intronic ncRNAs. In this case, a pre-alignment of ncRNAseq to an ncGenome might be important to avoid miscounting hosted ncRNAs in host-gene counts. By and large, in these data such miscounting is negligible, as both mRNA and intronic ncRNA typically arise from the same primary transcript, and thus generate similar processed transcript numbers (although differential stability may alter their steady-state stoichiometry), so the relative contribution of the short ncRNA to the long host gene coverage is minimal. This balance might be different in other systems.

Finally, as a proof of concept the ncRNAseq-ncGenome pipeline was applied to RNA from cytosolic and nuclear fractions of the most commonly studied breast cancer cell lines MCF-7 and MDA-MB-231. The analysis, presented in Figure 4, correctly identifies snoRNAs as enriched in the nuclear fraction ($\sim 2^3$-fold shift in median read count) and tRNAs as enriched in the cytoplasm ($\sim 2^4$-fold shift in median read count). Even larger-fold enrichments were observed for Y1 RNA, known to sequester the Ro autoantigen in the cytosol [25], and for the

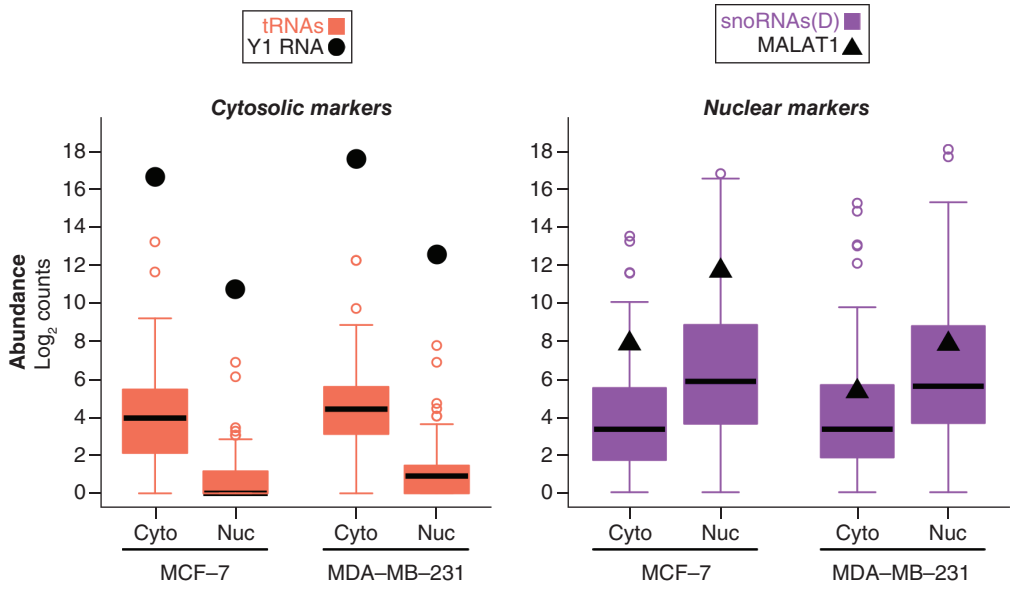

Figure 4. ncRNAseq-ncGenome pipeline correctly identifies markers of cytosolic and nuclear fractions. RNA from nuclear and cytosolic fractions of two breast cancer cell lines (MCF-7 and MDA-MB-231) was sequenced using the ncRNAseq modified protocol, and the resulting reads were aligned to the ncGenome. The distribution of $\log_2$ counts of tRNAs and snoRNAs is indicated in orange or purple, respectively, as is the abundance of known cytosolic (Y1 RNA) and nuclear (MALAT1 lncRNA) markers.

nuclear accumulation of the lncRNA MALAT1 [26]. Reassuringly, the higher levels of MALAT1 observed in the estrogen-receptor-positive MCF-7 compared with the estrogen-receptor-negative MDA-MB-231 are in agreement with estrogen-dependent enhancement of MALAT1 expression, for example, in prostate cancer and osteosarcoma cell lines [27,28]. Similar distributions were obtained in fractionation of seven additional cell lines (data not shown).

Conclusion & future perspective

In summary, we present a simple and inexpensive modification to the standard Illumina TruSeq protocol that allows the biochemical capture of mid-sized RNAs and cDNAs and enhances 5′ coverage of coding genes. In conjunction with a custom non-coding genome/transcriptome reference, it allows the enumeration of ncRNA expression in a lossless manner with respect to the coding transcriptome. The joint processing and analysis of the coding and the non-coding transcriptomes not only cut down the cost, labor and required sample material but also alleviate some of the difficulties in normalizing ncRNA expression obtained in isolation using one of several existing methods to sequence specific ncRNA classes, such as tRNAs or snoRNAs [29,30].

Due to their short length, which is insufficient for priming and extension of reverse transcription, short RNAs such as microRNAs or piRNAs are not yet captured using ncRNAseq and have to be sequenced separately (many kits are available for this purpose). Notably, a 3′ adapter can potentially be ligated to total RNA and thus add artificial length to short RNAs (preferentially), whose sequence can then be captured using ncRNAseq and computationally trimmed.

Although we exclusively used human cell lines for this paper, ncRNAseq could benefit the scope and accuracy of analyzing mid-sized RNAs in other systems (e.g., bacterial sRNA [31]) and assist in the analysis of RNA processing, ranging from pentatricopeptide repeat (PPR)-based processing of polycistronic RNA in plants and apicomplexans [32] to mapping non-contiguous rRNA in bacteria [33]. Future analyses and applications can be improved by iterative refinement of ncRNA aligners (to address polymorphisms and nucleotide misincorporation due to base modifications) and annotation of existing databases, as well as by *de novo* analysis of ncRNAseq reads that map non-uniquely—or do not map at all—to canonical reference genomes.

Executive summary

Background

- Despite their abundance and emerging roles in many biological processes (including human disease), mid-sized RNAs (30–300 nt) have not been extensively studied by high-throughput sequencing, mostly due to selective loss in library preparation.

Method

- In order to retain RNAs shorter than 300 nt and the cDNAs they give rise to, the authors reduce the fractionation time of the RNA and increase the efficiency of four precipitation steps in the Illumina TruSeq protocol: twice pre-ligation by adding isopropanol to AMPure beads and twice post-ligation by using a higher AMPure-to-sample ratio.
- The resulting sequencing reads are aligned first to a custom non-redundant curated 'non-coding genome', which allows the user to exert control, for example, on amalgamation of multi-copy genes or separation of non-coding intronic gene counts from coding host gene counts. The non-mapping reads are then passed to a standard genomic alignment in a lossless manner with respect to the coding transcriptome.

Results

- The modified method (ncRNAseq) leads to gains of counts up to >1000-fold for non-coding RNA genes shorter than 300 nt across the entire abundance range with no significant loss or skewing of coding-gene data.
- ncRNA genes with the highest read counts correspond to established abundant RNA families.
- Analysis of RNA from two fractionated cell lines correctly identifies the cytosolic/nuclear distribution of snoRNAs, tRNAs, MALAT1 and Y1 RNA.
- Retaining short cDNAs during library preparation also leads to better coverage of 5′ coding exons, including alternative transcription start sites.

Conclusion & future prospect

- ncRNAseq, a simple and inexpensive modification to the TruSeq protocol, allows routine investigation of gene expression of previously neglected mid-sized RNAs.
- ncRNAseq can enhance sequencing coverage of transcript 5′ ends, critical not only in mapping transcription start sites and alternative splicing events (of interest predominantly in mammalian studies) but also in studies of RNA processing in other classes or phyla.

Supplementary data

To view the supplementary data that accompany this paper please visit the journal website at: www.future-science.com/doi/suppl/10.2144/btn-2021-0035

Author contributions

The ncRNAseq method was conceived and optimized by A Git. Sequencing libraries were prepared by N Minshall and A Git and aligned by I Chernukhin and A Git. Analyses and visualizations were suggested by all authors. A Git drafted the manuscript, and it was edited and confirmed by all authors.

Acknowledgments

The authors wish to thank C Caldas for the use of breast cancer cell lines and C Duncan, J Mata and K Stott for critical reading of the manuscript.

Financial & competing interests disclosure

The study was funded by Breast Cancer Now grant number 2016NovPR816 to A Git. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

Data sharing statement

The manuscript neither reports original results of a clinical trial nor provides secondary analysis of clinical data that has been shared with the authors.

References

1. Xiao Y, Hu J, Yin W. Systematic identification of non-coding RNAs. In: *Advances in Experimental Medicine and Biology*. Springer, New York LLC, NY, USA, 9–18 (2018).
2. Tuck AC, Tollervey D. RNA in pieces. *Trends Genet.* 27(10), 422–432 (2011).

3. Marshall L, White RJ. Non-coding RNA production by RNA polymerase III is implicated in cancer. *Nat. Rev. Cancer* 8(12), 911–914 (2008).
4. Marshall L. Elevated RNA polymerase III transcription drives proliferation and oncogenic transformation. *Cell Cycle* 7(21), 3327–3329 (2008).
5. de Almeida RA, Fraczek MG, Parker S, Delneri D, O'Keefe RT. Non-coding RNAs and disease: the classical ncRNAs make a comeback. *Biochem. Soc. Trans.* 44(4), 1073–1078 (2016).
6. Deogharia M, Majumder M. Guide snoRNAs: drivers or passengers in human disease? *Biology (Basel)* 8(1), 495–501 (2019).
7. Werner S, Schmidt L, Marchand V *et al.* Machine learning of reverse transcription signatures of variegated polymerases allows mapping and discrimination of methylated purines in limited transcriptomes. *Nucleic Acids Res.* 48(7), 3734–3746 (2020).
8. Potapov V, Fu X, Dai N *et al.* Base modifications affecting RNA polymerase and reverse transcriptase fidelity. *Nucleic Acids Res.* 46(11), 5753–5763 (2018).
9. Deschamps-Francoeur G, Simoneau J, Scott MS. Handling multi-mapped reads in RNA-seq. *Comput. Struct. Biotechnol. J.* 18, 1569–1576 (2020).
10. Apollo repository. https://doi.org/10.17863/CAM.66304 (2021).
11. Ensembl. **ftp://ftp.ensembl.org/pub/release-96/fasta/homo_sapiens/ncrna/**
12. GtRNAdb: tRNAscan-SE analysis of complete genomes. **http://gtrnadb.ucsc.edu/genomes/eukaryota/Hsapi19/hg19-mature-tRNAs.fa**
13. EMBOSS explorer. **http://www.bioinformatics.nl/emboss-explorer/**
14. Liao Y, Shi W. Read trimming is not required for mapping and quantification of RNA-seq reads at the gene level. *NAR Genomics Bioinforma.* 2(3), lqaa068 (2020).
15. Liao Y, Smyth GK, Shi W. The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Res.* 47(8), e47 (2019).
16. Li H, Handsaker B, Wysoker A *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16), 2078–2079 (2009).
17. Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* 14(2), 178–192 (2013).
18. Miller DFB, Yan PX, Fang F *et al.* Stranded whole transcriptome RNA-Seq for all RNA types. *Curr. Protoc. Hum. Genet.* 84, 11.14.1–11.14.23 (2015).
19. Kuksa PP, Leung YY, Vandivier LE, Anderson Z, Gregory BD, Wang LS. *In silico* identification of RNA modifications from high-throughput sequencing data using HAMR. *Methods Mol. Biol.* 1562, 211–229 (2017).
20. Vandivier LE, Gregory BD. Reading the epitranscriptome: new techniques and perspectives. *Enzymes* 41, 269–298 (2017).
21. Enright CA, Stuart Maxwell E, Sollner-Webb B. 5′ETS rRNA processing facilitated by four small RNAs: U14, E3, U17, and U3. *RNA* 2(11), 1094–1099 (1996).
22. Dragon F, Compagnone-Post PA, Mitchell BM *et al.* A large nucleolar U3 ribonucleoprotein required for 18S ribosomal RNA biogenesis. *Nature* 417(6892), 967–970 (2002).
23. Jorjani H, Kehr S, Jedlinski DJ *et al.* An updated human snoRNAome. *Nucleic Acids Res.* 44(11), 5068–5082 (2016).
24. Bratkovič T, Bozič J, Rogelj B. Functional diversity of small nucleolar RNAs. *Nucleic Acids Res.* 48(4), 1627–1651 (2020).
25. Sim S, Weinberg DE, Fuchs G, Choi K, Chung J, Wolin SL. The subcellular distribution of an RNA quality control protein, the Ro autoantigen, is regulated by noncoding y RNA binding. *Mol. Biol. Cell* 20(5), 1555–1564 (2009).
26. Hutchinson JN, Ensminger AW, Clemson CM, Lynch CR, Lawrence JB, Chess A. A screen for nuclear transcripts identifies two linked noncoding RNAs associated with SC35 splicing domains. *BMC Genomics* 8, 39 (2007).
27. Aiello A, Bacci L, Re A *et al.* MALAT1 and HOTAIR long non-coding RNAs play opposite role in estrogen-mediated transcriptional regulation in prostate cancer cells. *Sci. Rep.* 6, 38414 (2016).
28. Hu Q, Li S, Chen C, Zhu M, Chen Y, Zhao Z. 17β-Estradiol treatment drives Sp1 to upregulate MALAT-1 expression and epigenetically affects physiological processes in U2OS cells. *Mol. Med. Rep.* 15(3), 1335–1342 (2017).
29. Pinkard O, McFarland S, Sweet T, Coller J. Quantitative tRNA-sequencing uncovers metazoan tissue-specific tRNA regulation. *Nat. Commun. 2020 111.* 11(1), 1–15 (2020).
30. Krishnan P, Ghosh S, Wang B *et al.* Profiling of small nucleolar RNAs by next generation sequencing: potential new players for breast cancer prognosis. *PLoS One* 11(9), e0162622 (2016).
31. Jørgensen MG, Pettersen JS, Kallipolitis BH. sRNA-mediated control in bacteria: an increasing diversity of regulatory mechanisms. *Biochim. Biophys. Acta–Gene Regul. Mech.* 1863(5), 194504 (2020).
32. Hicks JL, Lassadi I, Carpenter EF *et al.* An essential pentatricopeptide repeat protein in the apicomplexan remnant chloroplast. *Cell. Microbiol.* 12, e13108 (2019).
33. Evguenieva-Hackenberg E. Bacterial ribosomal RNA in pieces. *Mol. Microbiol.* 57(2), 318–325 (2005).