



(51) International Patent Classification:

C12Q 1/68 (2018.01) C40B 40/06 (2006.01)

C12P 19/34 (2006.01) C07H 21/04 (2006.01)

(21) International Application Number:

PCT/US2020/034127

(22) International Filing Date:

22 May 2020 (22.05.2020)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

62/851,969 23 May 2019 (23.05.2019) US

(71) Applicant: **THE ROCKEFELLER UNIVERSITY**
[US/US]; 1230 York Avenue, New York, NY 10065 (US).

(72) Inventors: **LIU, Shixin**; c/o The Rockefeller University,
1230 York Avenue, New York, NY 10065 (US). **JU, Xi-**
angwu; c/o The Rockefeller University, 1230 York Avenue,
New York, NY 10065 (US).

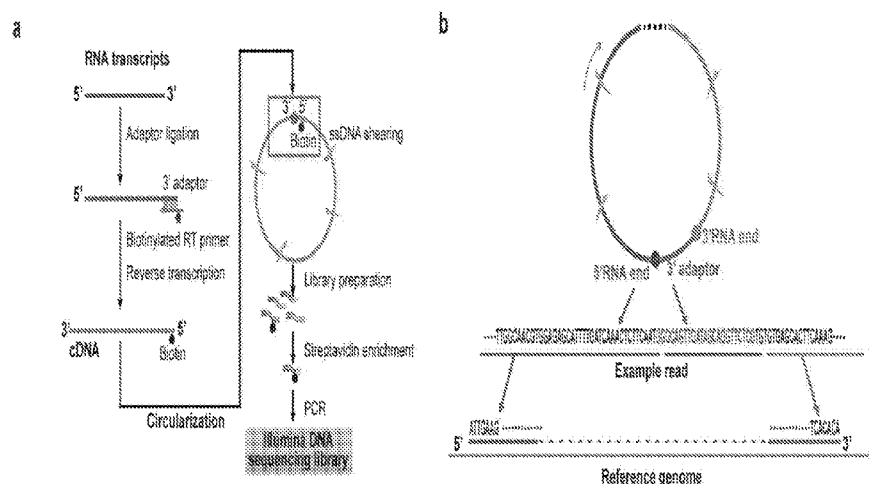
(74) Agent: **SCOLA, Jr., Daniel, A.** et al.; Hoffmann & Baron,
LLP, 6900 Jericho Tpke, Syosset, NY 11791 (US).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM,

(54) Title: METHODS OF FULL-LENGTH RNA PROFILING

FIGURE 1-1



(57) Abstract: The invention relates to the characterization of nucleic acid, particularly RNA, and to full-length transcript profiling of RNA. The invention relates to methods, aspects and kits for determining a transcriptome of an organism, including simultaneously capturing sequences of both the 5' and 3' termini of transcripts so as to characterize all transcripts or RNAs and to provide transcription start and termination sites of a transcriptome with nucleotide resolution.

TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW,
KM, ML, MR, NE, SN, TD, TG).

Published:

- *with international search report (Art. 21(3))*
- *before the expiration of the time limit for amending the
claims and to be republished in the event of receipt of
amendments (Rule 48.2(h))*
- *with sequence listing part of description (Rule 5.2(a))*

METHODS OF FULL-LENGTH RNA PROFILING

STATEMENT OF GOVERNMENT RIGHTS

[0001] This invention was made with government support under grant numbers R00GM107365 and DP2HG010510 awarded by National Institutes of Health (NIH). The government has certain rights in the invention.

FIELD OF THE INVENTION

[0002] The present invention relates generally to the characterization of nucleic acid, particularly RNA, and to full-length transcript profiling of RNA. The invention relates to methods, aspects and kits for simultaneously capturing and determining sequences of both the 5' and 3' termini of RNA transcripts so as to provide transcription start and termination sites of a transcriptome with single-nucleotide resolution.

BACKGROUND

[0003] It has become widely appreciated that RNA is not merely the messenger that relays genetic information from DNA to protein, but also itself carries out diverse and critical regulatory roles in cell physiology (Morris, KV & Mattick, JS (2014) Nat Rev Genet 15, 423-37). The function of an RNA transcript is fundamentally determined by its constituent sequence elements, including those residing at the 5' and 3' ends. Next generation RNA sequencing (RNA-seq) is a revolutionary tool for profiling a transcriptome—the set of all RNA molecules in a cell (Wang, Z et al (2009) Nat Rev Genet 10, 57-63). However, the most commonly used platform for transcriptomic analysis—Illumina-based short-read RNA-seq—requires strand fragmentation, which decouples the 5'-end sequence of an RNA molecule from its 3'-end sequence. As such, the resultant transcriptome map reports the ensemble average of RNA levels at each nucleotide position, but cannot resolve the end-to-end nucleotide composition of each individual transcript. Moreover, due to the inherent lack of end enrichment in the standard RNA-seq protocol, the transcript boundaries are usually poorly defined.

[0004] Various methods based on short-read sequencing have been developed to delineate the 5' or 3' extremities of transcripts (Sharma, CM et al (2010) Nature 464, 250-5 (2010); Wurtzel, O et al (2010) Genome Res 20, 133-41; Dar, D et al (2016) Science 352, aad9822;

Babski, J et al (2016) BMC Genomics 17, 629; Lalanne, JB et al (2018) Cell 173, 749-761 e38; Ettwiller, L et al (2016) BMC Genomics 17, 199; Matteau, D & Rodrigue, S (2015) Methods Mol Biol 1334, 143-59). However, these methods are not capable of concomitantly sequencing both ends of RNA. On the other hand, single-molecule long-read sequencing platforms such as Pacific Biosciences and Oxford Nanopore possess the ability to read an RNA molecule from one end to the other. Nonetheless, their read depth, error rate, and cost still compare unfavorably with the Illumina platform (Goodwin, S et al (2016) Nat Rev Genet 17, 333-51). Moreover, these long-read platforms typically select RNA from a limited size group, thus unable to cover the whole transcriptome—from short non-coding RNA (ncRNA) to long polycistronic messenger RNA (mRNA)—in a single assay.

[0005] Due to their small size and limited splicing, prokaryotic transcriptomes were once considered simple and well understood compared to their eukaryotic counterparts. This view is rapidly changing thanks to the growing list of RNA-based gene regulatory mechanisms employed by bacteria and archaea (Hor, J, Gorski, SA & Vogel, J (2018) Mol Cell 70, 785-799; Guell, M et al (2011) Nat Rev Microbiol 9, 658-69). However, counterintuitively, prokaryotic transcriptomic analyses have lagged behind eukaryotic ones. In particular, transcription termination sites, which mark the 3' ends of primary transcripts, have remained incompletely annotated even in model organisms such as *Escherichia coli* (Gama-Castro, S et al (2016) Nucleic Acids Res 44, D133-43).

[0006] While intramolecular ligation of 5' and 3' RNA termini allows for simultaneous capture of the sequences of both ends, and could provide a possible strategy for inferring full-length sequences of prokaryotic RNA given the scarcity of introns, the few existing methods that employ this strategy suffer from strong length bias and are therefore not suitable for full transcriptome analysis (Ruan, X & Ruan, Y (2012) Methods Mol Biol 809, 535-62; Pelechano, V, Wei, W & Steinmetz, LM (2013) Nature 497, 127-31). In addition, they rely on 3'-end polyadenylate tails for generating complementary DNA (cDNA), and thus are not readily applicable to prokaryotic transcripts.

[0007] Therefore, there remains a need for improved methods for transcriptome determination and analysis and capable of concomitantly sequencing both ends of RNA. There remains a need for methods and approaches providing simultaneous 5' and 3' end sequences, and enabling the comprehensive profiling of full-length transcripts for an organism including prokaryotes and eukaryotes, particularly in prokaryotes.

SUMMARY OF THE INVENTION

[0008] The invention relates to the characterization of nucleic acid, particularly RNA, and to full-length transcript profiling of RNA. The invention relates to methods, aspects and kits for determining a transcriptome of an organism. The invention method includes capturing and determining sequences of both the 5' and 3' termini of transcripts so as to characterize all transcripts or RNAs and to provide transcription start and termination sites of a transcriptome with nucleotide resolution. The invention provides methods and approaches for simultaneous 5'- and 3'-end capture, denoted herein as SEnd-seq.

[0009] In accordance with the invention, RNA is isolated, adaptors, including one or more adaptor, are ligated to at least the 3' ends of isolated RNA, the 3' end ligated RNAs are converted to complementary DNA, the cDNA is circularized, such as via intramolecular ligation, the ligated cDNA is then fragmented, and target cDNA fragments are enriched and sequenced. In view of the circularization of the cDNA and subsequent fragmentation, fragments comprising the 3'-adapted ends linked to the natural 5' ends are generated and can be sequenced. This provides a fragment sequence comprising the 5' end and transcription start site and also the 3' end and transcription termination site. Comparison of the sequence the 5' end and transcription start site and also the 3' end and transcription termination site to an available genome sequence permits determination of the transcript profile of the cell(s), organism, or subject from which the genomic RNA is isolated.

[00010] Genomic RNA includes all or any of the RNA material of an organism, subject, sample or cell(s).

[00011] In one embodiment, the present invention provides a method of transcript profiling for genomic RNA. This method includes the steps of:

- (a) isolating genomic RNA;
- (b) ligating an adaptor to 3' end of genomic RNA to provide 3'-end ligated genomic RNA;
- (c) converting the 3'-end ligated genomic RNA to cDNA;
- (d) circularizing the cDNA;
- (e) fragmenting the cDNA to provide fragmented cDNA;
- (f) sequencing the fragmented cDNA to obtain 5' - and 3' - end sequences; and
- (g) mapping the sequences to a reference genome to provide a full-length transcript profile of genomic RNA.

[00012] In one embodiment, the present invention provides a method of transcript profiling for genomic RNA. This method includes the steps of:

- (a) isolating genomic RNA;
- (b) ligating an adaptor to 3' end of genomic RNA to provide 3'-end ligated genomic RNA;

- (c) converting the 3'-end ligated genomic RNA to cDNA;
- (d) circularizing the cDNA;
- (e) fragmenting the cDNA;
- (f) enriching the cDNA fragments containing the 5'-3' junction;
- (g) sequencing the enriched cDNA fragments to obtain 5'- and 3'-end sequences; and
- (h) mapping the sequences to a reference genome to provide a full-length transcript profile of genomic RNA.

[00013] In accordance with the method, a method of full-length transcript profiling is provided. In accordance with the method, a method of full-length transcript profiling of primary transcript RNA is provided. In accordance with the method, a method of full-length transcript profiling for processed transcripts is provided. Processed transcripts include modified transcript RNA that has been processed to yield mature RNA products, including mRNA, tRNA, and rRNA. This processing includes 5' processing, 3' processing, and cleavage.

[00014] The method of the invention may further comprise enriching for primary transcripts wherein the method additionally comprises particularly enriching for or isolating 5'-triphosphorylated primary RNA, for instance by capping 5'-triphosphorylated primary RNA and isolating capped 5'-triphosphorylated primary RNA. The method of the invention may further comprise enriching for primary transcripts from the genomic RNA wherein the method additionally comprises capping 5'-triphosphorylated primary RNA, and isolating capped 5'-triphosphorylated primary RNA. In an aspect, the method further comprises: enriching primary transcripts from the genomic RNA obtained from step (a); wherein ligating an adaptor to the 3' end of genomic RNA of step (b) further comprises capping 5'-triphosphorylated primary RNA, and isolating the capped 5'-triphosphorylated primary RNA.

[00015] In an embodiment of the method, capping comprises capping with a moiety which can facilitate isolation or enrichment. In an aspect of the method, capping comprises capping with a biotin moiety. In one such aspect, the biotin capped 5'-triphosphorylated primary RNA is isolated or enriched via the biotin label.

[00016] In one such aspect, the biotin capped 5'-triphosphorylated primary RNA is isolated or enriched via streptavidin or another biotin binder or moiety capable of binding or associating specifically with biotin. In accordance with this method, isolating the capped 5'-triphosphorylated primary RNA may comprise isolating the primary RNA with streptavidin-coated magnetic beads.

[00017] The method of the invention may further comprise enriching for processed transcripts wherein the method additionally comprises ligating adaptors, including one or more adaptor,

to the 5' ends of isolated RNA, such that adaptors are ligated to each and both of the 3' ends and the 5' ends. In a particular aspect, unique adaptors are ligated to each of the 3' ends and the 5' ends. The method thus further comprises: enriching processed transcripts from the genomic RNA, such as obtained from step (a); wherein ligating an adaptor to 3' end of genomic RNA of step (b) comprises ligating adaptors to the 5' ends and 3' ends of genomic RNA of step (b), and converting genomic RNA to cDNA of step (c) comprises converting end-ligated processed RNA to cDNA. In an aspect of the method, enriching processed RNA comprises ligating 5'-monophosphorylated processed RNA to a 5' adaptor.

[00018] In accordance with the methods of the invention, the genomic RNA may comprise prokaryotic or eukaryotic RNA or viral RNA. The genomic RNA may comprise eukaryotic genomic RNA. The genomic RNA may comprise prokaryotic genomic RNA. The genomic RNA may comprise viral RNA. The eukaryote may be an animal, plant, fungus or protist. The eukaryote from which eukaryotic genomic RNA is isolated may be a mammal, such as a human, a cow, a chicken, a horse, a bat, a rat, a mouse, a laboratory animal. The prokaryote from which prokaryotic genomic RNA is isolated may be a bacteria or an archaea. The genomic RNA may be from a virus and may be viral genomic RNA. The virus from which viral genomic RNA is isolated may be an RNA virus.

[00019] The methods include wherein the prokaryotic genomic RNA comprises bacterial genomic RNA. The methods include wherein the prokaryotic genomic RNA comprises *Escherichia coli* genomic RNA. The methods include wherein the prokaryotic genomic RNA comprises pathogenic bacterial genomic RNA. The methods include wherein the prokaryotic genomic RNA comprises genomic RNA from drug resistant bacteria, multi-drug resistant bacteria, antibiotic-resistant bacteria. The methods include wherein the prokaryotic genomic RNA comprises *Mycobacterium* genomic RNA. The methods include wherein the prokaryotic genomic RNA comprises *Mycobacterium tuberculosis* genomic RNA. The methods include wherein the prokaryotic genomic RNA comprises *Borrelia burgdorferi* genomic RNA.

[00020] The methods include wherein the genomic RNA comprises genomic RNA from at least one prokaryotic species. The methods include wherein the genomic RNA comprises genomic RNA from one or more prokaryotic species. The methods include wherein the genomic RNA comprises genomic RNA from a microbiome. The methods include wherein the microbiome comprises gut, skin, animal rumen, or plant associated microbiomes. The methods include wherein the genomic RNA comprises genomic RNA from a bacterial biofilm.

[00021] In accordance with the methods, the sequences obtained may be compared to reference sequences, such as a reference genome. The sequences may be mapped to a reference

genome to provide a full-length transcript profile of genomic RNA. In accordance with the methods of the invention, mapping the sequences to the reference genome may comprise: merging paired-end reads into single-end reads; and inferring full-length sequences by mapping to the reference genome.

[00022] In accordance with the methods, mapping the sequences to the reference genome may further comprise identification of transcription start sites (TSS). In accordance with the methods, mapping the sequences to the reference genome may further comprise identification of transcription termination sites (TTS).

[00023] In accordance with the methods, circularizing the cDNA may comprise contacting the cDNA of step (b) with a single-stranded nucleic acid ligase. A single-stranded nucleic acid ligase may particularly include a ligase which is capable of ligating single-stranded RNA and DNA, and may particularly include a ligase which prefers single-stranded ligations, or preferentially ligates single-stranded molecules, prefers single-stranded substrates for ligation, or does not prefer double stranded substrates for ligation. In an aspect, the single-stranded nucleic acid ligase does not ligate double-stranded nucleic acid.

[00024] In an aspect, the single-stranded nucleic acid ligase provides more than 50% intramolecular ligation efficiency. In an aspect, the single-stranded nucleic acid ligase provides more than 50% intramolecular ligation efficiency. In an aspect, the single-stranded nucleic acid ligase provides more than 60% intramolecular ligation efficiency. In an aspect, the single-stranded nucleic acid ligase provides more than 70% intramolecular ligation efficiency. In an aspect, the single-stranded nucleic acid ligase provides more than 75% intramolecular ligation efficiency. In an aspect, the single-stranded nucleic acid ligase provides more than 80% intramolecular ligation efficiency. In an aspect, the single-stranded nucleic acid ligase provides more than 85% intramolecular ligation efficiency. In an aspect, the single-stranded nucleic acid ligase provides more than 90% intramolecular ligation efficiency. In an aspect, the single-stranded nucleic acid ligase provides more than 95% intramolecular ligation efficiency. In an aspect, the single-stranded nucleic acid ligase provides at least 50%, at least 60%, at least 65%, at least 70%, at least 75%, at least 80%, at least 85%, at least 90%, at least 95% intramolecular ligation efficiency.

[00025] In an embodiment of the methods, the single-stranded nucleic acid ligase comprises an RNA ligase. In an embodiment of the methods, single-stranded nucleic acid ligase comprises a thermostable RNA ligase. In an embodiment of the methods, the single-stranded nucleic acid ligase comprises Ts2126 ligase.

[00026] Methods are provided wherein the generated cDNA comprises a tag. The generated cDNA may comprise a tag by virtue of the 3' adaptor, by virtue of the 5' adaptor, or by virtue of the 3' and 5' adaptors. The adaptors may include a tag or label, such as a particular sequence or a lag or labeled nucleotide. In aspects the tag may be a label or molecule which can facilitate isolation, quantitation, or characterization. In an aspect, the tag may be biotin.

[00027] Methods are provided herein wherein fragmented cDNA is single-stranded DNA and is converted to double-stranded DNA.

[00028] In accordance with the methods hereof, spike-in RNA may be added prior to step (b). The spike-in RNA may be utilized, for example as a standard, or to confirm or assess the fidelity and/or accuracy of the methods.

[00029] In accordance with the methods hereof, certain RNAs or RNA types may be preferentially eliminated or removed prior to step (b). Thus, after isolation of genomic RNA and prior to ligating one or more adaptor certain RNAs or RNA types may be particularly or preferentially eliminated. In one such aspect, highly prevalent RNAs, particularly those which are not of interest or relevance in the method and transcriptome analysis being conducted, may be selectively eliminated or removed. Examples include highly prevalent RNA such as ribosomal RNA (rRNA) which can account for up to 80% of total cellular RNA, such as in mammalian cells or samples. Alternatively, tRNA, bacterial rRNA, mitochondrial rRNA, chloroplast rRNA may be depleted or preferentially removed prior to analysis of the genomic RNA. In mammalian cells and samples, globin mRNA, which is a highly prevalent RNA, may be removed.

[00030] In aspects of the methods, providing the full-length profile of a genomic transcript comprises identification of at least one transcript for at least one sequencing paired-end read.

[00031] In one embodiment, the present invention provides a method of making a DNA library of RNA transcripts. In accordance with this method, genomic RNA is isolated, an adaptor is ligated to the 3' end of genomic RNA, the 3' end ligated RNA is converted to cDNA, for example utilizing reverse transcription via a primer which specifically hybridizes to the 3' adaptor sequence, the cDNA is circularized and then subsequently fragmented to provide fragmented cDNA. The fragmented cDNA provides a DNA library of RNA transcripts. In aspects of this method, prior to converting the 3' end ligated RNA to cDNA, a 5' adaptor is added and ligated to free 5' ends. In accordance with this aspect, cDNA may include 3'-end ligated RNA and 5'-end ligated RNA. Circularization may then bring 3'-adaptor ligated ends to 5' ends by virtue of circularization thereby connecting the 3' end or transcription termination region with the 5' end or transcription start region. cDNA fragments or cDNA lengths are thus

comprised of 3' end linked to 5' end. In a particular aspect, cDNA fragments or cDNA lengths comprising of 3'-adaptor ligated ends are preferentially isolated by virtue of a tag or label in the 3' adaptor or in the primer for cDNA generation or reverse transcriptase primer.

[00032] This method includes the steps of: (a) isolating genomic RNA; (b) ligating an adaptor to 3' end of genomic RNA to provide 3'-end ligated genomic RNA; (c) converting the 3'-end ligated genomic RNA to cDNA; (d) circularizing the cDNA; and (e) fragmenting cDNA to provide fragmented cDNA, wherein the fragmented cDNA includes the DNA library of RNA transcripts.

[00033] The invention provides a method of making a DNA library of RNA transcripts, said method comprising the steps of:

- (a) isolating genomic RNA;
- (b) ligating an adaptor to 3' end of genomic RNA to provide 3'-end ligated genomic RNA;
- (c) converting the 3'-end ligated genomic RNA to cDNA;
- (d) circularizing the cDNA; and
- (e) fragmenting cDNA to provide fragmented cDNA.

[00034] The invention provides a method of making a DNA library of RNA transcripts, said method comprising the steps of:

- (a) isolating genomic RNA;
- (b) ligating an adaptor to 3' end of genomic RNA to provide 3'-end ligated genomic RNA;
- (c) converting the 3'-end ligated genomic RNA to cDNA;
- (d) circularizing the cDNA;
- (e) fragmenting the cDNA to provide fragmented cDNA; and
- (f) enriching the cDNA fragments containing the 5'-3' junction.

[00035] The method may further comprise incorporating the fragmented cDNA and/or the cDNA fragments containing the containing the 5'-3' junction or the 5'-3' junctions into a library system, including wherein the fragments are inserted into a vector(s) to generate a library of fragments representing RNA transcripts.

[00036] The method includes wherein fragmented cDNA is ssDNA and is converted to dsDNA.

[00037] The method includes wherein an adaptor is ligated to the fragmented cDNA.

[00038] The method includes wherein fragmented cDNA is amplified.

[00039] The method includes wherein spike-in RNA is added prior to step (b).

[00040] The method or methods may further comprise one or more step of enriching primary RNA from the genomic RNA prior to ligating any one or more adaptor. The method or methods may further comprise enriching primary RNA from the genomic RNA obtained from step (a); wherein ligating an adaptor to the 3' end of genomic RNA of step (b) comprises ligating an adaptor to the 3' end of primary RNA to provide 3'-end ligated primary transcripts, and converting genomic RNA to cDNA of step (c) comprises converting 3'-end ligated primary RNA to cDNA.

[00041] In an aspect of the method(s), enriching primary transcripts comprises capping 5'-triphosphorylated primary RNA and isolating the capped 5'-triphosphorylated primary RNA. Capping may comprise addition of a tag or label to the 5'-triphosphorylated RNA. For example, in an embodiment capping comprises capping with a biotin moiety. In an aspect, the biotin moiety may be desthiobiotin.

[00042] The methods may further comprise enriching processed RNA from the genomic RNA obtained from step (a), wherein ligating an adaptor to 3' end of genomic RNA of step (b) comprises ligating adaptors to the 5' ends and 3' ends of genomic RNA of step (b), and converting genomic RNA to cDNA of step (c) comprises converting end-ligated processed RNA to cDNA.

[00043] Enriching processed transcripts may comprise ligating 5'-monophosphorylated processed RNA to a 5' adaptor. In an embodiment, the 5' adaptor includes a label, tag or target sequence which can facilitate isolation, selection, and/or preferential amplification or priming for cDNA generation and/or reverse transcriptase.

[00044] Circularizing the cDNA may comprise contacting the cDNA, such as the cDNA of step (b), with a single-stranded nucleic acid ligase. A single-stranded nucleic acid ligase may particularly include a ligase which is capable of ligating single-stranded RNA and DNA, and may particularly include a ligase which prefers single-stranded ligations, or preferentially ligates single-stranded molecules, prefers single-stranded substrates for ligation, or does not prefer double stranded substrates for ligation. In an aspect, the single-stranded nucleic acid ligase does not ligate double-stranded nucleic acid.

[00045] In an aspect, the single-stranded nucleic acid ligase provides more than 95% intramolecular ligation efficiency. In an aspect, the single-stranded nucleic acid ligase provides at least 50%, at least 60%, at least 65%, at least 70%, at least 75%, at least 80%, at least 85%, at least 90%, at least 95% intramolecular ligation efficiency.

[00046] In an aspect, the single-stranded nucleic acid ligase comprises an RNA ligase. In an aspect, the single-stranded nucleic acid ligase comprises a thermostable RNA ligase. In an aspect, the single-stranded nucleic acid ligase comprises Ts2126 ligase.

[00047] The invention provides kits or assay systems for simultaneous 5'- and 3'-end capture by generating 5'-end and 3'-end linked sequences and for determining a transcriptome of an organism. The kits or assay systems enable capturing and determining sequences of both the 5' and 3' termini of transcripts so as to characterize all transcripts or RNAs and to provide transcription start and termination sites of a transcriptome with nucleotide resolution. In an embodiment, the kits provide components necessary or required for practicing the instant methods. The components may include one or more 3' adaptor, enzymes for ligating the adaptor(s) to the 3' end of isolated RNAs, enzymes for transcribing the isolated adaptor-ligated RNAs and/or for generating cDNA, and enzymes for circularizing the cDNA. In an embodiment, the enzyme for circularizing may be a ligase which is capable of ligating single-stranded RNA and DNA, and may particularly include a ligase which prefers single-stranded ligations, or preferentially ligates single-stranded molecules, prefers single-stranded substrates for ligation, or does not prefer double stranded substrates for ligation. In an embodiment, the ligase is a single-stranded ligase. In an embodiment, the single-stranded ligase does not ligate double-stranded nucleic acid.

[00048] Additional components may include one or more 5' adaptor, enzymes for ligating the adaptor(s) to the 5' end of isolated RNAs. Additional components may include primers capable of binding the 3' adaptor(s) and/or the 5' adaptor(s), such as reverse transcriptase primers. Further components may include appropriate stabilizing buffers and/or enzyme buffers. Other components may include components for RNA isolation. Other components may include one or more capping enzyme. Other components may include one or more binding agent, binding system, detecting agent or detecting system capable of detecting, binding, selecting, purifying the 3'-adaptor and/or 5'-adaptor ligated RNA and/or the circularized DNA comprising 5' end and 3' end such as by virtue of binding to, detecting, or selecting for adaptor sequence, tag, label or cDNA sequence, tag or label such as by virtue of one or more primer and/or adaptor sequence, tag, label. In an embodiment, biotin type labels or tags are employed along with a biotin binder, such as streptavidin.

[00049] Other objects and advantages will become apparent to those skilled in the art from a review of the ensuing detailed description, which proceeds with reference to the following illustrative drawings, and the attendant claims.

DESCRIPTION OF THE FIGURES

[00050] The patent or patent application contains at least one drawing executed in color. Copies of this patent or patent application publication with color drawing(s) will be provided by the U.S. Patent and Trademark Office upon request and payment of the necessary fee.

[00051] Figure 1. Simultaneous capture of 5'- and 3'-end sequences of bacterial transcripts by SEnd-seq. **A**, Workflow of SEnd-seq. **B**, An example read illustrating how to infer the full-length sequence of individual transcripts by extracting correlated 5'- and 3'-end sequences and mapping them to the reference genome. **C**, A sample data track of the log-phase *E. coli* transcriptome showing the comparison between standard RNA-seq and SEnd-seq. Dashed lines highlight the sharp boundaries of transcripts delineated by SEnd-seq, which are obscured in standard RNA-seq. **D**, SEnd-seq reads mapped to the *ssrA* gene in primary, total and processed RNA datasets. **E**, Ratio of *ssrA* transcripts with an intact, unprocessed 5' end in different datasets. **F**, Ratio of *ssrA* transcripts with an intact 3' end in different datasets.

[00052] Figure 2. Identification of transcription start sites (TSS). **A**, Venn diagram showing the number of TSS identified by SEnd-seq for *E. coli* cells growing in log phase versus stationary phase. **B**, Number of TSS located within intergenic regions or inside annotated genes (either in the sense orientation or in the antisense orientation). **C**, Distribution of the distance between an identified TSS and the start codon of its nearest annotated coding region (cutoff is 300 nt). **D**, Motif analysis of the +1 site, -10 element and -35 element from all TSS detected by SEnd-seq in log phase *E. coli* cells. **E**, Distribution of the number of alternative TSS for a given annotated gene. **F**, Log-phase SEnd-seq data track for the *cysK-ptsH-ptsI-crr* operon that shows multiple TSS (P-1 to P-7) and TTS (T-1 to T-3). TSS identified by dRNA-seq is shown on the top for comparison. **G,H**, Bar graphs displaying the differential usage of alternative TSS for the *cysK* (**G**) and *ptsH/I* (**H**) genes during different growth stages. **I**, SEnd-seq data track showing two TSS controlling the expression of the *yajQ* gene. **J**, Bar graphs displaying the amount of *yajQ* transcripts initiated from the upstream versus downstream TSS. Values are normalized to the upstream TSS transcript level for each experimental replicate. Data are mean $\pm$ s.d. from three independent replicates. **K,L**, Histogram of the percentage of detected transcripts initiated from the most downstream TSS for any gene that employ multiple TSS using cells harvested from the log phase (**K**) or stationary phase (**L**) of growth.

[00053] Figure 3. Identification of transcription termination sites (TTS). **A**, Venn diagram showing the number of identified TTS for log versus stationary phase *E. coli* cells. **B**, Distribution of the RNA folding energy for identified TTS sequences (blue bars) compared

with that for sequences of identical length randomly selected from the *E. coli* genome (red bars). **C**, (left) Pie chart showing the fraction of intrinsic and Rho-dependent terminators identified by SEnd-seq. (right) Nucleotide profiles for the 3'-end sequences of intrinsic and Rho-dependent TTS. Data are representative of two independent experiments. **D**, SEnd-seq data track for an example Rho-dependent terminator located downstream of the *fhuA* gene. When treated with the Rho inhibitor bicyclomycin (BCM), the fraction of readthrough transcripts significantly increased. **E**, Predicted secondary structure of the *fhuA* terminator. **F**, Average termination efficiency of all identified Rho-dependent terminators without or with BCM treatment. Error bars denote s.d. Data are representative of two independent experiments. **G**, SEnd-seq data track for an example intrinsic terminator located downstream of the *cspE* gene. **H**, Predicted secondary structure of the *cspE* terminator. **I**, Average termination efficiency of all identified intrinsic terminators without or with BCM treatment. Error bars denote s.d. Data are representative of two independent experiments. **J**, Scatter plot showing the span of termination efficiency for each TTS that is linked to multiple TSS. For example, a data point at 50% means that, for this TTS, the maximal termination efficiency and the minimal efficiency—depending on the choice of TSS—differ by 50%. The black bars indicate median values. **K**, An example SEnd-seq data track illustrating that the alternative usage of TSS can induce differential termination efficiencies at the same TTS. The fractions of readthrough transcripts initiated from any given TSS (P-1 to P-4) are indicated.

[00054] Figure 4. Pervasive bidirectional overlapping TTS revealed by SEnd-seq. A, SEnd-seq data track for an example convergent gene pair (*cfa-ribC*) exhibiting overlapping TTS. Standard RNA-seq data track is shown in green for comparison. (inset) Predicted secondary structure for the overlapping region. Data are representative of three independent experiments. **B**, SEnd-seq data track and predicted secondary structure of an example overlapping TTS between a coding gene (*sppA*; red reads) and a non-coding antisense RNA (blue reads). Data are representative of three independent experiments. **C**, Venn diagram showing the number of overlapping bidirectional terminators identified for log versus stationary phase *E. coli* cells. **D**, Pie chart showing the fraction of overlapping TTS located between a gene pair or between a gene and an antisense ncRNA. **E**, (left) Average termination efficiency for all identified overlapping bidirectional terminators in either orientation (positive direction in pink; negative direction in orange). (right) Average termination efficiency for those bidirectional TTS that are located between a pair of highly expressed genes. Error bars denote s.d. Data are representative of two independent experiments. **F-I**, Distributions of the length (**F**), folding energy (**G**), predicted stem size (**H**) and loop size (**I**) for the overlapping TTS. **J**,

(left) Schematic of the stem-loop structure formed in the overlapping region. (right) Nucleotide profiles for the 5' and 3' flanking sequences of the stem-loop within an overlapping region. Such profiling allows for classification of the overlapping TTS into three categories. **K**, Pie chart showing the fraction of each category described in (**J**).

[00055] Figure 5. Convergent transcription is required for bidirectional termination in vitro. **A**, SEnd-seq data track for the *yoaJ-yeaQ* gene pair showing an overlapping TTS. Data are representative of three independent experiments. **B**, Schematic of DNA templates harboring the *yoaJ-yeaQ* overlapping TTS region that are used for the in vitro transcription assay. **C**, Gel showing the RNA products transcribed from the different templates shown in (**B**) in the absence or presence of NusA. Data are representative of three independent experiments. **D**, Quantification of the fraction of readthrough transcripts for the different templates. Data are mean $\pm$ s.d. from three independent experiments. *P* values were determined by two-sided unpaired Student's *t*-tests. **E,F**, SEnd-seq data track for part of the *yeaQ* gene (**E**) and DNA templates derived from this region that lacks a terminator sequence (**F**). The templates contain either one or two promoters to allow unidirectional or convergent transcription, respectively. **G**, Gel showing predominant readthrough for unidirectional transcription (Forward and Reverse templates) and heterogeneous RNA products for convergent transcription (Dual template). Data are representative of three independent experiments.

[00056] Figure 6. Convergent transcription contributes to bidirectional termination in vivo. **A**, SEnd-seq data track (top) and schematic of in vivo genomic modification (bottom) for the *yccU-hspQ* convergent gene pair. To disrupt *hspQ* transcription, we replaced the promoter and part of the gene body of *hspQ* with two strong intrinsic terminators. Data are representative of three independent experiments. **B**, Predicted secondary structure for the overlapping TTS between *yccU* and *hspQ*. **C**, qPCR results showing the relative abundance of *yccU* readthrough transcripts across the overlapping region when *hspQ* transcription is abolished ($\Delta hspQ$). We also edited genes outside the convergent pair with the same procedure (Δhfq and $\Delta yeaQ$) as controls. Data are mean $\pm$ s.d. from three independent experiments. *P* values were determined by two-sided unpaired Student's *t*-tests. **D**, SEnd-seq data track around the *yccU-hspQ* region for the $\Delta yeaQ$ (top) or $\Delta hspQ$ strain (bottom). The fraction of *yccU* readthrough transcripts for each strain is indicated. Data are representative of two independent experiments. **E**, Model illustrating that head-on collisions between converging RNA polymerases drive bidirectional termination. The overlapping region produces an RNA hairpin that traps the transcription machinery, which is dislodged by another elongation complex traveling from the opposite direction—either through direct physical interaction or via torsional stress accumulated in the

DNA. This process occurs repeatedly, resulting in highly efficient termination in both directions.

[00057] Figure 7 provides additional information on the SEnd-seq workflow. A, A 100-nt ssDNA was circularized by the Ts2126 ligase and visualized on a 10% urea-PAGE gel. Circularized DNA is resistant to exonuclease (Exo I) treatment, while linear DNA is efficiently digested by Exo I. The lack of concatemeric products suggests predominant intramolecular ligation (circularization). **B,** Gel showing the circularization efficiency of a ssDNA ladder ranging from 100 nt to >2,000 nt. **C,** A representative IGV data track illustrating how to generate full-length RNA sequences from raw SEnd-seq reads such as the one shown in Fig. 1b. Data are representative of three independent experiments.

[00058] Figure 8 provides strategies for enriching primary and processed transcripts. A, 5'-triphosphorylated primary transcripts are exclusively capped with desthiobiotin and then isolated by streptavidin beads. To check the efficiency of processed RNA removal, we examined the abundance of the ribosomal RNA (rRNA)—which are predominantly 5' monophosphorylated—in the primary RNA dataset. Only less than 10% of the reads were mapped to rRNA as compared to 80% in the total RNA dataset. **B,** Alternatively, 5'-monophosphorylated processed RNA are exclusively ligated to a 5' adaptor. The rest of the workflow is the same as that shown in Fig. 1a.

[00059] Figure 9. Analysis of the transcriptome datasets yielded by SEnd-seq. A, *E. coli* growth curve indicating the time points at which cells were collected for log-phase and stationary-phase samples. **B,** Distribution of the transcript length (ribosomal RNA removed) recovered from SEnd-seq for log-phase (left) and stationary-phase (right) samples. **C,** Correlation for the coverage to each individual gene between two independent SEnd-seq replicates. **D,** Correlation for individual gene coverage between SEnd-seq and standard RNA-seq datasets. Pearson correlation coefficients are shown. **E,F,** Distribution of the nucleotide identity at the 5' end (**E**) and 3' end (**F**) of transcripts recovered from primary RNA SEnd-seq. The bias for A and G at the 5' end is consistent with their known enrichment at TSS. The enrichment for U near the 3' end is also expected given the U-tract in intrinsic terminators. Data are representative of three independent experiments.

[00060] Figure 10. Evaluation of the performance of SEnd-seq with spike-in RNA. A, Distribution of the spike-in RNA length recovered after SEnd-seq analysis for four different input species. **B,** SEnd-seq data track for a 680-nt spike-in RNA. The anticipated TSS controlled by a phage T7 promoter is shown at the bottom. **C,** Read count of the spike-in RNA in the total, primary and processed RNA datasets. The low count of the spike-in RNA—which

are 5' triphosphorylated—in the Processed column suggests successful depletion of primary RNA from the processed RNA sample. **D**, Correlation between the read count for different spike-in RNA species and their input amount.

[00061] Figure 11. Demonstration of single-nucleotide resolution afforded by SEnd-seq. SEnd-seq detects the major intermediates in the maturation pathway of the 16S rRNA. The 16S rRNA precursor is trimmed by RNase III and subsequently by RNase G on the 5' end (their cleavage sites are separated by 115 nt); its 3' end is trimmed by RNase III and another unknown enzyme (their cleavage sites are separated by 33 nt).

[00062] Figure 12. Comparison between SEnd-seq and other RNA 5'-end mapping methods. **A**, Venn diagram comparing TSS identified by SEnd-seq and by dRNA-seq. **B**, Venn diagram comparing TSS identified by SEnd-seq and by SMRT-Cappable-seq. Datasets obtained with log-phase samples were used for the comparison.

[00063] Figure 13. Validation of TSS by primer-extension assays. **A**, Workflow of the primer-extension assay. **B,C**, Examples of primary SEnd-seq and primer extension results showing consistent RNA 5' ends. Data are representative of two independent experiments. **D**, Distribution of the distance between the TSS position from primary SEnd-seq and that for the same TSS obtained from the primer-extension assay. Out of the 180 TSS tested—38 of which are newly identified sites—the vast majority exhibit an exact match between the two assays.

[00064] Figure 14. Growth-condition-dependent TSS usage revealed by SEnd-seq. **A**, SEnd-seq data track for *gapA* showing the usage of alternative TSS at different growth stages. Data are representative of two independent experiments. **B**, Quantification of the fraction of *gapA* transcripts starting at a given TSS. **C**, Expression level of the *gapA* gene in the log versus stationary phase reported by standard RNA-seq. Data are mean $\pm$ s.d.

[00065] Figure 15. Motif analysis of the sequences around the upstream TSS (A) and downstream TSS (B) for genes that employ multiple start sites.

[00066] Figure 16. Comparison between TTS detected by SEnd-seq and other RNA 3'-end mapping methods. **A**, Venn diagram comparing TTS identified by SEnd-seq and by SMRT-Cappable-seq. **b**, Venn diagram comparing TTS identified by SEnd-seq and by Term-seq (Babski, J et al (2016) BMC Genomics 17, 629).

[00067] Figure 17. SEnd-seq detects premature termination events. **A**, SEnd-seq data track for the *rpoS* gene, which encodes a general stress sigma factor. Consistent with previous results, our data showed that most *rpoS* transcripts prematurely terminate in the 5' UTR during normal log phase growth. In contrast, full-length transcription is greatly stimulated under heat shock conditions. Inhibition of the Rho activity by BCM also suppresses premature

termination. **B**, Number of *rpoS* reads mapped to the coding region over that mapped to the 5' UTR under normal, heat shock, and BCM-treated conditions. Data are shown as mean $\pm$ s.d. of three independent experiments. *P* values were determined by two-sided unpaired Student's *t*-tests.

[00068] Figure 18. Additional characterization of TTS identified by SEnd-seq. **A**, Distribution of termination efficiencies for all TTS identified by SEnd-seq. Log-phase samples were used for the analysis. Data are representative of two independent experiments. **B**, Termination efficiency as a function of the number of uracil residues found at the 3' flank of the hairpin in the absence or presence of the Rho inhibitor BCM. *n* denotes the number of terminators analyzed in each category. Data are mean $\pm$ s.d. *P* values were determined by two-sided unpaired Student's *t*-tests.

[00069] Figure 19. Analysis of transcription units (TU) defined by SEnd-seq. **A**, Length distribution of TU bound by a unique TSS and TTS. **B**, Distribution of the number of genes covered by each TU. **C**, An example of TU with TSS located inside the coding region of an annotated gene (*amtB*). **D**, An example of TU with an intragenic TSS that drives transcription of downstream genes (*dut* and *slmA*). Data are representative of three independent experiments.

[00070] Figure 20. Antisense transcripts detected by SEnd-seq. **A**, SEnd-seq data track showing examples of antisense transcripts (blue reads for the *insL1* gene, red reads for *entD* and *fepA*). **B**, Distribution of the position of detected antisense transcripts with respect to annotated genes. **C**, Length distribution of detected antisense transcripts. The scarcity of transcripts shorter than 100 nt is due to the procedure of sample preparation, which removes most of short RNA species including tRNA.

[00071] Figure 21. Direction arrangement of adjacent genes in the *E. coli* genome. **A**, Possible scenarios for the arrangement of adjacent genes. **B**, Statistics for the above scenarios in the *E. coli* genome.

[00072] Figure 22 provides examples of predicted stem-loop structure formed at the overlapping bidirectional TTS.

[00073] Figure 23. Effect of Rho inhibition on the termination efficiency of overlapping bidirectional TTS. **A**, SEnd-seq data track around the bidirectional TTS between the *ynaJ-uspE* convergent gene pair using primary RNA, total RNA, or total RNA treated with the Rho inhibitor BCM. Data are representative of two independent experiments. **B**, Average termination efficiency of bidirectional TTS featuring a highly expressed gene pair (*n* = 78) in the positive or negative direction for the different samples described in (A). Data are mean $\pm$ s.d.

[00074] Figure 24. Comparison of the termination efficiency of overlapping bidirectional TTS among different exonuclease knockout strains. **A**, SEnd-seq data track around the bidirectional TTS between the *yccU-hspQ* convergent gene pair for the wildtype, *pnp*-, *rnb*-, and *rnr*-knockout strains. Data are representative of two independent experiments. **B**, Termination efficiency of overlapping bidirectional TTS that feature a highly expressed gene pair ($n = 78$) in either orientation for the different strains described in (**A**). Data are mean $\pm$ s.d.

[00075] Figure 25. Example of a strong intrinsic terminator that causes efficient transcription termination in vitro. **A**, A DNA template harboring a T7A2 promoter and the intrinsic terminator downstream of the *gapA* gene was constructed for in vitro transcription. SEnd-seq reveals a high termination efficiency for the *gapA* terminator in vivo. **B**, Predicted secondary structure of the *gapA* terminator showing features of a strong intrinsic terminator (a GC-rich stem-loop and an 8-nt 3' U-tract). **C**, Gel showing the transcription products using the template described in (**A**). Data are representative of three independent experiments. **D**, Quantification of the fraction of readthrough transcripts.

[00076] Figure 26. Examples of bidirectional termination caused by convergent transcription in vitro. **A**, Three DNA templates that all contain the *macB-cspD* overlapping TTS were constructed. The first two templates (Forward and Reverse) contain one T7A2 promoter at either end. In the Dual construct, each end contains a promoter sequence. **B**, Predicted secondary structure of the overlapping bidirectional TTS between *macB* and *cspD*. **C**, Gel showing the size of RNA products from in vitro transcription with the different templates described in (**A**) in the absence or presence of NusA. Data are representative of three independent experiments. **D**, Fraction of readthrough transcripts for the different conditions described in (**C**). Data are mean $\pm$ s.d. *P* values were determined by two-sided unpaired Student's *t*-tests. **E-H**, The same experiments as in (**A-D**) were repeated with the *yfhL-ansP* overlapping TTS. Note that in this case, there is neither a strong 3' U-tract nor a strong 5' A-tract flanking the hairpin as shown in (**F**). Consistently, substantial readthrough was observed for both directions. Convergent transcription enhances bidirectional termination efficiency, which is further improved by NusA. Data are mean $\pm$ s.d. from three independent experiments. *P* values were determined by two-sided unpaired Student's *t*-tests.

[00077] Figure 27. Example demonstrating that convergent transcription contributes to bidirectional termination in vivo. **A**, SEnd-seq data track for the *ynaJ-uspE* convergent gene pair and schematic of in vivo genome editing. **B**, Predicted secondary structure of the overlapping bidirectional TTS between *ynaJ* and *uspE*. **C**, qPCR results showing the relative abundance of *ynaJ* readthrough transcripts across the overlapping region when *uspE*

transcription is abolished ($\Delta uspE$). Alternatively, transcription of *hfq* or *yeaQ* was disrupted as negative controls. Data are mean $\pm$ s.d. from three independent experiments. *P* values were determined by two-sided unpaired Student's *t*-tests. **D,E**, Reciprocal results showing that disrupting *ynaJ* transcription ($\Delta ynaJ$) increases the level of *uspE* readthrough across the overlapping TTS. Data are mean $\pm$ s.d. from three independent experiments. *P* values were determined by two-sided unpaired Student's *t*-tests.

[00078] Figure 28. RNAP occupancy is enriched around overlapping bidirectional TTS.

A,B, Example RNAP ChIP-seq and primary SEnd-seq data tracks showing the enrichment of RNAP occupancy around the overlapping bidirectional TTS between a convergent gene pair [*pepT-roxA* in (**A**), *yccU-hspQ* in (**B**)]. Data are representative of two independent experiments. **C**, Cumulative RNAP ChIP-seq signal around the overlapping bidirectional TTS summed over those sites identified by SEnd-seq using stationary-phase samples. ChIP-seq was conducted using antibodies against the RNAP β or β' subunit.

[00079] Figure 29 provides full-length gels. Boxes indicate the cropped areas used to generate the corresponding figures.

[00080] Figure 30 provides SEnd-seq data track of primary RNA and total RNA isolated from *B. burgdorferi*.

[00081] Figure 31 provides SEnd-seq data track of primary RNA from *B. burgdorferi* cultured at different temperatures.

[00082] Figure 32 provides SEnd-seq data track of primary RNA and total RNA isolated from a *M. tuberculosis* H37Rv strain.

[00083] Figure 33 provides comparison of the number of transcription start sites (TSS) and transcription termination sites (TTS) identified by SEnd-seq between *E. coli* and *Mtb*.

DETAILED DESCRIPTION

[00084] In accordance with the present invention there may be employed conventional molecular biology, microbiology, and recombinant DNA techniques within the skill of the art. Such techniques are explained fully in the literature and will be known and understood by the skilled artisan.

[00085] The present invention relates generally to full-length transcript profiling for genomic RNA. Novel methods and approaches are provided for simultaneous 5'- and 3'-end capture,

denoted herein as SEnd-seq. The general workflow of SEnd-seq is described herein and depicted in the examples and figures provided. Certain aspects of the workflow incorporate known components and/or enzymes which are known in the art and/or available to the skilled artisan. One described or utilized component and/or enzyme may be replaced or substituted with another suitable and known component by the skilled artisan.

[00086] One important step or aspect of the methods herein is the circularization of cDNA by a ligase, such as a single-stranded ligase that strongly favors intramolecular ligation. This step or aspect circularizes nucleic acid, particularly RNA or DNA such as cDNA generating from genomic RNA, of varying lengths, and particularly may circularize distinct lengths with uniformly high efficiencies. Once circularized, the circularized nucleic acid is fragmented such that fragments comprising a 3' end linked to its particular, corresponding and relevant 5' end are provided. After fragmentation, the pieces, fragments or nucleic acids containing the 5'-3' junction may be isolated and prepared for paired-end sequencing. Isolation of the pieces, fragments or nucleic acids containing the 5'-3' junction is facilitated via a label or tag marking or particularly associated with the pieces, fragments or nucleic acids containing the 5'-3' junction. For example, a biotin label or tag may be introduced via one or more primer or adaptor and can be utilized to isolate, select or enrich for cDNA representing RNAs of interest, such as fragments or nucleic acids containing the 5'-3' junction.

[00087] The 5'- and 3'-end sequences are subsequently determined or extracted. The 5'- and 3'-end sequences may be compared to known sequence, such as a genome sequence, and/or may be mapped to a reference genome. Mapping or comparison to known or available sequence(s) permits the determination of the relevant intervening sequence such that a full-length 5'-to-3' sequence can be determined or inferred. The full-length composition of individual transcripts may be inferred by connecting the two termini, particularly the 3'-end and 5'-end termini sequences.

[00088] The methods and systems provided herein are applicable to total RNA or a subset of RNA. In embodiments, workflows and/or methods have been developed to selectively enrich primary (5' triphosphorylated) or processed (5' monophosphorylated) transcripts prior to selecting or isolating cDNA and/or determining sequences.

[00089] In accordance with the invention, a method is provided which includes the steps of: (a) isolating genomic RNA; (b) ligating an adaptor to 3' end of genomic RNA to provide 3'-end ligated genomic RNA; (c) converting the 3'-end ligated genomic RNA to cDNA; (d) circularizing the cDNA; (e) fragmenting cDNA to provide fragmented cDNA; (f) sequencing

the fragmented cDNA to provide 5'- and 3'- end sequences; and (g) mapping the sequences to a reference genome to provide a full-length transcript profile of genomic RNA.

[00090] In accordance with the invention, a method is provided which includes the steps of: (a) isolating genomic RNA; (b) ligating an adaptor to 3' end of genomic RNA to provide 3'-end ligated genomic RNA; (c) converting the 3'-end ligated genomic RNA to cDNA; (d) circularizing the cDNA; (e) fragmenting the cDNA; (f) enriching the cDNA fragments containing the 5'-3' junction; (g) sequencing the enriched cDNA fragments to obtain 5'- and 3'-end sequences; and (h) mapping the sequences to a reference genome to provide a full-length transcript profile of genomic RNA.

[00091] Isolating Genomic RNA

[00092] Genomic RNA is isolated by any method known in the art. Methods and systems, kits, reagents, buffers for RNA isolation are well known and available to the skilled artisan. Exemplary kits, reagents include the PAXgene RNA kit, Tempus RNA, TriZol reagents and kits. As used herein, genomic RNA includes all or any of the RNA material of an organism, subject, sample or cell(s). In some embodiments, the genomic RNA represents a transcriptome. A transcriptome may refer to the all RNAs or just mRNAs, depending on the particular experiment or study. A transcriptome may refer to the sum total of all the messenger RNA molecules expressed from the genes of an organism, or the product of genome expression of an organism, or the complete set of RNA transcripts that are produced by the genome, including under particular circumstances or conditions or in a specific cell, cell(s) or tissue.

[00093] The organism may be a prokaryote or a eukaryote. The genomic RNA may comprise prokaryotic or eukaryotic RNA or viral RNA. The eukaryote may be an animal, plant, fungus or protist. The eukaryote from which eukaryotic genomic RNA is isolated may be a mammal, such as a human, a cow, a chicken, a horse, a bat, a rat, a mouse, a laboratory animal. The organism may be any animal. "Animal" as used herein refers to any animal, including fish, amphibians, reptiles, birds, and mammals, such as mice, rats, rabbits, goats, cats, dogs, cows, apes, and humans. Viral RNA may include a known or unknown human or mammalian pathogen or virus. The virus may be a hepatitis virus, influenza virus, SARS virus, Ebola virus, coronavirus. The virus may be a retrovirus such as HIV. The virus may be a single-stranded virus, such as influenza.

[00094] The prokaryote from which prokaryotic genomic RNA is isolated may be a bacteria or archaea. The genomic RNA described herein includes prokaryotic genomic RNA. The prokaryotic genomic RNA may be isolated from pathogenic bacteria, non-pathogenic bacteria,

drug-resistant or multi-drug resistant bacteria. The prokaryotic genomic RNA may be isolated from known or unknown or suspected pathogenic bacteria.

[00095] In any instance wherein genomic RNA is isolated from an unknown or undetermined organism or a source which may include an unknown or undetermined or uncharacterized or mutant or variant organism or a suspected organism, the sequences determined, such as the 5'- and 3'-end sequences may be mapped to or otherwise compared with a reference genome, a database of known genomes, a genome of a suspected organism, a genome of a standard or non-variant organism.

[00096] The organism may be an antigenically variant organism, such as a pathogen, or may display a high degree of allelic variation, or may have transcriptionally variant or altered transcripts, such as with growth, mutation or in response to certain environments, stressors, agents, which may be of interest for evaluating, characterizing and comparing.

[00097] Examples of prokaryotic genomic RNA includes genomic RNA isolated from any of the various bacteria or other prokaryotic pathogens which infect humans or other animals and/or which are significant from a clinical or disease management standpoint. Bacteria may be gram-positive or gram-negative bacteria. Relevant gram-positive bacteria include but are not limited to the genera *Actinomyces*, *Bacillus*, *Listeria*, *Lactococcus*, *Staphylococcus*, *Streptococcus*, *Enterococcus*, *Mycobacterium*, *Corynebacterium*, and *Clostridium*. Medically relevant species include *Streptococcus pyogenes*, *Streptococcus pneumoniae*, *Staphylococcus aureus*, and *Enterococcus faecalis*. Relevant bacteria or other pathogens may be of the *Escherichia*, *Actinomyces*, *Mycobacterium*, *Acinetobacter*, *Pseudomonas*, *Staphylococcus*, *Helicobacter*, *Neisseria*, *Streptococcus*, *Chlamydia*, *Vibrio*, *Bacilli*, *Clostridia*, *Sporiobacteriaceae* (includes *Borrelia*) and *Hemophilus* species, strains, class, phylum, genera or families. Relevant bacteria or other pathogens may be *Escherichia coli*, *Actinomyces israelii*, *Mycobacterium tuberculosis*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa*, *Staphylococcus aureus*, *Helicobacter pylori*, *Neisseria gonorrhoeae*, *Streptococcus pneumoniae*, *Borrelia burgdorferi* and *Hemophilus influenzae*.

[00098] The genomic RNA may be isolated from one or more prokaryotic species. The genomic RNA may be from a known or unknown prokaryotic or eukaryotic species. The genomic RNA may be from a sample of one or more known or unknown or undetermined species. The genomic RNA may be isolated from more than two, more than three, more than five, more than ten, or more than twenty-five prokaryotic species. In embodiments where genomic RNA is isolated from more than one species, the full-length transcript profile of the more than one species is provided in aggregate according to the methods described herein.

[00099] The genomic RNA may be metagenomics RNA, such as wherein the methods herein are utilized to comprehensively sequence and analyze all genetic material in complex samples. The genomic RNA may be metagenomic RNA, such as from a microbiome. As used herein, microbiome includes all of the microorganisms or a subset of the microorganisms in a particular environment (including the body or part of the body).

[000100] The genomic RNA may be isolated from a microbiome. The microbiome may be gut, skin, animal rumen, or plant associated. In one embodiment, the microbiome is from a human. In one embodiment, the microbiome is from a non-human animal.

[000101] In some embodiments, the genomic RNA initially isolated may be enriched for a subset of RNA(s) or for a particular type of RNA or a size range of RNA, for instance prior to contacting with any one or more adaptor. In some embodiments, certain RNA(s) or types of RNA may be removed, preferentially eliminated, or selected out before analysis, for instance prior to contacting with any one or more adaptor.

[000102] In some embodiments, the genomic RNA is enriched for primary transcripts. Primary transcripts include single-stranded ribonucleic acid (RNA) product that is synthesized by transcription of DNA within an organism. In embodiments where the genomic RNA is enriched for primary transcripts, the primary transcripts are subject to the methods described herein and below to obtain a full-length transcript profile of primary transcripts.

[000103] Enriching primary transcripts from the genomic RNA can be accomplished by any method known in the art. In some embodiments, to aid in the enrichment of primary transcripts, the 5'-triphosphorylated primary RNA may be capped. In an embodiment of the method, capping comprises capping with a moiety which can facilitate isolation or enrichment. In some embodiments, the 5'-triphosphorylated primary RNA may be capped with a tag or a label, such as a capture tag. The capped 5'-triphosphorylated primary RNA may then be isolated by way of the tag or label, such as via the capture tag. This serves to enrich or further enrich for primary RNA. Figure 8A provides an exemplary embodiment of enrichment of primary RNA transcripts. The capture tag may be a biotin or such other label which may be subject to detection or specific binding, such as desthiobiotin (DTB), as described herein and in Figure 8A. In an aspect of the method, capping comprises capping with a biotin moiety. In one such aspect, the biotin capped 5'-triphosphorylated primary RNA is isolated or enriched via the biotin label. In one such aspect, the biotin capped 5'-triphosphorylated primary RNA is isolated or enriched via streptavidin or another biotin binder or moiety capable of binding or associating specifically with biotin. In accordance with this method, isolating the capped 5'-

triphosphorylated primary RNA may comprise isolating the primary RNA with streptavidin-coated magnetic beads.

[000104] As used herein, “adaptor” refers to a nucleic acid sequence that can ligate to or label a nucleic acid sequence. The term may be spelled “adaptor” or “adapter” herein and in the art and refers to the same meaning. An adapter or adaptor, or a linker in genetic engineering, is a short, chemically synthesized, single-stranded or double-stranded sequence or oligonucleotide that can be ligated to the ends of other DNA or RNA molecules

[000105] In embodiments of the invention an adaptor may be ligated to the 3' end of a nucleic acid sequence, or the 5' end of a nucleic acid sequence, or to both the 5' and the 3' end. In a particular embodiment, the adaptor is ligated to an RNA sequence.

[000106] The adaptor may comprise or be comprised of RNA sequence. The adaptor may comprise defined and random sequence. The adaptor may comprise multiple N or random A, C, G and T (or U) sequences, or a string of N sequence. A string of defined sequences or random N sequences may serve as a barcode (Smith AM (2010) *Nucleic Acids Research* 38(13) e142, doi:10.1093/nar/gkq368). The adaptor may include a modified nucleotide. The modified nucleotide may be labelled or tagged for example. In an embodiment, the modified nucleotide may be labeled with a biotin, such as an internal biotin dT or other biotin-type labeled nucleotide. The adaptor may include a DNA chain-terminating nucleotide, such as dideoxycytidine (ddC). In such an instance, the chain terminating nucleotide serves as a nucleoside reverse transcriptase inhibitor and thus prevents reverse transcription to generate for example cDNA via the adaptor, thus all cDNA generated is via a RT primer specific for the adaptor-ligated RNA. This facilitates cDNA generation from a single strand or sense of the RNA. A 3' adaptor may include a 5' phosphate or a 5' phosphorylation modification. A 5' adaptor may include a 5' C3 spacer modification.

[000107] As will be appreciated by those of skill in the art, the length of the adaptor sequence can vary. The length will be selected based on any adaptor barcode or target sequence such as priming sequence for reverse transcription to cDNA, and based on a length which is stable and can be efficiently ligated. In some aspects, the adaptor sequence is about 1 nucleotide to about 100 nucleotides in length. In some aspects, the adaptor sequence is about 10 nucleotides to about 100 nucleotides in length. In other embodiments, the adaptor sequence is about 5 nucleotides to about 80 nucleotides. In other embodiments, the adaptor sequence is about 10 nucleotides to about 60 nucleotides. In other embodiments, the adaptor sequence is about 10 nucleotides to about 40 nucleotides. In other embodiments, the adaptor sequence is about 10 nucleotides to about 30 nucleotides. In other embodiments, the adaptor sequence is about 15

nucleotides to about 40 nucleotides. In other embodiments, the adaptor sequence is about 20 nucleotides to about 30 nucleotides. In other embodiments, the adaptor sequence is about 15 nucleotides to about 30 nucleotides.

[000108] In some embodiments, an adaptor includes a polynucleotide of known sequence that can serve as a primer hybridization site for a reverse transcription reaction, polymerase chain reaction (PCR), or primer extension reaction. The adaptor can also be used for labeling the 3' end of an RNA, providing barcode information to count the ligated RNA molecules, and/or anchoring the 5' end of RNA transcripts after circularization. Such adaptors are known in the art.

[000109] For example, art-known RNA library preparation methods for high throughput sequencing often start with the ligation of 5'- and 3'-adapters to add 'handles' that are used for priming during reverse transcription and PCR (Linsen SE et al (2009) Nat Methods 6:474-476). In methods of the art, adapters may be attached to RNA by ligases such as T4 RNA ligase using a single-stranded adapter ligation approach, a splinted ligation approach or by polyadenylation of the RNA 3' termini followed by 5'-end adapter ligation (Lau NC et al (2001) Science 294:858-862; Linsen SE et al (2009) Nat Methods 6:474-476; Hafner M et al (2011) RNA 17:1697-1712; Pfeffer S et al (2005) Nat Methods 2, 269-276).

[000110] Labels used herein may include any radioactive label, fluorescent label, a dye, a tag, an enzyme, an epitope, a peptide. Labels may include small modifications and RNA aptamers. Small molecule modification tags may include biotin, desthiobiotin and digoxigenin. Apatamers may include folded oligonucleotides that interact with a target molecule such as via electrostatic and hydrophobic interactions. Examples of RNA aptamers include PP7, S1, D8, tobramycin, streptomycin, MS2, Csy4 (H29A), Mango (Gemmill D et al (2020) Biochem Cell Biol 98:31-41). As used herein, a label includes labeling with a capture tag. As used herein, the term "capture tag" refer to a moiety that is capable of specifically binding to a binding partner for the capture tag non-covalently (i.e., is an "affinity tag"). An example of a capture tag includes a biotin moiety. As used herein, the term "biotin moiety" refers to an affinity agent that includes biotin or a biotin analogue such as desthiobiotin, oxybiotin, 2-iminobiotin, diaminobiotin, biotin sulfoxide, biocytin, and particularly wherein the biotin moiety can reversibly bind streptavidin or a streptavidin analogue.

[000111] Examples of pairs of suitable affinity tags/binding partners are numerous and include biotin/streptavidin, biotin/avidin, digoxigenin/anti-digoxigenin antibody, and fluorescein/anti-fluorescein antibody, although many others are known.

[000112] Methods and kits for generating capped RNA are known and available. The 5' end of RNA(s) may be capped with modified nucleotides, with detectable dyes, with enzymatic moieties, or with other tags or labels. Capping systems and methods include the vasccinia capping system (New England Biolabs) and Clean Cap (TriLink). In one embodiment, capping includes capping with a biotin or biotin-type label, such as desthiobiotin.

[000113] In one embodiment, the primary RNA is capped with desthiobiotin and is isolated with a solid support having streptavidin (or streptavidin analogue). In some embodiment, the solid support is streptavidin-coated magnetic beads.

[000114] In some embodiments, the genomic RNA is enriched for processed transcripts. Processed transcripts include modified transcript RNA that has been processed to yield mature RNA products, including mRNA, tRNA, and rRNA. This processing includes 5' processing, 3' processing, and cleavage.

[000115] In embodiments where the genomic RNA is enriched for processed transcripts, the processed transcripts are subject to the methods described herein and below to obtain a full-length transcript profile of processed transcripts. Enriching processed transcripts from the genomic RNA can be accomplished by any method known in the art.

[000116] In some embodiments, enriching processed transcripts from the genomic RNA includes ligating adaptors to the 5' ends and 3' ends of genomic RNA to provide end-ligated processed RNA, and converting end-ligated processed RNA to cDNA. The 5' adaptor may be ligated first after isolation of RNA and prior to ligation of a 3' adaptor. The 5' adaptor may be ligated second after ligation of a 3' adaptor.

[000117] In some embodiments, enriching processed RNA includes ligating 5'-monophosphorylated processed RNA to a 5' adaptor. Figure 8b details an exemplary embodiment of enrichment of processed RNA transcripts. The 5' adaptor may be ligated first after isolation of RNA, particularly in the case of processed RNA, and prior to ligation of a 3' adaptor.

[000118] As used herein, "enriched" refers to a composition wherein an object species has been partially purified such that, on a content basis, the content of the object species in the enriched composition is higher than level of the object species in the starting composition. In accordance with the present disclosure, the object species includes primary transcript or processed transcript, and the starting composition includes a composition of total or all genomic RNA.

[000119] In one embodiment, the concentration of the object species in the enriched composition is at least 1.5, at least 2, at least 3, at least 4, at least 5, at least 10, or at least 15 times higher than the object species in the non-enriched composition.

[000120] In one embodiment, the enriched RNA includes more than 45% enriched RNA, more than 50% enriched RNA, more than 75% enriched RNA, more than 80% enriched RNA, or more than 90% enriched RNA.

[000121] In accordance with the methods hereof, spike-in RNA may be added following or prior to RNA isolation and prior to addition and ligation of one or more adaptor, such as prior to step (b). The spike-in RNA may be utilized, for example as a standard, or to confirm or assess the fidelity and/or accuracy of the methods. For example, a nucleic acid sequence or sequences may be engineered to include a specific promoter sequence, such as a T7 promoter sequence. Addition of a polymerase recognizing the promoter sequence, for example T7 RNA polymerase, will specifically and particularly amplify RNA from the nucleic acid sequence(s). This ensures an amount of the nucleic acid sequence or sequences RNA(s) in a sample.

[000122] In addition or alternatively, certain RNAs or RNA types may be preferentially eliminated or removed prior to addition and ligation of one or more adaptor, such as prior to step (b). Thus, after isolation of genomic RNA and prior to ligating one or more adaptor, certain RNAs or RNA types may be particularly or preferentially eliminated. This will result in an RNA-depleted RNA sample. Highly prevalent RNAs, particularly those which are not of interest or relevance in the method and transcriptome analysis being conducted, may be selectively eliminated or removed. RNAs which are not of interest or relevant may be depleted. Examples include highly prevalent RNA such as ribosomal RNA (rRNA) which can account for up to 80% of total cellular RNA, such as in mammalian cells or samples. Methods and kits for removing rRNA are available and well known in the art. Suitable kits include Ribo-Zero rRNA Removal Kit (Illumina) and TruSeq Stranded Total RNA Gold. Another approach utilizes hybridization of rRNA-specific DNA 'scissor probes' and digestion of rRNA/DNA duplexes with RNase H (Uveno Y et al (2004) Appl Environ Microbiol 70(6):3650-3663). Alternatively or in addition, tRNA, bacterial rRNA, mitochondrial rRNA, chloroplast rRNA may be depleted or preferentially removed prior to analysis of the genomic RNA. Some commercial kits remove mitochondrial rRNA, bacterial rRNA, chloroplast rRNA for example. In mammalian cells and samples, globin mRNA, which is a highly prevalent RNA, may be removed. Kits for globin RNA removal are known and readily available. In instances where the RNA of an infectious agent is of interest, such as a virus, bacteria etc., host RNA(s) may

be depleted. Host RNA may be depleted for example through physiochemical purification, such as via isolation of viral particles to evaluate viral RNA.

[000123] Ligating an adaptor to 3' end of genomic RNA to provide 3'-end ligated genomic RNA

[000124] In accordance with the methods of the invention, the isolated genomic RNA described above and herein is ligated to an adaptor at the 3' end of the RNA to provide 3'-end ligated genomic RNA.

[000125] A similar procedure applies to ligating an adaptor to 5' end of genomic RNA to provide 5'-end ligated genomic RNA. In accordance with the methods of the invention, the isolated RNA, in one embodiment the enriched processed RNA described above and herein, is ligated to an adaptor at the 5' end of the RNA to provide 5'-end ligated genomic RNA.

[000126] Ligating may be accomplished by any known method. In an embodiment ligation is accomplished utilizing a ligase enzyme. In an embodiment, the ligase enzyme is capable of ligating single-stranded RNA and DNA. For example, the RNA described above may be contacted with T4 RNA Ligase.

[000127] Converting the 3'-end ligated genomic RNA to cDNA

[000128] The 3'-end ligated genomic RNA, enriched primary RNA, or enriched processed RNA is converted to cDNA by any known method. In an embodiment, the end-ligated genomic RNA includes 3'-end ligated and also 5'-end ligated RNA. Conversion of RNA to cDNA may be accomplished using any standard or known methods for transcription, particularly proper and efficient transcription, of RNA to its complementary DNA. Such methods include reverse transcription. This can be accomplished by contacting the 3'-end ligated genomic RNA with a reverse transcriptase. Reverse transcriptases (RTs) use an RNA template and a short primer complementary to the 3' end of the RNA to direct the synthesis of the first strand cDNA, which can be used directly as a template for the Polymerase Chain Reaction (PCR).

[000129] RTs catalyze the formation of DNA from an RNA template in reverse transcription. The RT utilized may be a virus or retrovirus reverse transcriptase enzyme. The RT utilized may be a bacterial reverse transcriptase enzyme. Examples of suitable reverse transcriptase includes *Avian myeloblastosis* virus reverse transcriptase, Moloney Murine Leukemia virus reverse transcriptase, Rous sarcoma virus reverse transcriptase, HIV type reverse transcriptase, and the bacterial *Eubacterium rectale* reverse transcriptase.

[000130] The primer used for the reverse transcription, the RT primer, will be capable of specifically binding or otherwise interacting with the one or more adaptor. For example, an RT primer having complementary sequence to and thus capable of specifically binding to and initiating reverse transcription from the 3' adaptor or 3' adaptor sequence is utilized. In instances where a 5' adaptor has also been ligated, a supplementary or distinct RT primer having complementary sequence to and thus capable of specifically binding to and initiating reverse transcription from the 5' adaptor or 5' adaptor sequence is utilized. In some embodiments, the primer used in the reverse transcription reaction is tagged or labeled or carries a sequence or modification which permits selection or isolation, such as by affinity purification. In a preferred embodiment, the primer is a biotinylated reverse transcription primer.

[000131] Circularizing the cDNA

[000132] The cDNA is circularized by ligating the two ends of the same DNA molecule. Ligation may be accomplished by contacting the cDNA with a nucleic acid ligase.

[000133] Circularizing the cDNA may comprise contacting the cDNA with a single-stranded nucleic acid ligase. A single-stranded nucleic acid ligase may particularly include a ligase which is capable of ligating single-stranded RNA and DNA, and may particularly include a ligase which prefers single-stranded ligations, or preferentially ligates single-stranded molecules, prefers single-stranded substrates for ligation, or does not prefer double-stranded substrates for ligation. In an aspect, the single-stranded nucleic acid ligase does not ligate double-stranded nucleic acid.

[000134] In some embodiments, the cDNA is circularized by way of a single-stranded nucleic acid ligase. In some embodiments, the single-stranded nucleic acid ligase does not ligate double-stranded nucleic acid.

[000135] In preferred embodiments the single-stranded nucleic acid ligase provides more than 50% intramolecular ligation efficiency, more than 75% intramolecular ligation efficiency, more than 80% intramolecular ligation efficiency, more than 85% intramolecular ligation efficiency, more than 90% intramolecular ligation efficiency, more than 95% intramolecular ligation efficiency.

[000136] As used herein, “intramolecular ligation” in the context of nucleic acid (polynucleotide) means ligation of two ends of the same nucleic acid molecule to provide a circularized nucleic acid molecule. Therefore, as used herein, “intramolecular ligation efficiency” means the amount of reactions that result in intramolecular ligation and provide a

circularized nucleic acid molecule. For example, a “50% intramolecular ligation efficiency” means that 50% of the ligation reactions are between two ends of the same nucleic acid molecule.

[000137] As used herein, circularization efficiency and intramolecular ligation efficiency are used interchangeably.

[000138] In some embodiments, the single-stranded nucleic acid ligase is an RNA ligase. The RNA ligase may be a thermostable RNA ligase. The RNA ligase may be a T4 RNA ligase. An example of a thermostable RNA ligase suitable for use in the invention includes Ts2126 ligase. This ligase is also known as thermostable RNA ligase 1 from bacteriophage TS2126.

[000139] High intramolecular ligation efficiency of the 5' and 3' RNA termini allows for simultaneous capture of the sequence of both ends, and is an important aspect of the method disclosed herein.

[000140] In some embodiments, the circularized cDNA is biotin labeled. The circularized DNA may be labeled by way of a label or tag included in the RT primer as an example. The circularized DNA may be labeled by way of a modified or labeled nucleotide, either in the RT primer or included in the reverse transcription reaction such that the cDNA(s) are thereby labeled.

[000141] Fragmenting cDNA to provide fragmented cDNA

[000142] The cDNA is fragmented to provide fragmented cDNA. In some embodiments, the peak length of the fragmented cDNA is between 100 nt and 500 nt, 100 nt and 400 nt, 100 nt and 300 nt, or 150 nt and 250 nt. In one embodiment, the fragmented cDNA has a peak length of about 200 nt in length.

[000143] Fragmentation of the cDNA may be accomplished by any method known in the art. Fragmentation may be accomplished utilizing physical fragmentation, enzymatic fragmentation or by chemical shearing. Physical fragmentation includes acoustic shearing, sonication, and hydrodynamic shear methods. The Covaris instrument is an acoustic device for breaking DNA into 100-5000 nt fragments, and small volumes of DNA can be sheared to 150-1000 nt in length. Hydroshear from Digilab utilizes hydrodynamic forces to shear DNA. Nebulizers (Life Tech) can also be used to atomize liquid using compressed air, shearing DNA into 100-3000 nt fragments. Enzymatic methods may utilize DNase I or other restriction endonuclease, such as RNase III, non-specific nuclease or transposase. Enzymatic methods to shear DNA into small pieces include DNase I, a combination of maltose binding protein (MBP)-T7 Endo I and a non-specific nuclease *Vibrio vulnificus* (Vvn), Fragmentase (New

England Biolabs) and Nextera tagmentation technology (Illumina). Chemical fragmentation may be accomplished using heat and divalent metal cation. Chemical shear is typically reserved for the breakup of long fragments and typically performed through the heat digestion of nucleic acid with a divalent metal cation (magnesium or zinc). The length (115-350 nt) can be adjusted by increasing or decreasing the time of incubation.

[000144] In an embodiment, fragmentation is accomplished by shearing. In some embodiments, fragmenting is accomplished by acoustic shearing. In some embodiments, fragmenting is accomplished by acoustic shearing in microTUBE 777 (Covaris, 520045) with Covaris S220 Focused-ultrasonicator under the condition of Peak145 for 90 seconds.

[000145] In some embodiments, the fragmented cDNA is converted to double-stranded DNA.

[000146] Sequencing the fragmented cDNA to provide 5'- and 3'-end sequences

[000147] The fragmented cDNA is sequenced to provide 5'- and 3'-end sequences.

[000148] The fragmented cDNA may be sequenced by any known method. In some embodiments, the fragmented cDNA is tagged prior to sequencing. Any tag known in the art, including a capture tag. Such tag aids in isolation of the fragmented cDNA.

[000149] In a preferred embodiment, the fragmented cDNA is tagged with a biotin moiety.

[000150] In a preferred embodiment the sequencing is performed by a paired-end method. A paired-end method provides paired-end reads of a target sequence. Paired-end reads often include one or more pairs of reads (e.g., two reads, a read mate pair) where each pair of reads is obtained from each end of a nucleic acid fragment that was sequenced.

[000151] Mapping the sequences to a reference genome to provide a full-length transcript profile of genomic RNA

[000152] The sequences obtained by sequencing are mapped a reference genome to provide a full-length transcript profile of genomic RNA. A reference genome may include a complete or partial available genome sequence or a predicted genome sequence or portion thereof. The sequences may be mapped or compared to a database of sequences first to determine or confirm the appropriate or applicable reference genome.

[000153] The reference genome may be the same genome or different genome as compared to the genome being analyzed. As an example, the reference sequence may be a genome sequence determined in the absence of an environmental stimulus or prior to resistance to an agent or therapy. The reference sequence may be a wild type or non-variant sequence and the sequences

derived herein and in accordance with the invention may be from a variant and presumably mutated sequence or organism.

[000154] As used herein, the term “reference genome” can refer to any particular known, sequenced, or characterized genome, whether partial or complete, of any organism which may be used to reference identified sequences from a subject. A reference genome sometimes refers to a segment of a reference genome (e.g., a chromosome or part thereof, e.g., one or more portions of a reference genome). As used herein, a reference sequence or reference genome often is an assembled or partially assembled genomic sequence from an individual or multiple individuals. In some embodiments, a reference genome is an assembled or partially assembled genomic sequence from one or more human individuals. In some embodiments, a reference genome comprises sequences assigned to chromosomes. The term “reference sequence” as used herein refers to one or more polynucleotide sequences of one or more reference samples. In some embodiments reference sequences comprise sequence reads obtained from a reference sample. In some embodiments reference sequences comprise sequence reads, an assembly of reads, and/or a consensus DNA sequence (e.g., a sequence contig). In some embodiments a reference sample is obtained from a reference subject substantially free of a genetic variation (e.g., a genetic variation in question). In some embodiments a reference sample is obtained from a reference subject comprising a known genetic variation. The term “reference” as used herein can refer to a reference genome, a reference sequence, reference sample and/or a reference subject. In some embodiments, sequence reads can be found and/or aligned with sequences in nucleic acid databases known in the art including, for example, GenBank, dbEST, dbSTS, EMBL (European Molecular Biology Laboratory) and DDBJ (DNA Databank of Japan). BLAST or similar tools can be used to search the identified sequences against a sequence database.

[000155] The sequences obtained above are mapped to a reference genome by any method known in the art. In one embodiment, the paired-end reads are merged into single-end reads, and the full-length sequences are inferred by mapping to the reference genome.

[000156] Mapping of the full-length sequences provides identification of transcription start sites (TSS) and transcription termination sites (TTS).

[000157] In some embodiments, spike-in RNA is added to the genomic RNA, enriched primary RNA, or enriched processed RNA to provide a method to detect and correct for error attributed to RNA handling. In some embodiments, the spike-in RNA is added prior to any 3'-end ligation to an adaptor.

[000158] The methods of full-length profiling for genomic RNA described herein do not suffer from the limitations of the prior art, for example RNA length bias. Prior art methods can have a bias towards the identification of short RNA molecules, or can be particularly directed to longer RNA molecules. See Table 1 below for a comparison of the methods described herein, as compared to a recent method of the prior art (Pelechano, V., Wei, W. & Steinmetz, L.M. Nature 497, 127-31 (2013)).

TABLE 1

	Embodiment of the invention [SEnd-seq]	Prior Art [TIF-seq]
Substrate	ssDNA (cDNA)	dsDNA
Circularization efficiency	very high	low to moderate
Percentage of useful data after sequencing	more than 50%	5-10%
Length Bias	low (from ~35nt to at least 4,500nt)	strong bias toward short RNA
Application	prokaryotic and eukaryotic transcriptome	eukaryotic transcriptome (relying on poly(A)-tailed RNA)
Workflow	simple	complicated

[000159] In one embodiment, the present invention provides a method of making a DNA library of RNA transcripts. This method includes the steps of: (a) isolating genomic RNA; (b) ligating an adaptor to 3' end of genomic RNA to provide 3'-end ligated genomic RNA; (c) converting the 3'-end ligated genomic RNA to cDNA; (d) circularizing the cDNA; and (e) fragmenting cDNA to provide fragmented cDNA, wherein the fragmented cDNA includes the DNA library of RNA transcripts.

[000160]

[000161] The invention provides kits or assay systems for simultaneous 5'- and 3'-end capture by generating 5' end and 3' end linked sequences and for determining a transcriptome of an organism. The kits or assay systems enable capturing and determining sequences of both the 5' and 3' termini of transcripts so as to characterize all transcripts or RNAs and to provide transcription start and termination sites of a transcriptome with single-nucleotide resolution. In an embodiment, the kits provide components necessary or required for practicing the instant methods. The components may include one or more 3' adaptor, enzymes for ligating the adaptor(s) to the 3' end of isolated RNAs, enzymes for transcribing the isolated adaptor-ligated

RNAs and/or for generating cDNA, and enzymes for circularizing the cDNA. In an embodiment, the enzyme for circularization may be a ligase which is capable of ligating single-stranded RNA and DNA, and may particularly include a ligase which prefers single-stranded ligations, or preferentially ligates single-stranded molecules, prefers single-stranded substrates for ligation, or does not prefer double-stranded substrates for ligation. In an embodiment, the ligase is a single-stranded ligase. In an embodiment, the single-stranded ligase does not ligate double-stranded nucleic acid.

[000162] Additional components may include one or more 5' adaptor, enzymes for ligating the adaptor(s) to the 5' end of isolated RNAs. Additional components may include primers capable of binding the 3' adaptor(s) and/or the 5' adaptor(s), such as reverse transcription primers. Further components may include appropriate stabilizing buffers and/or enzyme buffers. Other components may include components for RNA isolation. Other components may include one or more capping enzyme. Other components may include one or more binding agent, binding system, detecting agent or detecting system capable of detecting, binding, selecting, purifying the 3' adaptor and/or 5' adaptor ligated RNA and/or the circularized DNA comprising 5' end and 3' end such as by virtue of binding to, detecting, or selecting for adaptor sequence, tag, label or cDNA sequence, tag or label such as by virtue of one or more primer and/or adaptor sequence, tag, label. In an embodiment, biotin type labels or tags are employed along with a biotin binder, such as streptavidin.

[000163] Thus, the invention provides a kit for simultaneous 5'- and 3'-end RNA sequence capture from isolated RNA comprising:

- (a) at least one labeled or tagged adaptor for ligation to the 3' ends of the isolated RNA;
- (b) a ligase capable of ligating the adaptor to the 3' ends of the isolated RNA;
- (c) a primer for reverse transcription of the 3' end ligated RNA, wherein said primer is capable of specifically initiating transcription of the 3' end ligated RNA to generate cDNA;
- (d) a reagent capable of binding to or otherwise having affinity for circularized cDNA derived from the 3' end ligated RNA such that the reagent enriches for or otherwise isolates cDNA fragments comprising the 3' end ligated sequence in combination with its respective 5' end; and
- (e) directions for use of said kit.

[000164] The kit may further or additionally comprise a reverse transcriptase enzyme and/or a ligase capable of circularizing the generated cDNA. The kit may further or additionally comprise reagents and/or buffers appropriate or suitable for any enzyme or ligase component

of the kit. The kit may further comprise a means or one or more component for fragmentation of the cDNA after (c) such that cDNA fragments comprising the 3' end ligated sequence in combination with its respective 5' end are generated. The kit may further or additionally comprise at least one adaptor for ligation to the 5' ends of the isolated RNA and a ligase capable of ligating the adaptor to the 5' ends of the isolated RNA. The kit may further or additionally comprise reagents and optionally enzymes for the sequencing of the cDNA fragments comprising the 3' end ligated sequence in combination with its respective 5' end. Other components of the kit may be determined, or components may be added or deleted or replaced, as applicable by one of skill in the art.

[000165] In the specification, numerous specific details are set forth in order to provide a thorough understanding of the present embodiments. It will be apparent, however, to one having ordinary skill in the art that the specific detail need not be employed to practice the present embodiments. In other instances, well-known materials or methods have not been described in detail in order to avoid obscuring the present embodiments.

[000166] Throughout this specification, quantities are defined by ranges, and by lower and upper boundaries of ranges. Each lower boundary can be combined with each upper boundary to define a range. The lower and upper boundaries should each be taken as a separate element.

[000167] Reference throughout this specification to “one embodiment,” “an embodiment,” “one example,” or “an example” or “one aspect”, “an aspect” means that a particular feature, structure or characteristic described in connection with the embodiment or example is included in at least one embodiment of the present embodiments. Thus, appearances of the phrases “in one embodiment,” “in an embodiment,” “one example,” or “an example” or “one aspect”, “an aspect” in various places throughout this specification are not necessarily all referring to the same embodiment or example. Furthermore, the particular features, structures or characteristics may be combined in any suitable combinations and/or sub-combinations in one or more embodiments or examples. In addition, it is appreciated that the figures provided herewith are for explanation purposes to persons ordinarily skilled in the art and that the drawings are not necessarily drawn to scale.

[000168] As used herein, the terms “comprises,” “comprising,” “includes,” “including,” “has,” “having,” or any other variation thereof, are intended to cover a non-exclusive inclusion. For example, a process, article, or apparatus that comprises a list of elements is not necessarily limited to only those elements but may include other elements not expressly listed or inherent to such process, article, or apparatus.

[000169] Further, unless expressly stated to the contrary, “or” refers to an inclusive “or” and not to an exclusive “or”. For example, a condition A or B is satisfied by any one of the following: A is true (or present) and B is false (or not present), A is false (or not present) and B is true (or present), and both A and B are true (or present).

[000170] Additionally, any examples or illustrations given herein are not to be regarded in any way as restrictions on, limits to, or express definitions of any term or terms with which they are utilized. Instead, these examples or illustrations are to be regarded as being described with respect to one particular embodiment and as being illustrative only. Those of ordinary skill in the art will appreciate that any term or terms with which these examples or illustrations are utilized will encompass other embodiments which may or may not be given therewith or elsewhere in the specification and all such embodiments are intended to be included within the scope of that term or terms. Language designating such nonlimiting examples and illustrations includes, but is not limited to: “for example,” “for instance,” “e.g.,” and “in one embodiment.”

[000171] In this specification, groups of various parameters containing multiple members are described. Within a group of parameters, each member may be combined with any one or more of the other members to make additional sub-groups. For example, if the members of a group are a, b, c, d, and e, additional sub-groups specifically contemplated include any one, two, three, or four of the members, e.g., a and c; a, d, and e; b, c, d, and e; etc.

[000172] The invention may be better understood by reference to the following non-limiting Examples, which are provided as exemplary of the invention. The following examples are presented in order to more fully illustrate the preferred embodiments of the invention and should in no way be construed, however, as limiting the broad scope of the invention.

EXAMPLE 1

FULL-LENGTH RNA PROFILING VIA THE SEnd-seq METHOD REVEALS PERVASIVE BIDIRECTIONAL TRANSCRIPTION TERMINATION IN BACTERIA

[000173] The ability to determine full-length nucleotide composition of individual RNA molecules is essential for understanding the architecture and function of a transcriptome. However, experimental approaches capable of capturing the sequences of both 5' and 3' termini of the same transcript remain scarce. In the present study, simultaneous 5' and 3' end sequencing (SEnd-seq)—a high-throughput and unbiased method that simultaneously maps transcription start and termination sites with single-nucleotide resolution—is presented. Using this method,

a comprehensive view of the *Escherichia coli* transcriptome was obtained, which displays an unexpected level of complexity. SEnd-seq notably expands the catalogue of transcription start sites and termination sites, defines unique transcription units and detects prevalent antisense RNA. Strikingly, the results of the present study unveil widespread overlapping bidirectional terminators located between opposing gene pairs. Furthermore, it has been shown that convergent transcription is a major contributor to highly efficient bidirectional termination both in vitro and in vivo. This finding highlights an underappreciated role of RNA polymerase conflicts in shaping transcript boundaries and suggests an evolutionary strategy for modulating transcriptional output by arranging gene orientation.

[000174] METHODS

[000175] Oligonucleotides used in this study and example are provided in Tables 2-5. Sequences are represented in the 5' to 3' orientation. Table 2 provides example oligonucleotides used for SEnd-seq library preparation. Table 3 provides oligonucleotides for *in vitro* RNA synthesis by *E. coli* RNA polymerase. To incorporate the T7A2 promoter to both ends of a target DNA, the genome DNA template was first amplified by the half promoter set of primers. The cleaned template was further amplified by the two full-length T7A2 promoter primers (T7A2 full-1 forward, T7A2 full-2 reverse), which contain the same promoter sequence. Table 4 provides oligonucleotides for in vitro RNA synthesis by phage T7 RNA polymerase. Table 5 provides oligonucleotides for genome editing and qPCR. The template of kanamycin-resistance cassette flanked by FLP recognition target sites was originally amplified by PCR from the *E. coli* gene deletion strain generated in a previous paper (8) with TR kana stop site set primers, which would yield a strong intrinsic terminator at either end after modification. The different genome target primer sets were then used to amplify the template for promoter deletion. For *pnp*-, *rnr*- and *rnb*- knockout strains, the corresponding primer set was used to directly amplify DNA template for target gene deletion.

[000176] Bacterial strains and growth conditions. *E. coli* K-12 MG1655 and K-12 SIJ_488 (Addgene #68246; a gift from Alex Nielsen) were cultured in LB media (10 g/l tryptone, 5 g/l yeast extract, 10 g/l NaCl, pH 7.4) under aerobic conditions at 37 °C. To inhibit Rho activity, cells were cultured in LB media with 50 µg/ml bicyclomycin (Santa Cruz, sc-391755) at 37 °C for 15 min at indicated growth condition. Δpnp , Δrnb and Δrnr strains were generated using a previously reported protocol based on the arabinose inducible lambda Red recombineering system and the rhamnose inducible flippase recombinase (52). PCR primers (Table 5) were used to amplify the kanamycin-resistant gene in pKD13 and the DNA product was transformed into the K-12 SIJ_488 strain. After selection for positive colonies, the inserted kanamycin-

resistant gene was excised by culturing with L-rhamnose. To knockout a gene in a convergent gene pair, two strong intrinsic terminators were put into the insert DNA to replace the promoter region of the target gene.

[000177] SEnd-seq pipeline:

[000178] Cellular RNA isolation. The overnight culture medium was diluted 1:50 into fresh media and grown to an OD₆₀₀ of 0.4 to 0.6 for the log phase sample or an OD₆₀₀ over 2.0 for the stationary phase sample. *E. coli* cells were quenched by adding 0.5× vol of cold Stop Buffer (5% phenol in ethanol) to the culture medium immediately before harvest and placed on ice for 15 min. Cell pellets were collected by centrifugation (6,000 rpm for 5 min at 4 °C), thoroughly resuspended in 100 µl of lysozyme solution [2 mg/ml in TE buffer (10 mM Tris-HCl and 1 mM EDTA)], and incubated for 2 min. The cells were then immediately lysed by adding 1 ml of TRIzol Reagent (Invitrogen, 15596) and subsequently pipetted vigorously until the solution was clear. After incubation for 5 min at room temperature, 200 µl of chloroform was added and the sample was gently inverted several times until reaching homogeneity. The sample was then incubated for 15 min at room temperature before centrifugation at 12,000 g for 10 min. The upper phase (~600 µl) was gently collected and mixed at a 1:1 ratio with 100% isopropanol. The sample was incubated for 1 hr at -20 °C and then centrifuged at 14,000 rpm for 10 min at 4 °C. The pellet was washed twice with 1 ml of 75% ethanol, air dried for 5 min, and dissolved in nuclease-free water. RNA integrity was assessed with 1% agarose gel and Agilent 2100 Bioanalyzer System.

[000179] 3' adaptor ligation. RNA with or without 5' adaptor ligation (see below) was subjected to 3' adaptor ligation by mixing 12 µl of RNA (<5 µg) with 1 µl of 100 µM 3' adaptor (Table 2), 0.5 µl of 50 mM ATP, 2 µl of dimethyl sulfoxide, 5 µl of 50% PEG8000, 1 µl of RNase Inhibitor (New England BioLabs, M0314), and 1 µl of High Concentration T4 RNA Ligase 1 (New England BioLabs, M0437). After incubation at 23 °C for 5 hr, the reaction was diluted to 40 µl with water and purified twice with 1.5× vol of Agencourt RNAClean XP beads (Beckman Coulter, A63987) to remove excess RNA adaptors. The sample was subsequently eluted in 12 µl of water.

[000180] rRNA removal and reverse transcription. The eluted RNA was subjected to an optional step of rRNA removal with Ribo-Zero rRNA Removal Kit (Illumina, MRZB12424). The RNA was then recovered by ethanol precipitation. 11.5 µl of eluted RNA was incubated with 0.5 µl of 100 µM biotinylated reverse transcription primer (Table 2) and 1 µl of 10 mM Deoxynucleotide Solution Mix (dNTPs) (New England BioLabs, N0447) at 65 °C for 5 min, and then placed on ice for 2 min. 1 µl of the maturase reverse transcriptase from *Eubacterium*

rectale (recombinantly purified from *E. coli*, a gift from Anna Marie Pyle, Yale University; ref. 35), 4 µl of 5× maturase buffer, 2 µl of 100 mM DTT and 0.5 µl of RNase Inhibitor were added to the reaction and incubated at 42 °C for 90 min. The reaction was then terminated by incubation at 85 °C for 10 min. Following reverse transcription, 10 µl of 1 N NaOH solution was added and incubated at 70 °C for 15 min to remove the RNA templates. After neutralization by adding 10 µl of 1 N HCl solution, the reaction was diluted to 100 µl with TE buffer and cleaned twice with 100 µl of TE-saturated phenol:chloroform:isoamyl alcohol (25:24:1, vol/vol) (Thermo Fisher Scientific, 15593031). The cDNA was purified by ethanol precipitation, dissolved in TE buffer and cleaned once with 1.5× vol of Agencourt AMPure XP beads (Beckman Coulter, A63881). The cDNA was then eluted with 30 µl of water and subjected to 5' phosphorylation by adding 2 µl of T4 Polynucleotide Kinase (New England BioLabs, M0201), 4 µl of PNK Reaction Buffer and 4 µl of 10 mM ATP. After incubation at 37 °C for 60 min and 65 °C for 20 min, the cDNA was cleaned with 1.5× vol of AMPure beads again and eluted with 20 µl of 0.1× TE buffer. The cDNA concentration was determined by the Qubit ssDNA Assay Kit (Invitrogen, Q10212).

[000181] Enrichment of primary transcripts. Primary transcripts were enriched following a protocol adapted from a previously published method (8). 5 µg of total RNA was mixed with 5 µl of 10× VCE Buffer (New England BioLabs, M2080) in a total volume of 50 µl, incubated for 2 min at 70 °C, and then placed on ice. 5 µl of 3'-Desthiobiotin-GTP (New England BioLabs, N0761) and 5 µl of Vaccinia virus Capping Enzyme (New England BioLabs, M2080) were added to the reaction and incubated at 37 °C for 30 min. After purification with 1.5× RNAClean beads, the capped RNA was eluted and subjected to 3' adaptor ligation as described above. The RNA was cleaned twice with 1.5× RNAClean beads and then enriched with Hydrophilic Streptavidin Magnetic Beads (New England BioLabs, S1421). After washing thoroughly four times with Binding Buffer (10 mM Tris-HCl pH 7.5, 2 M NaCl, 1 mM EDTA) and three times with Washing Buffer (10 mM Tris-HCl pH 7.5, 0.25 M NaCl, 1 mM EDTA), the RNA was eluted with 26 µl of Biotin Buffer (10 mM Tris-HCl pH 7.5, 0.5 M NaCl, 1 mM EDTA, 1 M biotin) and incubated at 37 °C for 25 min on a rotator. Then 14 µl of Binding Buffer was added and incubated for another 4 min. The RNA was cleaned with 1.5× RNAClean beads and eluted in 12 µl of H₂O. The 5' capped and 3' ligated RNA was reverse transcribed by the maturase as described above.

[000182] Enrichment of processed transcripts. Processed RNA in a total RNA sample was selectively ligated to a 5' adaptor. Briefly, 5 µg of total RNA was incubated for 2 min at 70 °C

and then placed on ice. 1 μ l of 100 μ M 5' adaptor (Table 2), 0.5 μ l of 50 mM ATP, 2 μ l of dimethyl sulfoxide, 5 μ l of 50% PEG8000, 1 μ l of RNase Inhibitor and 1 μ l of High Concentration T4 RNA Ligase 1 were added to the sample. After incubation at 23 °C for 5 hr, the sample was diluted with water and cleaned twice with 1.5 $\times$ vol of Agencourt RNAClean XP beads. After the SEnd-seq pipeline, we used a custom shell script to search for the adaptor-labeled reads, thereby specifically extracting processed RNA ends.

[000183] Circularization. 50 ng of cDNA was mixed with 2 μ l of CutSmart Buffer (New England BioLabs, B7204), 2 μ l of 50 mM MnCl₂, 2 μ l of 0.1 M DTT, 2 μ l of 5 M betaine (Affymetrix, 77507) and 2 μ l of Ts2126 RNA Ligase I (16). The reaction was incubated at 37 °C for 5-16 hr. Subsequently, the reaction was supplemented with 1 μ l of 10 mM dNTPs and diluted to 100 μ l with TE buffer and 0.1% SDS. Then 100 μ l of TE-saturated phenol:chloroform:isoamyl alcohol (25:24:1, vol/vol) was added and incubated for 1 hr with occasional vortexing. After centrifugation, the water phase was cleaned again with phenol:chloroform:isoamyl alcohol. Finally, the circularized cDNA was ethanol precipitated and dissolved in 130 μ l of TE buffer.

[000184] Library preparation. Circularized cDNA was fragmented by acoustic shearing in microTUBE (Covaris, 520045) with Covaris S220 Focused-ultrasonicator under the condition of Peak145 for 90 sec. After ethanol precipitation, the ssDNA was converted to dsDNA by the Second Strand cDNA Synthesis Kit (New England BioLabs, E6114) at 16 °C for 2 hr. The product was cleaned with 1.8 $\times$ vol of AMPure beads and eluted in 50 μ l of 0.1 $\times$ TE buffer. The DNA ends were prepared and ligated to the Illumina sequencing adaptor with the NEBNext Ultra II DNA Library Prep Kit (New England BioLabs, E7645). The ligated product was cleaned twice with 1 $\times$ vol of AMPure beads and eluted in 50 μ l of 0.1 $\times$ TE buffer. Biotin-labeled DNA strands were bound to the Dynabeads M-280 Streptavidin (Invitrogen, 11205D) and cleaned four times with Washing Buffer (5 mM Tris-HCl pH 7.5, 1 M NaCl, 0.5 mM EDTA) and twice with TE buffer. The beads were re-suspended thoroughly with the Q5 High-Fidelity 2 $\times$ Master Mix (New England BioLabs, M0492). The DNA library was then amplified for 13 (total RNA SEnd-seq) to 17 cycles (primary RNA SEnd-seq) following the manufacturer's protocol. The final library was cleaned twice with 1 $\times$ vol (50 μ l) of AMPure beads, and its concentration and size distribution were determined with Agilent 2200 TapeStation (Agilent, 5067-5576).

[000185] Spike-in RNA preparation. A T7 promoter sequence was incorporated upstream of four DNA sequences with different lengths taken from the bacteriophage λ genome. After PCR amplification and gel excision/cleanup, the DNA templates were subjected to in vitro

transcription by T7 RNA Polymerase (New England BioLabs, M0251). DNA was removed by adding 1 μ l of TURBO DNase (Life Technologies, AM2238) and incubated at 37 °C for 15 min. Full-length RNA products were purified by polyacrylamide gel electrophoresis. After cleanup and concentration measurement, all spike-in RNA species were pooled together. Typically the spike-in RNA mix was added to the total bacterial RNA at a mass ratio of 1:1000. **[000186]** RNA-seq. For standard RNA-seq, ~5 μ g of RNA was treated with TURBO DNase and recovered by ethanol precipitation. Ribosomal RNA was depleted with the Ribo-Zero rRNA Removal Kit. The sequencing library was prepared with the TruSeq Stranded mRNA Library Prep Kit (Illumina, RS-122-2101) following the manufacturer's instructions.

[000187] RNAP ChIP-seq. The ChIP-seq workflow is adapted from a previously published ChIP-microarray study (53). Briefly, cells were grown to the stationary stage and crosslinked by the addition of formaldehyde (1% final concentration) with continued shaking at 37 °C for 10 min before quenching with glycine (100 mM final concentration). Cells were then lysed and DNA was sheared by sonication followed by treatment with micrococcal nuclease (New England BioLabs, M0247S) and RNase A (Thermo Fisher Scientific, EN0531). Antibodies against the RNAP β or β' subunit (BioLegend 663903 or 662904) were used for immunoprecipitation. RNAP-DNA crosslinks were enriched by protein A/G beads (Thermo Fisher Scientific, 26159). Enriched immunoprecipitated DNA and input DNA sequencing libraries were prepared with NEBNext Ultra II DNA Library Prep Kit.

[000188] Primer extension assay. ~5 μ g of RNA was treated with TURBO DNase, cleaned three times with phenol:chloroform:isoamyl alcohol (25:24:1, vol/vol), and recovered by ethanol precipitation. Subsequently the RNA was denatured at 70 °C for 2 min and then treated with Terminator 5'-Phosphate-Dependent Exonuclease (Illumina, TER51020) at 30 °C for 1 hr. After ethanol precipitation, the recovered RNA was treated with RppH (New England BioLabs, M0356S) at 37 °C for 1 hr. The RNA was cleaned by 1.5 $\times$ vol of Agencourt RNAClean XP beads (Beckman Coulter, A63987) and ligated to a 5' adaptor as described above. After reaction, the RNA was cleaned with 1.5 $\times$ vol of Agencourt RNAClean XP beads. The eluted RNA was then reverse transcribed to cDNA with pooled RT primers by the maturase. Subsequently, 10 μ l of 1 N NaOH solution was added and incubated at 70 °C for 15 min to remove the RNA templates. The second strand DNA was synthesized with an oligo complementary to the 5' adaptor. The resultant dsDNA was used for sequencing library preparation and sequencing was performed on MiSeq.

[000189] Data analysis. Sequencing data collection and processing. SEnd-seq data were collected by the Illumina MiSeq or NextSeq 500 platform in a paired-end mode (150 nt $\times$ 2).

After quality filter and adaptor trimming, the paired-end reads were merged to single-end reads by using the FLASH software. The correlated 5'-end and 3'-end sequences were extracted by the custom script `fasta_to_paired.sh`. The full-length sequences were inferred by mapping to the reference *E. coli* genome NC_000913.3 by using Bowtie 2. Reads with an insert length greater than 10,000 nt were discarded. For each sample we obtained over 2 million usable reads (i.e., those harboring at least 15 nucleotides on each end of the same transcript). RNA-seq and ChIP-seq data were collected by the Illumina MiSeq or NextSeq 500 platform in a paired-end mode (75 nt ×2). After quality filter, the sequencing data were analyzed by the Rockhopper software (54). The wig files and SAM files were further analyzed by custom Perl scripts. The results were visualized with the Integrative Genome Viewer (IGV).

[000190] Gene coverage quantification. For SEnd-seq data, each read was first mapped to the genome. Each position within the intervening region of the read (from the start site to the end site) was considered as effective coverage. For RNA-seq data, the coverage of each nucleotide position was directly extracted from the wig files generated by the Rockhopper software (54). Gene coverage was quantified by summing the coverage of all nucleotide positions spanned by each gene. Only genes longer than 200 nt were used for the correlation analysis between SEnd-seq and RNA-seq.

[000191] TSS identification. Transcription start sites (TSS) were identified from the primary transcript SEnd-seq data with a custom Perl script. Only positions with more than 10 reads starting at that position and with an increase of at least 50% in read coverage from its upstream to its downstream were retained. Candidate TSS positions within 5 bases in the same orientation were clustered together, and the position with the largest amount of read increase was used as the representative TSS position. Motif analysis around the TSS regions (-40 nt to +1 nt) was performed by MEME (55).

[000192] TTS identification. Based on previous work and our observation that transcripts with intact, unprocessed 3' termini are enriched in the primary RNA SEnd-seq dataset, we reasoned that transcription termination sites (TTS) should be reproducible between the total RNA and primary RNA datasets. In practice, we first identified from the total RNA SEnd-seq data positions with more than 10 reads ending at that position (outside of rRNA genes) and with a reduction of more than 40% in read coverage from its upstream to its downstream. We then cross-checked the site in the primary SEnd-seq dataset and with RNase-knockout strains (Δpnp , Δrnb , Δrnr). Candidate TTS positions within 5 bases in the same orientation were clustered together, and the position with the largest amount of read reduction was used as the representative TTS position. Only the TTS sites identified from at least two samples were used

for further analysis. The terminators are classified into Rho-dependent terminators (those showing a readthrough percentage increase of > 30% upon BCM treatment), intrinsic terminators (those showing a readthrough percentage of < 30% in the control sample, a readthrough percentage increase of < 15% upon BCM treatment, and harboring at least five uracils out of the eight bases in the 3' flank region of the terminator hairpin), or undefined.

[000193] Overlapping bidirectional TTS identification. Overlapping bidirectional termination sites were identified by screening for two opposing TTS with a custom Perl script. Only those with an overlapping region shorter than 60 nt and yielding a stem-loop structure were retained for further analysis. Highly expressed convergent gene pairs were defined as those with > 20 read counts for each gene in the pair.

[000194] RNA secondary structure analysis. The sequence from 45-nt upstream to 9-nt downstream of an identified TTS was used for RNA secondary structure prediction with RNAfold (57) combined with custom Perl scripts.

[000195] Motif analysis. The -45 nt to +9 nt TTS regions and overlapping bidirectional TTS regions were used for motif analysis. Nucleotide logos around TTS were generated by WebLogo (58).

[000196] Transcription unit annotation. Transcription units were identified by a custom Perl script based on the defined TSS, TTS and read coverage. Only those with a continuous coverage of more than 5 reads were retained for further analysis. We also excluded units with a length shorter than 80 nt.

[000197] ChIP-seq data analysis. The RNAP ChIP-seq signal at each nucleotide position was calculated and normalized to the input sample data using a custom script. The normalized ChIP/input ratio was used for downstream analysis.

[000198] Previously deposited datasets. dRNA-seq datasets (SRR1411276 and SRR1411277 for log and stationary phase *E. coli* RNA, respectively) (19) and SMRT-Cappable-seq dataset (GSE117273) (20) were used for comparison with the SEnd-seq results from this study.

[000199] In vitro transcription. DNA templates for T7 RNAP were amplified from the FLuc Control Template (New England BioLabs, E2040S). DNA templates for *E. coli* RNAP were prepared by PCR from the *E. coli* genomic DNA with indicated primer sets (Supplementary Table 5). The T7A2 promoter sequence was incorporated at one or both ends of the template. Purified *E. coli* RNAP and sigma factor σ^{70} holoenzyme (a gift from the Darst Lab at The Rockefeller University) was used for in vitro transcription reactions. The reaction mixture included 4 μ l of 5 $\times$ Reaction Buffer (200 mM Tris-HCl, 600 mM KCl, 40 mM MgCl₂, 4 mM DTT, 0.04% Triton X-100, pH 7.5 at 25 °C), 0.5 μ l of RNase Inhibitor, 0.5 pmol of DNA

template and 2 pmol of *E. coli* RNAP holoenzyme. When applicable, 20 pmol of NusA (a gift from the Landick Lab at University of Wisconsin-Madison) was added to the reaction mixture. The mixture was incubated at 37 °C for 30 min before rNTPs (50 μM each) were added to initiate transcription. After 5 min of reaction (unless noted otherwise), reinitiation of transcription was prevented by adding heparin (Sigma-Aldrich, H4784) to a final concentration of 100 μg/ml. After incubation with 0.3 μl of TURBO DNase for 10 min, the RNA was separated by 5% urea polyacrylamide gel electrophoresis, stained by SYBR Gold Nucleic Acid Gel Stain (Thermo Fisher Scientific, S11494), scanned by Axygen Gel Documentation System (Corning, GD1000), and quantified by ImageJ (National Institutes of Health).

[000200] Quantitative PCR. First-strand cDNA was reverse transcribed from the total RNA of indicated samples with the High-Capacity cDNA Reverse Transcription Kit (Applied Biosystems, 4368813) and strand-specific RT primers (Supplementary Table 5). Control cDNA was reverse transcribed with random primers and the same amount of input RNA. Quantitative RT-PCR was performed using the SYBR Green PCR Master Mix (Applied Biosystems, 4309155) and QuantStudio 6 Flex Real-Time PCR System (Thermo Fisher Scientific). The relative abundance of RNA is represented as the signal ratio between the target transcript and the reference *rnpB* gene from the same sample using the formula: $2^{-(\Delta CT)}$ ($\Delta CT = CT_{\text{target}} - CT_{\text{rnpB}}$; CT stands for cycle threshold).

[000201] Statistics. Data are shown as mean ± s.d. unless noted otherwise. P values were determined by two-sided unpaired Student's t-tests using GraphPad Prism 6. The difference between two groups was considered statistically significant when the P value is less than 0.05 (*P < 0.05; **P < 0.01; ***P < 0.001; ****P < 0.0001; ns, not significant).

[000202] Data availability. SEnd-seq and standard RNA-seq datasets from this study have been deposited in the Gene Expression Omnibus (GEO) with the accession number GSE117737. The custom scripts used in this study are available on Github (github.com/LiuLab-codes/SEnd_seq_analysis). Other data that support the findings of this study are available from the corresponding author upon request.

[000203] RESULTS

[000204] Simultaneous 5'- and 3'-end capture by SEnd-seq. The general workflow of SEnd-seq is depicted in Fig. 1a. The key step is the circularization of cDNA by a single-stranded ligase that strongly favors intramolecular ligation (16) (Figure 7A). Importantly, this step circularizes DNA of varying lengths with uniformly high efficiencies (Figure 7B). After fragmentation, the biotin-labeled pieces containing the 5'-3' junction are isolated and prepared for paired-end sequencing. The 5'- and 3'-end sequences are extracted and mapped to the

reference genome (Figure 1B). The full-length composition of individual transcripts is then inferred by connecting the two termini (Figure 7C). Besides total RNA SEnd-seq, we also developed workflows to selectively enrich primary (5' triphosphorylated) or processed (5' monophosphorylated) transcripts (Figure 8).

[000205] Evaluation of the performance of SEnd-seq. We applied SEnd-seq to *E. coli* cells collected under different growth conditions (Figure 9A, 9B). The read coverage on each gene is highly correlated between SEnd-seq replicates (Figure 9C), and between SEnd-seq and standard RNA-seq (Figure 9D). The transcriptome dataset yielded by SEnd-seq exhibits no severe nucleotide bias at either the 5' or 3' end of RNA (Figure 9E, 9F).

[000206] The advantage of SEnd-seq over standard RNA-seq in mapping the boundaries of individual RNA molecules is apparent in a direct comparison of their respective data tracks (Figure 1C). To assess the ability of SEnd-seq to reproduce the precise ends of input transcripts, we added a mixture of in vitro synthesized RNA to the cellular RNA and subjected them together to SEnd-seq analysis. Correct lengths were recovered for all tested spike-in RNA species (Supplementary Figure 4A-4C), indicating minimal sample deterioration during the procedure. The read coverage on each spike-in RNA species matches the ratio at which it was added to the mixture (Figure 10D), arguing against any significant length bias of SEnd-seq.

[000207] We also demonstrated that SEnd-seq is able to recover the boundaries of endogenous transcripts with single-nucleotide resolution. For example, intact 5' and 3' ends of the 452-nt *ssrA* RNA precursor were enriched in the primary RNA sample, whereas the processed and total RNA datasets predominantly yielded the mature 365-nt *ssrA* species with its termini exactly corresponding to the known RNase cleavage sites (17) (Figure 1D-1F). As another example, the 1,861-nt 16S rRNA precursor and the major intermediates in its maturation pathway were successfully detected by SEnd-seq (Figure 11).

[000208] Identification of transcription start sites. The single-nucleotide resolution afforded by SEnd-seq allows us to precisely annotate transcription start sites (TSS) and termination sites (TTS) in the same assay. Using primary RNA datasets, we identified 4,358 and 4,038 TSS for log-phase and stationary-phase *E. coli* cells respectively, among which 2,884 are common sites (Figure 2A, data not shown). These sites are located both within intergenic regions and inside gene bodies (Figure 2B, 2C). Most of them display a characteristic bacterial promoter sequence in the 5' flank (18) (Figure 2D).

[000209] SEnd-seq not only reproduced the vast majority of TSS previously annotated by other 5'-end mapping methods (19,20), but also identified thousands of TSS unknown until now (Figure 12). A subset of these start sites was selected and validated by primer-extension assays

(Figure 13). We found 2,133 genes that feature alternative TSS upstream of their coding regions (Figure 2E, 2F), indicating that those genes are each controlled by multiple promoters. In many cases, the usage of alternative TSS is dependent on the growth condition (Figure 2G, 2H and Figure 14). For the genes that employ multiple TSS, we analyzed the fraction of transcripts initiated from the upstream TSS versus the downstream TSS [e.g., *yajQ* (Figure 2I, 2J)]. We found that the most downstream TSS (i.e., the one closest to the start codon) tends to make the largest contribution to the overall RNA expression level (Figure 2K, 2L). The upstream and downstream TSS regions share a similar bacterial promoter -10 element, while exhibiting minor differences in the -35 element (Figure 15).

[000210] Identification of transcription termination sites. Two major transcription termination mechanisms have been well documented in bacteria: intrinsic termination that is mediated by a hairpin structure formed in the nascent RNA followed by a U-rich tract, and factor-dependent termination that relies on the Rho ATPase (21). The identification of TTS is more challenging than TSS because of the lack of chemical distinction between bona fide termination sites and processed 3' ends, resulting in much fewer annotated TTS in the existing database. To exclude post-processing cleavage sites, we created single-deletion *E. coli* strains in which each of the three genes (*pnp*, *rnb*, *rnr*) that encodes a major 3'-5' exoribonuclease is knocked out (22). Only those RNA 3' ends that were not affected by any of these knockouts were annotated as TTS, notwithstanding the caveat that these RNases likely play redundant roles. We identified 1,285 TTS that are common between log-phase and stationary-phase *E. coli* cells, as well as 255 growth-stage-specific ones (Figure 3A, data not shown). SEnd-seq recaptures most of the TTS annotated by other 3'-end mapping methods (20, 23), but also finds a large number of previously unknown sites (Figure 16). We found that TTS predominantly reside within intergenic regions (89%), although there are cases where termination occurs prematurely within the 5' untranslated region (UTR) of a gene (Figure 17).

[000211] TTS sites identified here tend to form stable secondary structures (Figure 3B). The termination efficiency, derived from the level of readthrough transcripts across the termination site, varies widely (Figure 18A). We assigned 709 TTS as Rho-dependent terminators based on their sensitivity to the Rho-specific inhibitor bicyclomycin (BCM) (24) (Figure 3C-3F). Among the other TTS, which are less sensitive to BCM treatment, many display sequence characteristics of an intrinsic terminator (a GC-rich hairpin followed by a 7-8 nt U-rich tract) (21) (Figure 3G-3I). As the number of uridines decreases, the termination efficiency drops—consistent with previous results (25)—and can be further reduced by Rho inhibition (Figure 18B). This result suggests that the intrinsic and Rho-dependent termination mechanisms are

not mutually exclusive and can act on the same site. Alternatively, such apparent overlap could result from RNase trimming following Rho action downstream of the hairpin (23, 26), despite that the aforementioned exonuclease knockout did not substantially change the 3'-end pattern of these sites.

[000212] Taking advantage of the ability of SEnd-seq to simultaneously determine the 5' and 3' ends of the same transcript, we asked whether the TSS selection—especially for those genes that employ multiple start sites—influences the termination efficiency at the corresponding TTS. We found 71 TTS whose termination efficiency alters by at least 40% depending on the choice of TSS (Figure 3J, 3K), implying crosstalk between the two termini as previously proposed (27, 28).

[000213] Annotation of transcription units and antisense transcripts. The concomitant mapping of TSS and TTS enabled us to define 3,578 unique transcription units (TU) in the *E. coli* transcriptome (Figure 19A, 19B; data not shown). Most TU have their boundaries located within intergenic regions. We did detect 323 TU with TSS in a gene-coding region, yielding a shorter RNA product (Figure 19C). We also found 452 TU with an intragenic TSS that drives transcription of a downstream gene (Figure 19D).

[000214] The ability of SEnd-seq to comprehensively profile full-length RNA of different sizes also allowed us to analyze the genome-wide distribution of antisense transcripts, whose prevalence and importance in bacteria are increasingly being appreciated (29, 30). We found that a substantial fraction of transcripts (~15%) are derived from the complementary strand of protein-coding genes. These antisense transcripts are mostly located toward the 3' end of a coding region or within a 3' UTR, and have a wide range of lengths (Figure 20).

[000215] Prevalent overlapping bidirectional TTS revealed by SEnd-seq. As demonstrated above, SEnd-seq provides an unprecedented inventory of the *E. coli* transcriptome. In the following we focus on one of the most striking findings that emerged from the SEnd-seq dataset. There are 658 pairs of neighboring genes in *E. coli* that are orientated in a head-to-head manner (Supplementary Figure 15). Unexpectedly, we discovered that two opposing TTS frequently overlap with each other between a pair of convergent genes (284 out of 658 pairs) (Figure 4A-4D; data not shown). In addition, we found 115 cases in which TTS of an unopposed gene overlaps with that of an antisense RNA (Figure 4B, 4D). These overlapping regions are largely hidden from the standard RNA-seq dataset due to its lack of coverage around RNA ends (Figure 4A, 4B).

[000216] Overlapping bidirectional TTS are on average ~80% efficient in both directions. The termination efficiency tends to be even higher for the sites that are sandwiched between two

highly expressed genes (Figure 4E). The length of the overlapping region ranges from 18 to 60 nt (Figure 4F). The vast majority of these overlapping sequences are predicted to form RNA stem-loop structures (Figure 4G-4I). However, only a minor fraction (~16%) exhibit features of a canonical bidirectional intrinsic terminator (25, 31), i.e., a short GC-rich hairpin flanked by an A-tract and a U-tract on either side (Figure 4J, 4K). Most overlapping regions feature a nonspecific flanking sequence on at least one side of the hairpin. Moreover, the stems tend to be longer than those of typical intrinsic terminators and often contain mismatches and bulges (Figure 22). These bidirectional terminators do not appear to be primarily Rho-dependent either, as the BCM inhibitor only confers a minor effect on their termination efficiency (Figure 23).

[000217] We found that the patterns of these overlapping regions in the RNase-knockout strains (Δpnp , Δrnb , Δrnr) are largely similar to those in the wildtype strain (Figure 24), suggesting that the boundaries of these regions are genuine termination sites rather than products of RNase trimming. In further support of this notion, the overlapping sequences identified here almost always contain single-stranded regions flanking the stem loop, unlike decay products that are usually processed until the edge of the protective hairpin stem (32).

[000218] Convergent transcription drives bidirectional termination in vitro. Since neither intrinsic termination nor Rho-mediated termination can fully explain the widespread occurrence of overlapping TTS between convergent TU pairs, we postulated that head-on collisions between opposing transcription machineries may cause termination in both directions. To test this hypothesis, we performed in vitro transcription assays with *E. coli* RNA polymerase (RNAP) on synthetic DNA templates harboring a convergent gene pair. We copied the genomic sequence around the *yoaJ-yeaQ* locus into the template (Figure 5A, 5B). This region contains a 34-nt overlapping TTS sequence and displays strong bidirectional termination in vivo. When a T7A2 promoter that controls transcription initiation by *E. coli* RNAP was placed at one end of the template, unidirectional transcription was permitted, which resulted in significant readthrough (Figure 5C, 5D). This result confirms the notion that the overlapping TTS sequence alone cannot cause efficient termination. In comparison, in vitro transcription using a strong intrinsic terminator yielded much lower readthrough (Figure 25).

[000219] Importantly, when a promoter was incorporated in both ends of the template in order to support convergent transcription, the readthrough level was significantly reduced (Figure 5C, 5D). The sizes of the RNA products are consistent with termination occurring at positions demarcating the overlapping region. Similar results were obtained with sequences taken from

other convergent gene pairs (Figure 26). These in vitro results strongly suggest that RNAP conflicts alone—without other cellular factors—can induce bidirectional termination.

[000220] NusA is known to stimulate bacterial transcription termination (33). We examined the influence of NusA on convergent transcription and found that NusA further enhanced the bidirectional termination efficiency (Figure 5C, 5D and Figure 26). Therefore, the effect of NusA and the effect of RNAP conflicts on the termination efficiency can be additive.

[000221] How do transcription complexes originating from stochastic initiation events always meet at the overlapping region? Given that the formation of RNA hairpins often contributes to RNAP pausing (34), we posited that the stem-loop structures formed in the overlapping regions—although they do not lead to termination per se—cause RNAP to pause for an extended period of time such that another polymerase traveling from the opposite direction causes interference at the pausing site. To test this idea, we conducted in vitro transcription assays with DNA templates that lack an overlapping TTS sequence (Figure 5E, 5F). As expected, unidirectional transcription yielded predominantly readthrough transcripts (Figure 5G). Interestingly, when convergent transcription was allowed, readthrough decreased but the RNA products were heterogeneous in length (Figure 5G), indicating promiscuous collision sites. This is in contrast to the uniform RNA products released from templates harboring an overlapping TTS sequence (Figure 5C). Therefore, the overlapping TTS sequence—and hence the pausing signal—is required for synchronizing the converging transcription complexes, causing them to interfere with each other at well-defined positions.

[000222] Convergent transcription contributes to bidirectional termination in vivo. To seek further evidence that converging transcription elongation complexes contribute to their own termination inside the cell, we performed in vivo genome editing to disrupt transcription from one direction in an opposing gene pair. We targeted the *yccU-hspQ* convergent pair, which displays a 40-nt overlapping TTS (Figure 6A, 6B). To disrupt *hspQ* transcription, we created the $\Delta hspQ$ strain by deleting the promoter sequence for *hspQ* and inserting two strong intrinsic terminators around the original TSS of *hspQ*. We then assessed the extent of *yccU* readthrough across the overlapping region with strand-specific qPCR. As predicted, the $\Delta hspQ$ strain showed a significant increase in the abundance of *yccU* readthrough transcripts (Figure 6C). Disrupting the transcription of other genes at distal genomic locations did not confer the same effect on *yccU* readthrough (Δhfq and $\Delta yeaQ$ in Figure 6C). Furthermore, we performed SEnd-seq with the $\Delta hspQ$ strain and examined the transcript profile around the *yccU-hspQ* region (Fig. 6d). First of all, *hspQ* transcription was indeed abolished. Secondly, the *yccU* readthrough level markedly increased in the $\Delta hspQ$ dataset compared to the control dataset (45% vs. 6.5%).

Similar results were obtained from genome editing experiments on other convergent gene pairs (Figure 27).

[000223] Together, these in vitro and in vivo results support a model in which the stem-loop structure formed near the 3' ends of two converging transcription units causes pausing of the elongation complex and, subsequently, transcription termination when an opposite elongation complex collides into it (Figure 6E). This model predicts that RNAP occupancy is enriched at the overlapping bidirectional TTS due to pausing. We thus performed RNAP ChIP-seq experiments using antibodies against the β or β' subunit. Indeed, stronger ChIP signals were observed around the overlapping TTS sites compared to nearby regions (Figure 28).

[000224] DISCUSSION

[000225] Despite the reinvigorated interest of the scientific community in RNA biology and the myriad RNA-seq technologies, methods capable of defining the boundaries of all transcripts in a transcriptome still remain scarce. TIF-seq, which was developed to analyze eukaryotic transcript isoforms (15), ligates the termini of dsDNA—as opposed to ssDNA in SEnd-seq—and displays a strong bias toward short transcripts. Recently, a method based on PacBio long-read sequencing was reported (20). But this method involves size-selection steps that remove any RNA shorter than 1,000 nt, and therefore is blind to all small RNA and a significant fraction of mRNA. In contrast, SEnd-seq comprehensively profiles RNA of different sizes in a single assay with reduced length bias. It is worth noting that the conversion from RNA to full-length cDNA in SEnd-seq is critically dependent on the performance of reverse transcription. A highly processive reverse transcriptase was used in this study (see Zhao and Pyle; 35). Continued enzyme engineering could further enhance the transcriptome coverage of SEnd-seq.

[000226] SEnd-seq enabled us to determine the correlated occurrence of TSS and TTS and to discern the crosstalk between promoters and terminators that control the same transcript. Future experiments are needed to elucidate the origin of such crosstalk. Our method uses the sequences of 5' and 3' termini to infer the full-length composition of each distinct transcript. Thus it is most ideally suited for studying organisms with limited splicing. SEnd-seq could also be employed for meta-transcriptomics analysis with RNA pooled from multi-species communities.

[000227] The sharp transcript boundaries defined by SEnd-seq led us to identify a widespread but previously underappreciated mechanism of transcription termination driven by head-on interference between transcription complexes. The unique ability of SEnd-seq to determine the 5'-end origin of terminated RNA and the full sequence of the overlapping region helped to

uncover this mechanism. Transcriptional interference resulting from convergent promoters has been well documented in bacteria (36-38). However, studies of transcriptional interference have thus far mainly focused on its negative impact on gene activity due to promoter occlusion or random RNAP collisions during elongation (39). The present work shows that such interference can be exploited to precisely terminate transcription, thereby limiting undesired readthrough and fine-tuning the transcriptional output. Moreover, although overlapping bidirectional terminators have been reported for a few individual genes (40, 41), the extent to which they occur genome-wide was unexplored. Here we show that this phenomenon is pervasive, which raises the intriguing scenario that head-to-head gene pairs are functionally related, akin to co-directional genes within the same polycistronic operon. In the cases where an opposing gene is absent, antisense transcription can also suppress the readthrough of sense transcription, which adds to the functional repertoire of non-coding RNA.

[000228] In this work we used the strong T7A2 promoter for the in vitro transcription experiments, where we observed efficient bidirectional termination. In vivo, the likelihood of RNAP head-on encounter is influenced by additional factors, notably the promoter strength (42). For highly expressed convergent gene pairs, the frequent physical interference between RNAP is likely a major contributor to the bidirectional termination, although we do not exclude alternative, but not mutually exclusive, mechanisms that may play a role in shaping the transcript 3' boundaries, such as antisense-RNA-mediated attenuation (43). Moreover, given the known effect of ribosome movement on RNAP pause release (44, 45), the uncoupling between transcription and translation downstream of the stop codon may enhance RNAP pausing and termination at intergenic bidirectional TTS. With regard to RNAP collisions, further studies are required to elucidate whether termination is induced by direct contacts between the converging motors or by the accumulation of torsional stress in DNA when they approach (46, 47). Finally, considering that convergent genes and polymerase conflicts are also found in eukaryotes (48-50), it will be interesting to investigate whether the transcription termination mechanism documented here is conserved across kingdoms of life.

TABLE 2

Oligonucleotides for SEnd-seq library preparation

Name	Sequence
5' adaptor	/5SpC3/rArArA rCrUrC rUrCrC rArCrG rUrNrN rNrNrN rN (SEQ ID NO:1)

3' adaptor -1 (barcode 1)	/5Phos/rNrNrNrNrArArCrCrUrGrCrUrArUrCrArArC rUrG/3ddC/ (SEQ ID NO:2)
3' adaptor -2 (barcode 2)	/5Phos/rNrNrNrNrGrCrUrUrCrCrUrGrCrUrArUrCrA rArCrUrG/3ddC/ (SEQ ID NO:3)
3' adaptor -3 (barcode 3)	/5Phos/rNrNrNrNrNrNrArUrA rGrCrA rArCrC rUrGrC rUrArU rCrArA rCrUrG /3ddC/ (SEQ ID NO:4)
Biotinylated RT primer	GGG CAG T/iBiodT/G ATA GCA GG (SEQ ID NO:5)

/5SpC3/: 5' C3 Spacer modification

/5Phos/: 5' Phosphorylation modification

/3ddC/: 3' Dideoxycytidine (ddC) modification

/iBiodT/: Internal biotin dT

'r': ribonucleotide

Sequences are represented in the 5'-3' orientation. All of these oligonucleotides were HPLC purified after synthesis.

TABLE 3

Oligonucleotides for in vitro RNA synthesis by *E. coli* RNA polymerase

Name	Sequence
T7A2 half-1 1878781 forward	GCT AGG TAA CAG ACG CAA TCA ATG GGT GCG GTA TTA TTT TTC CGG TTG TAA TC (SEQ ID NO:6)
T7A2 half-2 1879297 reverse	GCT AGG TAA CAT GAT GAT TGT GCG TGC TAC AGC GAC TAA CCC GCT AAT GCA AAA CAG (SEQ ID NO:7)
T7A2 full-1 forward	TGG TCG GCG TTT TTA AAA CAG GTA TTG ACA ACA TGA AGT AAC ATG CAG TAA GAT ACA AAT CGC TAG GTA ACA GAC GCA ATC AAT GGG TGC (SEQ ID NO:8)
T7A2 full-2 reverse	CGG CGC CAT TTT TAA AAC AGG TAT TGA CAA CAT GAA GTA ACA TGC AGT AAG ATA CAA ATC GCT AGG TAA CAT GAT GAT TGT GCG TGC TAC (SEQ ID NO:9)
T7A2 half-1 922200 forward	GCT AGG TAA CAG ACG CAA TCA ATG GGT GCC TCT GCT CGA CGG TCA CCG GG (SEQ ID NO:10)
T7A2 half-2 922600 reverse	GCT AGG TAA CAT GAT GAT TGT GCG TGC TAC ATG TCG AAG CAT GGA AAA GGG (SEQ ID NO:11)
T7A2 half-1 1879000 forward	GCT AGG TAA CAG ACG CAA TCA ATG GGT GCT AAG CGT TAA CTT TTA ATC TTC (SEQ ID NO:12)
T7A2 half-2 1879300 reverse	GCT AGG TAA CAT GAT GAT TGT GCG TGC TAC ATC AGC GAC TAA CCC GCT AAT G (SEQ ID NO:13)
T7A2 half-1 1028290 forward	GC TAG GTAACAGACGCAAT CAATGGGTGCCCG GGT TAA ATG TGG TGA TGG (SEQ ID NO:14)
T7A2 half-2 1028540 reverse	GC TAG GTAACATGATGAT TGTGC GTGCT ACG GAC GAT AAC GGC CTA CCG G (SEQ ID NO:15)
T7A2 half-1 1863700 reverse	GC TAG GTAACAGACGCAAT CAATGGGTGCGTG GTA CGA CAA CGA AAC CGG (SEQ ID NO:16)
T7A2 full-1 2699900 forward	GCT AGG TAA CAG ACG CAA TCA ATG GGT GCA CCA CGC GGA TAA AAT TTA ATA TTC (SEQ ID NO:17)
T7A2 full-2 2700210 reverse	GCT AGG TAA CAT GAT GAT TGT GCG TGC TAC CGA ATG ACT GAT TGC CGA TAC C (SEQ ID NO:18)

T7A2 full-1 1524276 forward	GCT AGG TAA CAG ACG CAA TCA ATG GGT GCA TGG CAA AAC GCT GTA TGT CAG (SEQ ID NO:19)
T7A2 full-2 1524600 reverse	GCT AGG TAA CAT GAT GAT TGT GCG TGC TAC CTG GTT ATA GGC TGG TTT GGT G (SEQ ID NO:20)
1863950 reverse	GGA TGA TCG ACC ACA ATG AGG TCC (SEQ ID NO:21)

The numbers indicate genomic positions.

TABLE 4

Oligonucleotides for in vitro RNA synthesis by phage T7 RNA polymerase

Name	Sequence
Spike-in RNA (C gene 500nt) forward	TAC GTA ATA CGA CTC ACT ATA GGG GTG ACA GCA GAG CTG CGT AAT CTC (SEQ ID NO:22)
Spike-in RNA (C gene 500nt) reverse	GCC GGG AGG CGG CAC TGG CAA G (SEQ ID NO:23)
Spike-in RNA (B gene 1000nt) forward	TAC GTA ATA CGA CTC ACT ATA GGG ATG AAA ACG CCC ACC ATT CCC AC (SEQ ID NO:24)
Spike-in RNA (B gene 1000nt) reverse	CCG TAT CCT GAG CCG TCT GCA G (SEQ ID NO:25)
Spike-in RNA (A gene 1500nt) forward	TAC GTA ATA CGA CTC ACT ATA GGG GTG AAT ATA TCG AAC AGT CAG G (SEQ ID NO:26)
Spike-in RNA (A gene 1500nt) reverse	GGC CAC CGG CTT TCC GTA GAC GG (SEQ ID NO:27)
Spike-in RNA (H gene 2000nt) forward	TAC GTA ATA CGA CTC ACT ATA GGG ATG GCT GAA CCG GTA GGC GAT CTG (SEQ ID NO:28)
Spike-in RNA (H gene 2000nt) reverse	TCA TCA TGG ACA GCA CGG AAC GGG (SEQ ID NO:29)

TABLE 5

Oligonucleotides for genome editing and qPCR

Name	Sequence
TR kana stop site forward	GCA TAC ATA AAA AAC CTG CCA GCG ATG GCA GGT TTT TTT TAT TCC GGG GAT CCG TCG ACC (SEQ ID NO:30)
TR kana stop site reverse	AAC CAC GTT TAA AAA AGG TGC TCA ATG AGC ACC TTT TTT CTG TAG GCT GGA GCT GCT TCG (SEQ ID NO:31)
yoak PD_stop forward	TTT CGT CGC TGC GAG AAC GCA TTC AGC AGC AAA TAC GTG GGC ATA CAT AAA AAA CCT GCC (SEQ ID NO:32)
yoak PD_stop reverse	CTT TAT GCT CCC GGG GCA GCA TAG CCG CCA ATA AAA AAC CAA CCA CGT TTA AAA AAG GTG (SEQ ID NO:33)
ynaj PD_stop forward	ATC GAA TAC ATG AAA GCC AGC ATC CGG GCC AGG GTG GAG CGC ATA CAT AAA AAA CCT GCC (SEQ ID NO:34)
ynaj PD_stop reverse	AAA TTT CTT CCC TTT CGC TGA CTT CAG TTT CGC CAT AAT CAA CCA CGT TTA AAA AAG GTG (SEQ ID NO:35)

uspe PD_stop forward	TCT TCT GGC AGA CCT TTT TCT ACG TGC GTC ATG TTT TCA TGC ATA CAT AAA AAA CCT GCC (SEQ ID NO:36)
uspe PD_stop reverse	AAA TTT TTC TGA TTT ATT GAT CTG GCA GAA GGT TCA TCA CAA CCA CGT TTA AAA AAG GTG (SEQ ID NO:37)
hspq PD_stop forward	CCA GCT CAT CAG GCG ACG GTT CAG AGA GCG AAT AAA CCG GGC ATA CAT AAA AAA CCT GCC (SEQ ID NO:38)
hspq PD_stop reverse	TAT TGC ACG GCA GAA CGT TGA GCT GAA CAA ACT GGA TCT GAA CCA CGT TTA AAA AAG GTG (SEQ ID NO:39)
yoak PD_S pcr forward	ACC CAT CAT AAT AAT CGT TGT T (SEQ ID NO:40)
yoak PD_S pcr reverse	TTC ATT ACA GCG GTC GTA (SEQ ID NO:41)
yoak PD_S RT primer	TAA TCT TCA TTA CAG CG (SEQ ID NO:42)
ynaj PD_S pcr forward	GTA AAG ACC AGG CTA TAA ACA A (SEQ ID NO:43)
ynaj PD_S pcr reverse	TAT TCC GCT GTC TGA GTG (SEQ ID NO:44)
ynaj RT primer	AAT TTC TCT TTG GTT TG (SEQ ID NO:45)
uspe PD_S qpcr forward	TCT TCT TCG TCA TCC AGT T (SEQ ID NO:46)
uspe PD_S qpcr reverse	CGG TAT TTC AGC AGC ATT (SEQ ID NO:47)
uspe PD_S RT primer	GGT GAT TCC TGA TTT GG (SEQ ID NO:48)
hspq PD_S qpcr forward	ATG GCT GTT CAG GAT GTT (SEQ ID NO:49)
hspq PD_S qpcr reverse	GTA GTG ATG GAG GAC GAT AA (SEQ ID NO:50)
hspq PD_S RT primer	CGG TGA ATG ACG AGC TTC (SEQ ID NO:51)
pnp del forward	CCC CCC GCC GCA GCG GAG GGC AAA TGG CAA CCT TAC TCG CAT TCC GGG GAT CCG TCG ACC (SEQ ID NO:52)
pnp del reverse	TTC CCA CTG ACC GTT AAC TAT CAG GAG CGT ACC TAC GCT GTG TAG GCT GGA GCT GCT TCG (SEQ ID NO:53)
rnr del forward	ATC CTT TCC AGG AAC GCG AAG CTG AAA AAT ACG CGA ATC CAT TCC GGG GAT CCG TCG ACC (SEQ ID NO:54)
rnr del reverse	CCT CTT CTT TTA AAG AGG GTA TTG ATC ACT CTG CCA CTT TTG TAG GCT GGA GCT GCT TCG (SEQ ID NO:55)
rnb del forward	CGC GAC CGG GCG CGC AAT AAT GCT GCG GGT TTC CAT GCG GAT TCC GGG GAT CCG TCG ACC (SEQ ID NO:56)

[000229] REFERENCES

1. Morris, K.V. & Mattick, J.S. The rise of regulatory RNA. Nat Rev Genet 15, 423-37 (2014).
2. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10, 57-63 (2009).
3. Sharma, C.M. et al. The primary transcriptome of the major human pathogen *Helicobacter pylori*. Nature 464, 250-5 (2010).

4. Wurtzel, O. et al. A single-base resolution map of an archaeal transcriptome. *Genome Res* 20, 133-41 (2010).
5. Dar, D. et al. Term-seq reveals abundant ribo-regulation of antibiotics resistance in bacteria. *Science* 352, aad9822 (2016).
6. Babski, J. et al. Genome-wide identification of transcriptional start sites in the haloarchaeon *Haloferax volcanii* based on differential RNA-Seq (dRNA-Seq). *BMC Genomics* 17, 629 (2016).
7. Lalanne, J.B. et al. Evolutionary Convergence of Pathway-Specific Enzyme Expression Stoichiometry. *Cell* 173, 749-761 e38 (2018).
8. Ettwiller, L., Buswell, J., Yigit, E. & Schildkraut, I. A novel enrichment strategy reveals unprecedented number of novel transcription start sites at single base resolution in a model prokaryote and the gut microbiome. *BMC Genomics* 17, 199 (2016).
9. Matteau, D. & Rodrigue, S. Precise Identification of Genome-Wide Transcription Start Sites in Bacteria by 5'-Rapid Amplification of cDNA Ends (5'-RACE). *Methods Mol Biol* 1334, 143-59 (2015).
10. Goodwin, S., McPherson, J.D. & McCombie, W.R. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* 17, 333-51 (2016).
11. Hor, J., Gorski, S.A. & Vogel, J. Bacterial RNA Biology on a Genome Scale. *Mol Cell* 70, 785-799 (2018).
12. Guell, M., Yus, E., Lluch-Senar, M. & Serrano, L. Bacterial transcriptomics: what is beyond the RNA horiz-ome? *Nat Rev Microbiol* 9, 658-69 (2011).
13. Gama-Castro, S. et al. RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. *Nucleic Acids Res* 44, D133-43 (2016).
14. Ruan, X. & Ruan, Y. Genome wide full-length transcript analysis using 5' and 3' paired-end-tag next generation sequencing (RNA-PET). *Methods Mol Biol* 809, 535-62 (2012).
15. Pelechano, V., Wei, W. & Steinmetz, L.M. Extensive transcriptional heterogeneity revealed by isoform profiling. *Nature* 497, 127-31 (2013).
16. Lama, L. & Ryan, K. Adenylation of small RNA sequencing adapters using the TS2126 RNA ligase I. *RNA* 22, 155-61 (2016).
17. Lin-Chao, S., Wei, C.L. & Lin, Y.T. RNase E is required for the maturation of *ssrA* RNA and normal *ssrA* RNA peptide-tagging activity. *Proc Natl Acad Sci U S A* 96, 12406-11 (1999).
18. Ruff, E.F., Record, M.T., Jr. & Artsimovitch, I. Initial events in bacterial transcription initiation. *Biomolecules* 5, 1035-62 (2015).
19. Conway, T. et al. Unprecedented high-resolution view of bacterial operon architecture revealed by RNA sequencing. *MBio* 5 (2014).

20. Yan, B., Boitano, M., Clark, T.A. & Ettwiller, L. SMRT-Cappable-seq reveals complex operon variants in bacteria. *Nat Commun* 9, 3676 (2018).
21. Ray-Soni, A., Bellecourt, M.J. & Landick, R. Mechanisms of Bacterial Transcription Termination: All Good Things Must End. *Annu Rev Biochem* 85, 319-47 (2016).
22. Hui, M.P., Foley, P.L. & Belasco, J.G. Messenger RNA degradation in bacterial cells. *Annu Rev Genet* 48, 537-59 (2014).
23. Dar, D. & Sorek, R. High-resolution RNA 3'-ends mapping of bacterial Rho-dependent transcripts. *Nucleic Acids Res* 46, 6797-6805 (2018).
24. Zwiefka, A., Kohn, H. & Widger, W.R. Transcription termination factor rho: the site of bicyclomycin inhibition in *Escherichia coli*. *Biochemistry* 32, 3564-70 (1993).
25. Chen, Y.J. et al. Characterization of 582 natural and synthetic terminators and quantification of their design constraints. *Nat Methods* 10, 659-64 (2013).
26. Wang, X. et al. Processing generates 3' ends of RNA masking transcription termination events in prokaryotes. *Proc Natl Acad Sci U S A* (2019).
27. Goliger, J.A., Yang, X.J., Guo, H.C. & Roberts, J.W. Early transcribed sequences affect termination efficiency of *Escherichia coli* RNA polymerase. *J Mol Biol* 205, 331-41 (1989).
28. Telesnitsky, A.P. & Chamberlin, M.J. Sequences linked to prokaryotic promoters can affect the efficiency of downstream termination sites. *J Mol Biol* 205, 315-30 (1989).
29. Thomason, M.K. et al. Global transcriptional start site mapping using differential RNA sequencing reveals novel antisense RNAs in *Escherichia coli*. *J Bacteriol* 197, 18-28 (2015).
30. Dornenburg, J.E., Devita, A.M., Palumbo, M.J. & Wade, J.T. Widespread antisense transcription in *Escherichia coli*. *MBio* 1 (2010).
31. Peters, J.M., Vangeloff, A.D. & Landick, R. Bacterial transcription terminators: the RNA 3'-end chronicles. *J Mol Biol* 412, 793-813 (2011).
32. Dar, D. & Sorek, R. Extensive reshaping of bacterial operons by programmed mRNA decay. *PLoS Genet* 14, e1007354 (2018).
33. Mondal, S., Yakhnin, A.V., Sebastian, A., Albert, I. & Babitzke, P. NusA-dependent transcription termination prevents misregulation of global gene expression. *Nat Microbiol* 1, 15007 (2016).
34. Zhang, J. & Landick, R. A Two-Way Street: Regulatory Interplay between RNA Polymerase and Nascent RNA Structure. *Trends Biochem Sci* 41, 293-310 (2016).
35. Zhao, C., Liu, F. & Pyle, A.M. An ultraprocessive, accurate reverse transcriptase encoded by a metazoan group II intron. *RNA* 24, 183-195 (2018).
36. Callen, B.P., Shearwin, K.E. & Egan, J.B. Transcriptional interference between convergent promoters caused by elongation over the promoter. *Mol Cell* 14, 647-56 (2004).

37. Horowitz, H. & Platt, T. Regulation of transcription from tandem and convergent promoters. *Nucleic Acids Res* 10, 5447-65 (1982).
38. Elledge, S.J. & Davis, R.W. Position and density effects on repression by stationary and mobile DNA-binding proteins. *Genes Dev* 3, 185-97 (1989).
39. Shearwin, K.E., Callen, B.P. & Egan, J.B. Transcriptional interference--a crash course. *Trends Genet* 21, 339-45 (2005).
40. Sameshima, J.H., Wek, R.C. & Hatfield, G.W. Overlapping transcription and termination of the convergent *ilvA* and *ilvY* genes of *Escherichia coli*. *J Biol Chem* 264, 1224-31 (1989).
41. Postle, K. & Good, R.F. A bidirectional rho-independent transcription terminator between the *E. coli* *tonB* gene and an opposing gene. *Cell* 41, 577-85 (1985).
42. Sneppen, K. et al. A mathematical model for transcriptional interference by RNA polymerase traffic in *Escherichia coli*. *J Mol Biol* 346, 399-409 (2005).
43. Brantl, S. & Wagner, E.G. An antisense RNA-mediated transcriptional attenuation mechanism functions in *Escherichia coli*. *J Bacteriol* 184, 2740-7 (2002).
44. Landick, R., Carey, J. & Yanofsky, C. Translation activates the paused transcription complex and restores transcription of the *trp* operon leader region. *Proc Natl Acad Sci U S A* 82, 4663-7 (1985).
45. Proshkin, S., Rahmouni, A.R., Mironov, A. & Nudler, E. Cooperation between translating ribosomes and RNA polymerase in transcription elongation. *Science* 328, 504-8 (2010).
46. Ma, J., Bai, L. & Wang, M.D. Transcription under torsion. *Science* 340, 1580-3 (2013).
47. Crampton, N., Bonass, W.A., Kirkham, J., Rivetti, C. & Thomson, N.H. Collision events between RNA polymerases in convergent transcription studied by atomic force microscopy. *Nucleic Acids Res* 34, 5416-25 (2006).
48. Hobson, D.J., Wei, W., Steinmetz, L.M. & Svejstrup, J.Q. RNA polymerase II collision interrupts convergent transcription. *Mol Cell* 48, 365-74 (2012).
49. Prescott, E.M. & Proudfoot, N.J. Transcriptional collision between convergent genes in budding yeast. *Proc Natl Acad Sci U S A* 99, 8796-801 (2002).
50. Eszterhas, S.K., Bouhassira, E.E., Martin, D.I. & Fiering, S. Transcriptional interference by independently regulated genes occurs in any relative arrangement of the genes and is influenced by chromosomal integration position. *Mol Cell Biol* 22, 469-79 (2002).
51. Creecy, J.P. & Conway, T. Quantitative bacterial transcriptomics with RNA-seq. *Curr Opin Microbiol* 23, 133-40 (2015).
52. Jensen, S.I., Lennen, R.M., Herrgard, M.J. & Nielsen, A.T. Seven gene deletions in seven days: Fast generation of *Escherichia coli* strains tolerant to acetate and osmotic stress. *Sci Rep* 5, 17874 (2015).

53. Peters, J.M. et al. Rho directs widespread termination of intragenic and stable RNA transcription. *Proc Natl Acad Sci U S A* 106, 15406-11 (2009).
54. McClure, R. et al. Computational analysis of bacterial RNA-Seq data. *Nucleic Acids Res* 41, e140 (2013).
55. Bailey, T.L. et al. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* 37, W202-8 (2009).
56. Celesnik, H., Deana, A. & Belasco, J.G. Initiation of RNA decay in *Escherichia coli* by 5' pyrophosphate removal. *Mol Cell* 27, 79-90 (2007).
57. Lorenz, R. et al. ViennaRNA Package 2.0. *Algorithms Mol Biol* 6, 26 (2011).
58. Crooks, G.E., Hon, G., Chandonia, J.M. & Brenner, S.E. WebLogo: a sequence logo generator. *Genome Res* 14, 1188-90 (2004).
59. Kim, D. et al. Comparative analysis of regulatory elements between *Escherichia coli* and *Klebsiella pneumoniae* by genome-wide transcription start site profiling. *PLoS Genet* 8, e1002867 (2012).
60. Deutscher, M.P. Maturation and degradation of ribosomal RNA in bacteria. *Prog Mol Biol Transl Sci* 85, 369-91 (2009).
61. Conway, T. et al. Unprecedented high-resolution view of bacterial operon architecture revealed by RNA sequencing. *MBio* 5 (2014).
62. Sedlyarova, N. et al. sRNA-Mediated Control of Transcription Termination in *E. coli*. *Cell* 167, 111-121 e13 (2016).

EXAMPLE 2

SEnd-seq APPLIED TO *BORRELIA BURGDORFERI*

[000230] Lyme disease is a type of animal-borne disease caused by spirochetes of *Borrelia burgdorferi* (1), which affects multiple organs such as skin, nervous system, and heart in humans (2). Lyme disease is prevalently reported in Europe and North America, with more than 30,000 cases reported to the US Centers for Disease Control and Prevention (CDC) each year (3). Lyme symptoms vary from person to person. Some patients are asymptomatic, whereas some develop brain or heart damage at a late stage (4). Treatment of Lyme disease consists of several antibiotics. However, studies show an approximately 20 percent treatment failure rate (1). Moreover, there is still no reliable and fast way to test people for the bacteria that cause Lyme disease, as clinical symptoms do not always appear (5).

[000231] *B. burgdorferi* is often isolated from mosquitoes, fleas, flies and ticks, with ticks playing the main role of transmitting the infection to humans (1). *B. burgdorferi* adapts to

disparate environments when transmitted between ticks and mammals. In naturally and experimentally infected murine hosts, *B. burgdorferi* has been found in a variety of tissues, including heart, bladder, joints, and ears (3). It is interesting and important to understand how *B. burgdorferi* survives in different environments.

[000232] The genome of *B. burgdorferi* contains a linear chromosome of 910,725 base pairs and at least 17 linear and circular plasmids with a combined size of more than 533,000 base pairs. Its genome is one of the most complex of any bacterium (6). The chromosome contains 853 genes. Of 430 genes on the commonly existing 11 plasmids, most have no known biological function (7).

[000233] In order to survive diverse environments, *B. burgdorferi* alters its pattern of gene expression in response to environmental signals, including temperature and nutrients (8). Central to this response is the RpoN-RpoS alternative sigma factor cascade that transcriptionally controls the expression of numerous genes required for the transmission from ticks and the establishment of infection in mammals (9). Recently, a large number of small RNAs have been identified in *B. burgdorferi*, many of which are believed to be involved in gene regulation (10). However, the functions of these RNAs remain largely unknown. A better characterization of the differential *B. burgdorferi* transcriptomes under different conditions would improve our understanding of the pathogen's adaptive mechanisms (11), such as those responsible for its dissemination and colonization in host cells. Such studies also promise to identify novel targets for the development of new treatments and diagnostics.

[000234] Recently we devised and reported on a new RNA sequencing method, termed *simultaneous 5' and 3' end sequencing* (SEnd-seq) (Example 1 and ref 12), which concurrently captures both ends of cellular transcripts. This method is able to map the correlated occurrence of transcription start sites (TSS) and termination sites (TTS) across a whole transcriptome with single-nucleotide resolution. By annotating gene-coding and non-coding RNA transcripts in *B. burgdorferi*, SEnd-seq is expected to provide unprecedented insights into the principles of gene regulation by this complex bacterial pathogen.

[000235] METHODS

[000236] Bacterial strains and growth conditions. *B. burgdorferi* strain B31 was cultivated at 34 °C under 5% CO₂ in a complete Barbour-Stoenner-Kelly II medium (BSK-II) with 6% (vol/vol) rabbit serum (13). Cell densities and growth phases were monitored by visualization under dark-field microscopy and by counting using a Petroff-Hausser counting chamber. To

mimic the infectious process in human, cells passaged from normal culture were grown at 23 °C, and then cultured at 34 °C again after one passage. Samples in mid-log phase were collected from different temperature stages for SEnd-seq.

[000237] RNA isolation. Bacterial cells were grown to mid-log phase ($\sim 10^7$ /mL) and harvested by spinning the culture medium at 8,200 g for 30 min at 4 °C. After PBS washing, the cell pellet was resuspended with 1 mL 60 °C warmed Trizol Reagent (Invitrogen, 15596) by pipetting up and down. After additional incubation at room temperature for 5 min, the cells were frozen at -80 °C. For RNA extraction, 0.2 mL of 1-bromo-3-chloropropane (Sigma B9673) was added to the thawed sample. The sample was then gently inverted several times until reaching homogeneity, and incubated for 15 min at room temperature before centrifugation at 12,000 g for 10 min. The upper phase (~ 600 μ L) was gently collected and mixed at a 1:1 ratio with 100% isopropanol. The mixture was incubated for 1 hr at -20 °C and then centrifuged at 14,000 rpm for 10 min at 4 °C. The pellet was washed twice with 1 mL of 75% ethanol, air dried for 5 min, and dissolved in nuclease-free water. RNA integrity was assessed with 1% agarose gel and Agilent 2100 Bioanalyzer System.

[000238] Library preparation for total RNA SEnd-seq. The protocol for SEnd-seq library preparation was slightly modified from our previously provided method (see Example 1 and ref 12). In order to label the 5' end of processed RNA, 5-10 μ g of total RNA in 12 μ L volume was used to incubate with a 5' adaptor ligation mix (1 μ L of 100 μ M 5' adaptor, 0.5 μ L of 50 mM ATP, 2 μ L of dimethyl sulfoxide, 5 μ L of 50% PEG8000, 1 μ L of RNase Inhibitor and 1 μ L of High Concentration T4 RNA Ligase 1) at 23 °C for 5 hr. Then the sample was diluted with water and cleaned twice with 1.5 $\times$ vol of Agencourt RNAClean XP beads (Beckman Coulter, A63987). The eluted RNA was ligated to a 3' adaptor in the same way as the 5'-adaptor ligation. The 5' and 3' adaptors of Example 1 were utilized for this study. After incubation at 23 °C for 5 hr, the reaction was diluted to 40 μ L with water and purified twice with 1.5 $\times$ vol of Agencourt RNAClean XP beads to remove excess RNA adaptors. The sample was subsequently eluted with 0.1 $\times$ TE buffer (10 mM Tris-HCl pH 7.5 and 1 mM EDTA) and subjected to rRNA removal with RiboMinus Transcriptome Isolation Kit (ThermoFisher, K155004). After RNA recovery by ethanol precipitation, the RNA was reverse transcribed with maturase from *Eubacterium rectale* (recombinantly purified from *E. coli*, a gift from Anna Marie Pyle, Yale University) (14) and biotinylated primers. After thorough cleaning, the cDNA was circularized by the Ts2126 RNA Ligase I (15). Double-stranded DNA was generated with DNA polymerase I (New England BioLabs, M0209S) and subsequently fragmented by dsDNA Fragmentase (New England BioLabs, M0348S) at 37 °C for 15 min. The reaction was stopped by adding 5

μl of 0.5 M EDTA and heated at 65 °C for 15 min in the presence of 50 mM DTT. The DNA was diluted to 40 μl with TE buffer and purified with 1× vol of AMPure beads. The eluted DNA was used for sequencing library preparation with the NEBNext Ultra II DNA Library Prep Kit (New England BioLabs, E7645). The DNA library was amplified for 12 (for total RNA SEnd-seq) to 15 cycles (for primary RNA SEnd-seq) following the manufacturer's protocol.

[000239] Library preparation for primary transcripts. 5 μg of total RNA was used for primary transcript enrichment with our previous described method (Example 1 and ref 12). The eluted RNA was used for SEnd-seq library preparation with 15 PCR cycles of amplification.

[000240] Data analysis. SEnd-seq data were collected by the Illumina NextSeq 500 platform in a paired-end mode (150 nt ×2). The raw data were processed as described in Example 1 and as reported (12). The full-length sequences were inferred by mapping to the reference *B. burgdorferi* B31 genome GCF_000008685.2 using Bowtie 2. Reads with an insert length greater than 10,000 nt were discarded. TSS, TTS and gene coverage data were extracted as described in Example 1 and as reported (12).

[000241] RESULTS

[000242] We cultured *B. burgdorferi* at different temperatures to mimic bacteria grown in the hosts of flea or human with different body temperatures, then studied the transcriptomic response to the temperature shift. 1,890 transcription start sites (TSS) were enriched from different growth conditions, about half of which are located inside gene bodies. On the other hand, the TTS number is low and most of these TTS sites are also located in gene coding regions. In addition, we observed pervasive antisense RNA transcripts robustly expressed across different conditions (Figure 30). The length and location of the antisense transcripts indicate that they may play important roles in transcriptional regulation. *B. burgdorferi* contains a varied number of plasmids besides the main chromosome. Genes coded in the plasmids were reported to help the bacterium survive in different hosts or environments. Indeed, we found that many genes in these plasmids were differentially expressed when the cells were cultured at different temperatures (Figure 31). Some small RNAs were exclusively expressed under a specific condition.

[000243] REFERENCES

- 1 Biesiada, G., Czepiel, J., Lesniak, M. R., Garlicki, A. & Mach, T. Lyme disease: review. *Arch Med Sci* **8**, 978-982, doi:10.5114/aoms.2012.30948 (2012).

- 2 Sanchez, E., Vannier, E., Wormser, G. P. & Hu, L. T. Diagnosis, Treatment, and Prevention of Lyme Disease, Human Granulocytic Anaplasmosis, and Babesiosis: A Review. *JAMA* **315**, 1767-1777, doi:10.1001/jama.2016.2884 (2016).
- 3 Perronne, C. Lyme and associated tick-borne diseases: global challenges in the context of a public health threat. *Front Cell Infect Microbiol* **4**, 74, doi:10.3389/fcimb.2014.00074 (2014).
- 4 Lacout, A., El Hajjam, M., Marcy, P. Y. & Perronne, C. The Persistent Lyme Disease: "True Chronic Lyme Disease" rather than "Post-treatment Lyme Disease Syndrome". *J Glob Infect Dis* **10**, 170-171, doi:10.4103/jgid.jgid_152_17 (2018).
- 5 Marques, A. R. Laboratory diagnosis of Lyme disease: advances and challenges. *Infect Dis Clin North Am* **29**, 295-307, doi:10.1016/j.idc.2015.02.005 (2015).
- 6 Fraser, C. M. *et al.* Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. *Nature* **390**, 580-586, doi:10.1038/37551 (1997).
- 7 Brisson, D., Drecktrah, D., Eggers, C. H. & Samuels, D. S. Genetics of *Borrelia burgdorferi*. *Annu Rev Genet* **46**, 515-536, doi:10.1146/annurev-genet-011112-112140 (2012).
- 8 Samuels, D. S. Gene regulation in *Borrelia burgdorferi*. *Annu Rev Microbiol* **65**, 479-499, doi:10.1146/annurev.micro.112408.134040 (2011).
- 9 Ouyang, Z., Blevins, J. S. & Norgard, M. V. Transcriptional interplay among the regulators Rrp2, RpoN and RpoS in *Borrelia burgdorferi*. *Microbiology* **154**, 2641-2658, doi:10.1099/mic.0.2008/019992-0 (2008).
- 10 Lybecker, M. C. & Samuels, D. S. Small RNAs of *Borrelia burgdorferi*: Characterizing Functional Regulators in a Sea of sRNAs. *Yale J Biol Med* **90**, 317-323 (2017).
- 11 Arnold, W. K. *et al.* RNA-Seq of *Borrelia burgdorferi* in Multiple Phases of Growth Reveals Insights into the Dynamics of Gene Expression, Transcriptome Architecture, and Noncoding RNAs. *PLoS One* **11**, e0164165, doi:10.1371/journal.pone.0164165 (2016).
- 12 Ju, X., Li, D. & Liu, S. Full-length RNA profiling reveals pervasive bidirectional transcription terminators in bacteria. *Nat Microbiol* **4**, 1907-1918, doi:10.1038/s41564-019-0500-z (2019).
- 13 Zuckert, W. R. Laboratory maintenance of *Borrelia burgdorferi*. *Curr Protoc Microbiol* **Chapter 12**, Unit 12C 11, doi:10.1002/9780471729259.mc12c01s4 (2007).
- 14 Zhao, C., Liu, F. & Pyle, A. M. An ultraprocessive, accurate reverse transcriptase encoded by a metazoan group II intron. *Rna* **24**, 183-195, doi:10.1261/rna.063479.117 (2018).
- 15 Lama, L. & Ryan, K. Adenylation of small RNA sequencing adapters using the TS2126 RNA ligase I. *RNA* **22**, 155-161, doi:10.1261/rna.054999.115 (2016).

EXAMPLE 3

SEnd-seq APPLIED TO MYCOBACTERIA TUBERCULOSIS

[000244] Despite more than one century of active research, widely used vaccine and therapy drugs, tuberculosis (TB) is still the leading infectious disease threatening human lives, killing more people than HIV/AIDS (1). This is attributed to the fact that the causative

pathogen, *Mycobacterium tuberculosis* (Mtb), has an exceptional ability to respond effectively to host defense and drug treatment, allowing it to survive inside the host for long periods. With time Mtb cells can develop resistance to drugs. Nowadays, multidrug-resistant TB (MDR-TB) accounts for 4.1% of all new TB cases and 19% of previously treated cases worldwide (2).

[000245] Mtb spreads from person to person almost exclusively through aerosol droplets that contain Mtb bacteria. They are then quickly phagocytized by the lung macrophages at the beginning of infection (3). The survived Mtb can replicate and recruit more inflammatory and immunity cells in the lung to establish the complex granulomas. Most infected individuals remain in a latent state of infection, in which no clinical symptoms are present. A small percentage of these people eventually progress and develop active disease, which can lead to the release of Mtb and generate infectious droplets that transmit the disease.

[000246] In Mtb's life cycle, it can persist in the presence of reactive oxygen intermediates as well as acidity within the phagosomes of macrophages, low oxygen stress, as well as toxic lipases and proteases released by dead immune cells within the centers of caseating granulomas. Also, it can alter its activity and survive in the granulomas for many years. Mtb achieves this feat by enacting integrated control of gene expression in response to changing environments. Thus, transcriptome-wide analyses are key to understanding Mtb gene regulation, pathogenesis, and persistence. However, features of the Mtb transcriptome deviate significantly from the classical *Escherichia coli* paradigm and remain poorly studied. In order to combat the emerging antibiotic resistance and find new drug targets, proper identification, annotation, and categorization of transcription start sites (TSSs) and termination sites (TTSSs) in Mtb is urgently needed.

[000247] In addition, many characteristics specific to the Mtb transcriptome have been reported, such as pervasive leaderless transcripts (4), antisense transcripts (5), and small RNAs (6). However, their functions and effects on gene regulation are largely unknown. Besides, as transcription, translation, and RNA degradation are coupled in time and space in bacteria, a high-resolution transcriptomic map could provide useful information on how cells regulate their gene expression.

[000248] RNA sequencing (RNA-seq) is a powerful high-throughput method to study the dynamic changes of Mtb gene expression in different environments. However, the published RNA-seq datasets mainly focused on the gene expression levels. Recently we devised and described a new RNA-seq method termed SEnd-seq (Example 1), which can simultaneously

detect both ends of cellular transcripts with single-nucleotide resolution. SEnd-seq has the potential to provide a higher-resolution transcriptomic map that will allow us to better understand the complex life cycle of *Mtb*.

[000249] METHODS

[000250] Bacterial strains and growth conditions. *M. tuberculosis* H37Rv was grown in a Middlebrook 7H9 medium supplemented with 0.5% glycerol, 0.05% tyloxapol, 0.2g/L casamino acids, 0.024g/L pantothenic acid and 10% OADC (BD 212351) (minimal medium). The leucine auxotroph of *M. tuberculosis* (Δ leuD) (7) was grown in minimal medium with the addition of 50 mg/L of L-leucine (Sigma). Solid cultures were grown on 7H10 agar supplemented as described above except tyloxapol. *Mycobacterium smegmatis* MC²155 was grown in the Middlebrook 7H9 medium supplemented with 0.2% glycerol, 0.05% Tween80 and 10% albumin-dextrose-catalase (ADC).

[000251] When necessary, antibiotic 50 mg/L rifampicin (Sigma), 30 mg/L linezolid (Sigma), 40 mg/L clarithromycin (Sigma), 300 mg/L streptomycin, 100 ng/L anhydrotetracycline (ATc) or 20 mg/L kanamycin (Sigma) was added. Bacterial growth was monitored by measuring the optical densities of the broth cultures over time. All liquid cultures were grown at 37 °C in Nalgene sterile square PETG media bottle with constant agitation.

[000252] RNA isolation. Bacterial cells were quenched by adding 1× vol of GTC buffer (600g/L Guanidium thiocyanate, 5 g/L N-laurylsarcosine and 7.1 g/L sodium citrate and 0.7% 2-mercaptoethanol) to the culture medium immediately before harvest and placed at room temperature for 15 min. Cell pellets were collected by centrifugation (4,000 g for 10 min at 4 °C), then thoroughly resuspended in 100 µl of TE buffer (10 mM Tris-HCl and 1 mM EDTA, pH 8.0). After adding 1 mL of TRIzol Reagent (Invitrogen, 15596) and 300 mg of glass beads (Sigma G1145), the cells were immediately lysed by bead beating with the Precellys Evolution homogenizer at the highest speed for 4× 45s cycles. After removal of the beads by spinning at 12,000 rpm for 5 min, the liquid phase was transferred to a new tube. 200 µl of chloroform was added and the sample was gently inverted several times until reaching homogeneity. The sample was then incubated for 15 min at room temperature before centrifugation at 12,000 g for 10 min. The upper phase (~600 µL) was gently collected and mixed at a 1:1 ratio with 100% isopropanol. The mixture was incubated for 1 hr at -20 °C and then centrifuged at 14,000 rpm for 10 min at 4 °C. The pellet was washed twice with 1 mL of 75% ethanol, air dried for 5 min, and dissolved in nuclease-free water. RNA integrity was assessed with 1% agarose gel and Agilent 2100 Bioanalyzer System.

[000253] Library preparation for total RNA SEnd-seq. The protocol for SEnd-seq library preparation was slightly modified from aspects of our method described in Example 1 and ref 8, such that processed RNA was labeled at the 5' end. In order to label the 5' end of processed RNA, 5-10 µg of total RNA in 12 µl volume was used to incubate with a 5' adaptor ligation mix (1 µl of 100 µM 5' adaptor, 0.5 µl of 50 mM ATP, 2 µl of dimethyl sulfoxide, 5 µl of 50% PEG8000, 1 µl of RNase Inhibitor and 1 µl of High Concentration T4 RNA Ligase 1) at 23 °C for 5 hr. Then the sample was diluted with water and cleaned twice with 1.5× vol of Agencourt RNAClean XP beads (Beckman Coulter, A63987). The eluted RNA was ligated to a 3' adaptor in the same way as the 5'-adaptor ligation. The 5' and 3' adaptors of Example 1 were utilized for this study. After incubation at 23 °C for 5 hr, the reaction was diluted to 40 µl with water and purified twice with 1.5× vol of Agencourt RNAClean XP beads to remove excess RNA adaptors. The sample was subsequently eluted with 0.1× TE buffer (10 mM Tris-HCl pH 7.5 and 1 mM EDTA) and subjected to rRNA removal with RiboMinus Transcriptome Isolation Kit (ThermoFisher, K155004). After RNA recovery by ethanol precipitation, the RNA was reverse transcribed with maturase from *Eubacterium rectale* (recombinantly purified from *E. coli*, a gift from Anna Marie Pyle, Yale University) (9) and biotinylated primers. After thorough cleaning, the cDNA was circularized by the TS2126 RNA Ligase I. Double-stranded DNA was generated with DNA polymerase I (New England BioLabs, M0209S) and subsequently fragmented by dsDNA Fragmentase (New England BioLabs, M0348S) at 37 °C for 15 min. The reaction was stopped by adding 5 µl of 0.5 M EDTA and heated at 65 °C for 15 min in the presence of 50 mM DTT. The DNA was diluted to 40 µl with TE buffer and purified with 1× vol of AMPure beads. The eluted DNA was used for sequencing library preparation with the NEBNext Ultra II DNA Library Prep Kit (New England BioLabs, E7645). The DNA library was amplified for 12 (for total RNA SEnd-seq) to 15 cycles (for primary RNA SEnd-seq) following the manufacturer's protocol.

[000254] Library preparation for primary transcripts. 5 µg of total RNA was used for primary transcript enrichment in accordance with the method described in Example 1 and reported in ref 8. The eluted RNA was used for SEnd-seq library preparation with 15 PCR cycles of amplification.

[000255] CRISPRi plasmid construction. The CRISPR-Cas9 system from *Streptococcus thermophilus* (dCas9_{Sth1}) (10) (a gift from the Rock Lab at The Rockefeller University) was used to modulate the expression level of target genes in the Mtb cell. Individual sgRNAs were designed and cloned into the dCas9_{Sth1} expression vector pIRL58. After verification by Sanger sequencing, the plasmid was transformed into Mtb cells by electroporation with the BioRad

GenePluser (2,500V, 700 ohm, 25 uF). Single colonies were then picked in the presence of kanamycin selection. Target gene knockdown was induced by 100 ng/L anhydrotetracycline. The knockdown efficiency was tested by qPCR with primers targeting the downstream part of the target gene.

[000256] DNase treatment and quantitative PCR. 2-10 µg of total RNA was treated with 0.5 µl of TURBO DNase (Life Technologies, AM2238) at 37 °C for 15 min. The RNA was then cleaned three times with 100 µl of TE-saturated phenol:chloroform:isoamyl alcohol (25:24:1, vol/vol). After recovery by ethanol precipitation, the RNA was reverse transcribed to cDNA with the high-capacity cDNA reverse transcription Kit (ThermoFisher, 4368814). qPCR was conducted with synthesized primers and SYBR green master mix (ThermoFisher, 4309155) on a QuantStudio 6 Flex Real-Time PCR System (Thermo Fisher Scientific). The relative abundance of RNA is represented as the signal ratio between the target transcript and the reference 16S rRNA gene from the same sample using the formula: $2^{-(\Delta CT)}$ ($\Delta CT = CT_{\text{target}} - CT_{16S}$; CT stands for cycle threshold).

[000257] In vitro transcription. DNA templates for Mtb RNAP were amplified by PCR from *M. tuberculosis* H37Rv genomic DNA with indicated primer sets. The AP3 promoter sequence was incorporated at one or both ends of the template. Purified Mtb RNAP and σ^A /RbpA holoenzyme (1:5 ratio, a gift from the Darst Lab at The Rockefeller University) was used for in vitro transcription reactions. The reaction mixture included 2 µl of 10× transcription buffer (100 mM Tris-HCl pH 7.9, 0.5 M KCl, 100 mM MgCl₂, 10 mM DTT, 50 µg/mL BSA), 0.5 µl of RNase Inhibitor, 0.5 pmol of DNA template and 2 pmol of Mtb RNAP holoenzyme. When applicable, 20 pmol of CarD (a gift from the Darst Lab at The Rockefeller University) was added to the reaction mixture. The mixture was incubated at 37 °C for 15 min before rNTPs (200 µM each) were added to initiate transcription. After 10 min of reaction (unless noted otherwise), reinitiation of transcription was prevented by adding heparin (Sigma-Aldrich, H4784) to a final concentration of 100 µg/ml. After incubation with 0.3 µl of TURBO DNase for 10 min, the RNA was separated by 5% urea polyacrylamide gel electrophoresis, stained by SYBR Gold Nucleic Acid Gel Stain (Thermo Fisher Scientific, S11494), scanned by Axygen Gel Documentation System (Corning, GD1000), and quantified by ImageJ (National Institutes of Health).

[000258] Data analysis. SEnd-seq data were collected by the Illumina MiSeq, NextSeq 500 or Novaseq 6000 platform in a paired-end mode (150 nt ×2). The raw data were processed as described above in Example 1 and as reported (8). The full-length sequences were inferred by mapping to the reference genome using Bowtie 2. Reads with an insert length greater than

10,000 nt were discarded. The TSS, TTS and gene coverage data were extracted as described above in Example 1 and as reported (8).

[000259] RESULTS

[000260] We have collected initial SEnd-seq datasets for *Mtb* upon different growth conditions, such as log phase, stationary phase and treatment of different antibiotic drugs (Figure 32). In contrast to *E. coli*, *Mtb* generally activates the transcription of most of its genes in a single condition, and the boundary between different genes is not clear. It is interesting to note there are a lot of antisense RNA transcripts within the *Mtb* transcriptome, and the conflicts between sense and antisense transcripts could stop transcription elongation of either direction. *E. coli* and *Mtb* share very similar genome size, however many more TSSs were found in the *Mtb* transcriptome, whereas much less TTSs were detected (Figure 33). Therefore, we expect that the regulatory mechanisms of RNA transcription such as elongation and termination in *Mtb* cells are very different from *E. coli*.

[000261] REFERENCES

- 1 Khan, M. K., Islam, M. N., Ferdous, J. & Alam, M. M. An Overview on Epidemiology of Tuberculosis. *Mymensingh Med J* **28**, 259-266 (2019).
- 2 Drug-resistant tuberculosis. *World Health Organization* (2018).
- 3 Nunes-Alves, C. *et al.* In search of a new paradigm for protective immunity to TB. *Nat Rev Microbiol* **12**, 289-299, doi:10.1038/nrmicro3230 (2014).
- 4 Cortes, T. *et al.* Genome-wide mapping of transcriptional start sites defines an extensive leaderless transcriptome in *Mycobacterium tuberculosis*. *Cell Rep* **5**, 1121-1131, doi:10.1016/j.celrep.2013.10.031 (2013).
- 5 Arnvig, K. B. *et al.* Sequence-based analysis uncovers an abundance of non-coding RNA in the total transcriptome of *Mycobacterium tuberculosis*. *PLoS Pathog* **7**, e1002342, doi:10.1371/journal.ppat.1002342 (2011).
- 6 Gerrick, E. R. *et al.* Small RNA profiling in *Mycobacterium tuberculosis* identifies MrsI as necessary for an anticipatory iron sparing response. *Proc Natl Acad Sci U S A* **115**, 6464-6469, doi:10.1073/pnas.1718003115 (2018).
- 7 Rustad, T. R. *et al.* Global analysis of mRNA stability in *Mycobacterium tuberculosis*. *Nucleic Acids Res* **41**, 509-517, doi:10.1093/nar/gks1019 (2013).
- 8 Ju, X., Li, D. & Liu, S. Full-length RNA profiling reveals pervasive bidirectional transcription terminators in bacteria. *Nat Microbiol* **4**, 1907-1918, doi:10.1038/s41564-019-0500-z (2019).
- 9 Zhao, C., Liu, F. & Pyle, A. M. An ultraprocessive, accurate reverse transcriptase encoded by a metazoan group II intron. *Rna* **24**, 183-195, doi:10.1261/rna.063479.117 (2018).
- 10 Rock, J. M. *et al.* Programmable transcriptional repression in mycobacteria using an orthogonal CRISPR interference platform. *Nat Microbiol* **2**, 16274, doi:10.1038/nmicrobiol.2016.274 (2017).

PAGE INTENTIONALLY LEFT BLANK

CLAIMS

1. A method of full-length transcript profiling of RNA, said method comprising the steps of:
 - (a) isolating genomic RNA;
 - (b) ligating an adaptor to 3' end of genomic RNA to provide 3'-end ligated genomic RNA;
 - (c) converting the 3'-end ligated genomic RNA to cDNA;
 - (d) circularizing the cDNA;
 - (e) fragmenting the cDNA;
 - (f) enriching the cDNA fragments containing the 5'-3' junction;
 - (g) sequencing the enriched cDNA fragments to obtain 5'- and 3'-end sequences; and
 - (h) mapping the sequences to a reference genome to provide a full-length transcript profile of genomic RNA.
2. The method according to claim 1, further comprising:

enriching primary transcripts from the genomic RNA obtained from step (a);

wherein ligating an adaptor to the 3' end of genomic RNA of step (b) further comprises capping 5'-triphosphorylated primary RNA, and isolating the capped 5'-triphosphorylated primary RNA.
3. The method according to claim 2, wherein capping comprises capping with a biotin moiety.
4. The method according to claim 3, wherein isolating the capped 5'-triphosphorylated primary RNA comprises isolating the primary RNA with streptavidin-coated magnetic beads.
5. The method according to any of claims 1-4, further comprising:

enriching processed transcripts from the genomic RNA obtained from step (a);

wherein ligating an adaptor to 3' end of genomic RNA of step (b) comprises ligating adaptors to the 5' ends and 3' ends of genomic RNA of step (b), and

converting genomic RNA to cDNA of step (c) comprises converting end-ligated processed RNA to cDNA.

6. The method according to claim 5, wherein enriching processed RNA comprises ligating 5'-monophosphorylated processed RNA to a 5' adaptor.
7. The method according to any of claims 1-6, wherein the genomic RNA comprises prokaryotic genomic RNA.
8. The method according to claim 7, wherein the prokaryotic genomic RNA comprises bacterial genomic RNA.
9. The method according to any of claims 7-8, wherein the prokaryotic genomic RNA comprises pathogenic bacterial genomic RNA.
10. The method according to any of claims 7-9, wherein the prokaryotic genomic RNA comprises *Mycobacterium tuberculosis* genomic RNA.
11. The method according to any of claims 7-10, wherein the prokaryotic genomic RNA comprises genomic RNA from multi-drug resistant bacteria.
12. The method according to any of claims 1-11, wherein the genomic RNA comprises genomic RNA from one or more prokaryotic species.
13. The method according to any of claims 1-11, wherein the genomic RNA comprises genomic RNA from a microbiome.
14. The method according to claim 13, wherein the microbiome comprises gut, skin, animal rumen, or plant associated microbiomes.
15. The method according to any of claims 1-14, wherein mapping the sequences to the reference genome comprises:
 - merging paired-end reads into single-end reads; and
 - inferring full-length sequences by mapping to the reference genome.
16. The method according to any of claims 1-15, wherein mapping the sequences to the reference genome further comprises identification of transcription start sites (TSS).
17. The method according to any of claims 1-16, wherein mapping the sequences to the reference genome further comprises identification of transcription termination sites (TTS).

18. The method according to any of claims 1-14, wherein circularizing the cDNA comprises contacting the cDNA of step (b) with a single-stranded nucleic acid ligase.
19. The method according to claim 18, wherein the single-stranded nucleic acid ligase does not ligate double-stranded nucleic acid.
20. The method according to any of claims 18-19, wherein the single-stranded nucleic acid ligase provides more than 75% intramolecular ligation efficiency.
21. The method according to any of claims 18-19, wherein the single-stranded nucleic acid ligase provides more than 80% intramolecular ligation efficiency.
22. The method according to any of claims 18-19, wherein the single-stranded nucleic acid ligase provides more than 90% intramolecular ligation efficiency.
23. The method according to any of claims 18-22, wherein the single-stranded nucleic acid ligase comprises an RNA ligase.
24. The method according to any of claims 18-23, wherein the single-stranded nucleic acid ligase comprises a thermostable RNA ligase.
25. The method according to any of claims 18-24, wherein the single-stranded nucleic acid ligase comprises Ts2126 ligase.
26. The method according to any of claims 1-14, wherein said cDNA comprises a tag.
27. The method according to claim 28, wherein said tag comprises a biotin moiety.
28. The method according to any of claims 1-14, wherein fragmented cDNA is single-stranded DNA and is converted to double-stranded DNA.
29. The method according to any of claims 1-14, wherein spike-in RNA is added prior to step (b).
30. The method according to any of claims 1-14, wherein providing the full-length profile of a genomic transcript comprises identification of at least one transcript for at least one sequencing paired-end read.
31. A method of making a DNA library of RNA transcripts, said method comprising the steps of:
 - (a) isolating genomic RNA;

(b) ligating an adaptor to 3' end of genomic RNA to provide 3' end ligated genomic RNA;

(c) converting the 3'-end ligated genomic RNA to cDNA;

(d) circularizing the cDNA; and

(e) fragmenting cDNA to provide fragmented cDNA.

32. The method according to claim 31, wherein fragmented cDNA is ssDNA and is converted to dsDNA.

33. The method according to claim 31 or 32, wherein an adaptor is ligated to the fragmented cDNA.

34. The method according to any of claims 31-33, wherein fragmented cDNA is amplified.

35. The method according to any of claims 31-34, wherein spike-in RNA is added prior to step (b).

36. The method according to any of claims 31-35, further comprising:

enriching primary RNA from the genomic RNA obtained from step (a);

wherein ligating an adaptor to the 3' end of genomic RNA of step (b) comprises ligating an adaptor to the 3' end of primary RNA to provide 3'-end ligated primary transcripts, and converting genomic RNA to cDNA of step (c) comprises converting 3' end ligated primary RNA to cDNA.

37. The method according to claim 36, wherein enriching primary transcripts comprises: capping 5'-triphosphorylated primary RNA and isolating the capped 5'-triphosphorylated primary RNA.

38. The method according to claim 37, wherein capping comprises capping with a biotin moiety.

39. The method according to any of claims 31-35, further comprising:

enriching processed RNA from the genomic RNA obtained from step (a);

wherein ligating an adaptor to 3' end of genomic RNA of step (b) comprises ligating adaptors to the 5' ends and 3' ends of genomic RNA of step (b), and

converting genomic RNA to cDNA of step (c) comprises converting end-ligated processed RNA to cDNA.

40. The method according to claim 39, wherein enriching processed transcripts comprises ligating 5'-monophosphorylated processed RNA to a 5' adaptor.

41. The method according to any of claims 31-40, wherein circularizing the cDNA comprises contacting the cDNA of step (b) with a single-stranded nucleic acid ligase.

42. The method according to claim 41, wherein the single-stranded nucleic acid ligase does not ligate double-stranded nucleic acid.

43. The method according to any of claims 41-42, wherein the single-stranded nucleic acid ligase provides more than 75% intramolecular ligation efficiency.

44. The method according to any of claims 41-43, wherein the single-stranded nucleic acid ligase provides more than 80% intramolecular ligation efficiency.

45. The method according to any of claims 41-44, wherein the single-stranded nucleic acid ligase provides more than 90% intramolecular ligation efficiency.

46. The method according to any of claims 41-45, wherein the single-stranded nucleic acid ligase comprises an RNA ligase.

47. The method according to any of claims 41-46, wherein the single-stranded nucleic acid ligase comprises a thermostable RNA ligase.

48. The method according to any one of claims 41-47, wherein the single-stranded nucleic acid ligase comprises Ts2126 ligase.

49. A kit for simultaneous 5'- and 3'-end RNA sequence capture from isolated RNA comprising:

- (a) at least one labeled or tagged adaptor for ligation to the 3' ends of the isolated RNA;
- (b) a ligase capable of ligating the adaptor to the 3' ends of the isolated RNA;
- (c) a primer for reverse transcription of the 3'-end ligated RNA, wherein said primer is capable of specifically initiating transcription of the 3'-end ligated RNA to generate cDNA;

- (d) a reagent capable of binding to or otherwise having affinity for circularized cDNA derived from the 3'-end ligated RNA such that the reagent enriches for or otherwise isolates cDNA fragments comprising the 3'-end ligated sequence in combination with its respective 5' end; and
- (e) directions for use of said kit.

50. The kit of claim 49 further comprising a reverse transcriptase enzyme capable of generating cDNA from the 3'-end ligated RNA.

51. The kit of claim 49 or 50 further comprising a ligase capable of circularizing the generated cDNA of (c).

52. The kit of claim 51 further comprising a means or one or more component for fragmentation of the circularized cDNA after (c) such that cDNA fragments comprising the 3'-end ligated sequence in combination with its respective 5' end are generated.

53. The kit of any of claims 49-52 further comprising at least one adaptor for ligation to the 5' ends of the isolated RNA and a ligase capable of ligating the adaptor to the 5' ends of the isolated RNA before (c).

FIGURE 1-1

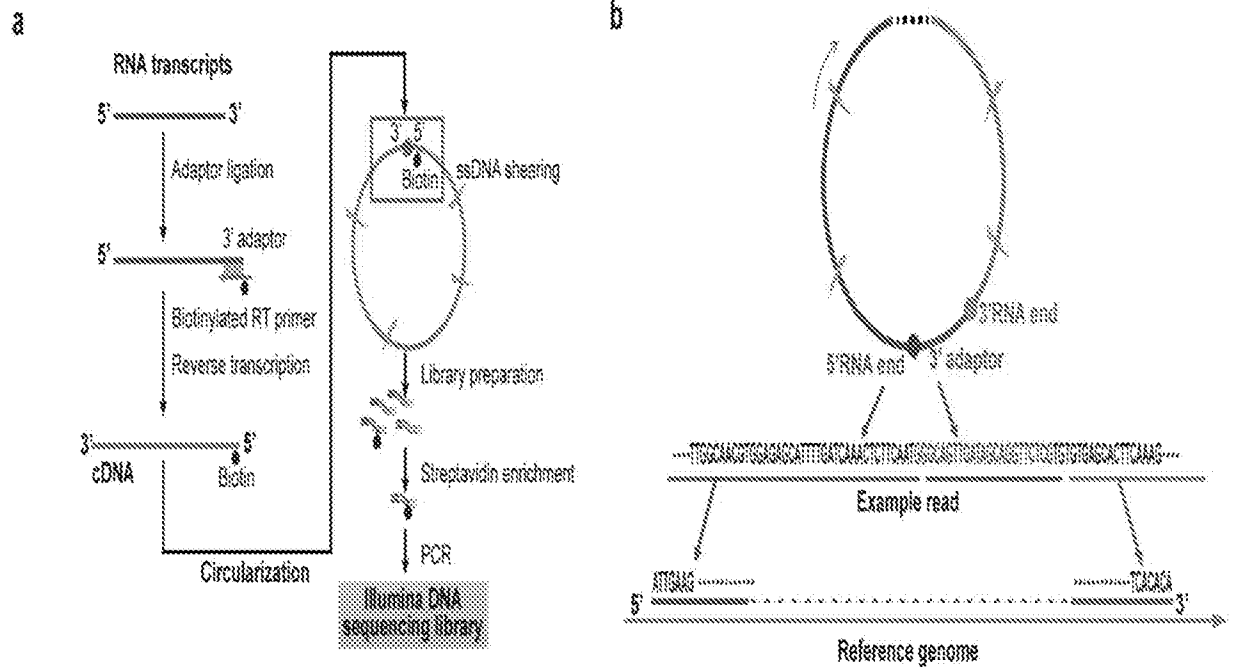


FIGURE 1-2

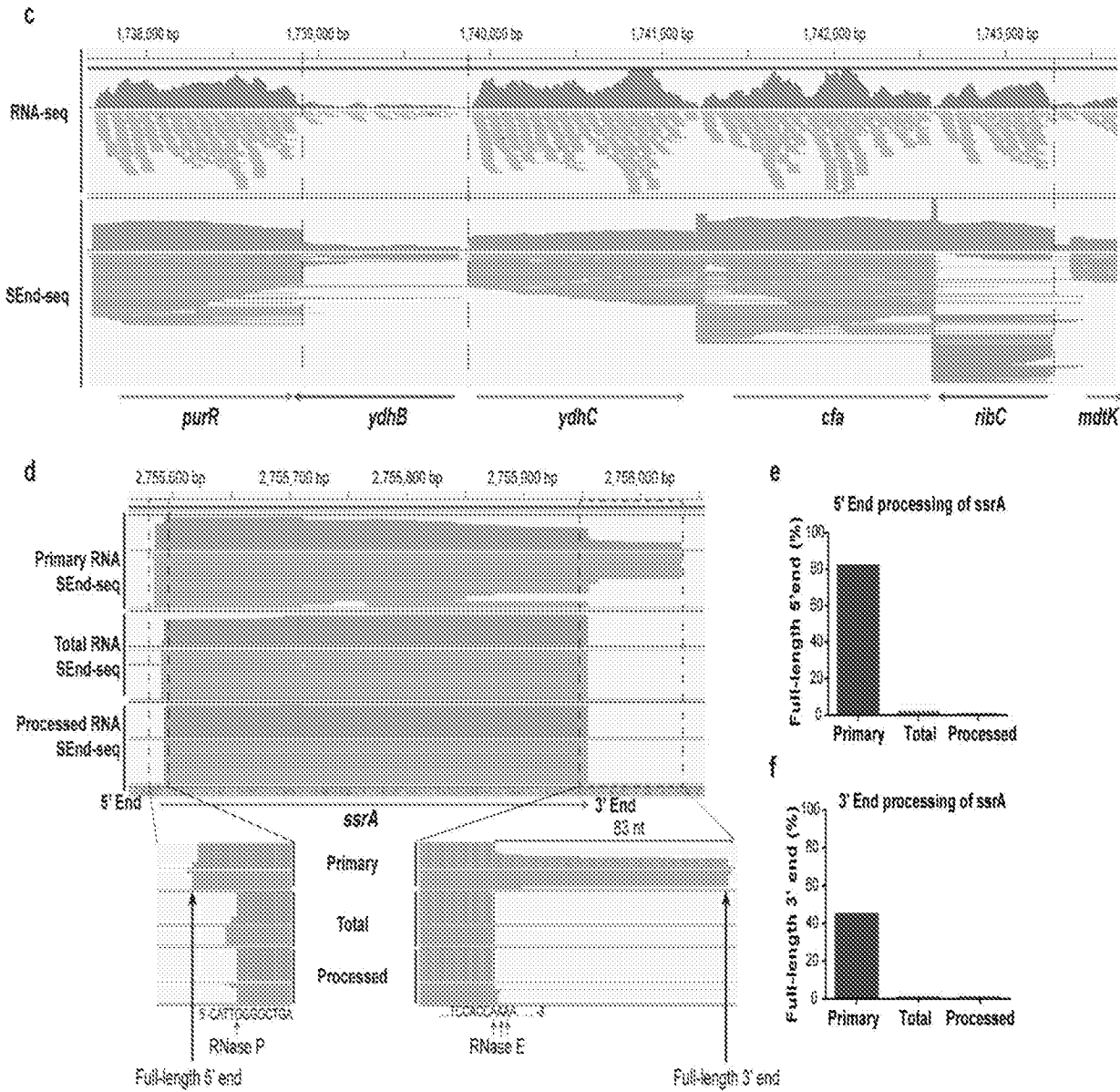


FIGURE 2-1

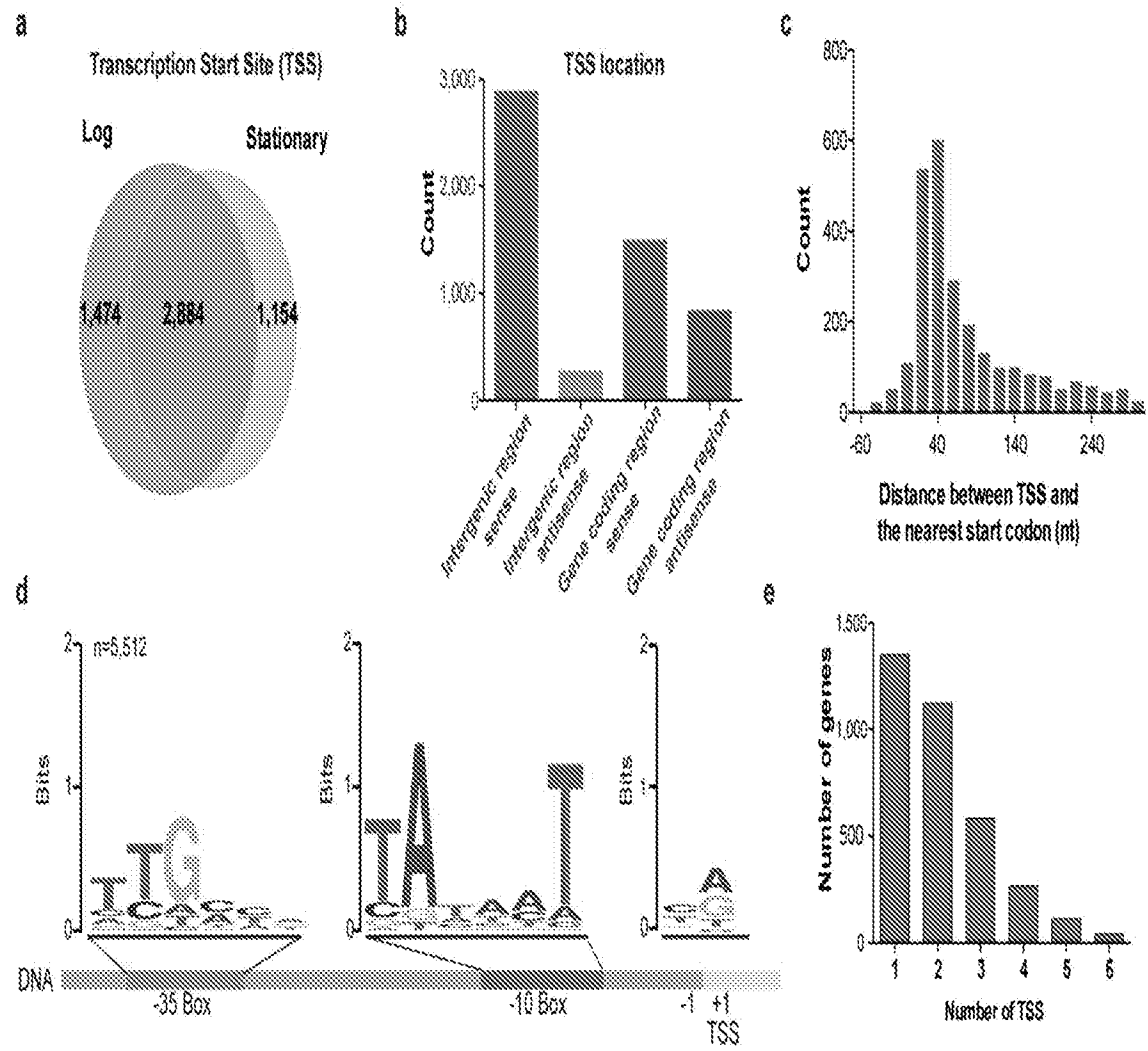


FIGURE 2-2

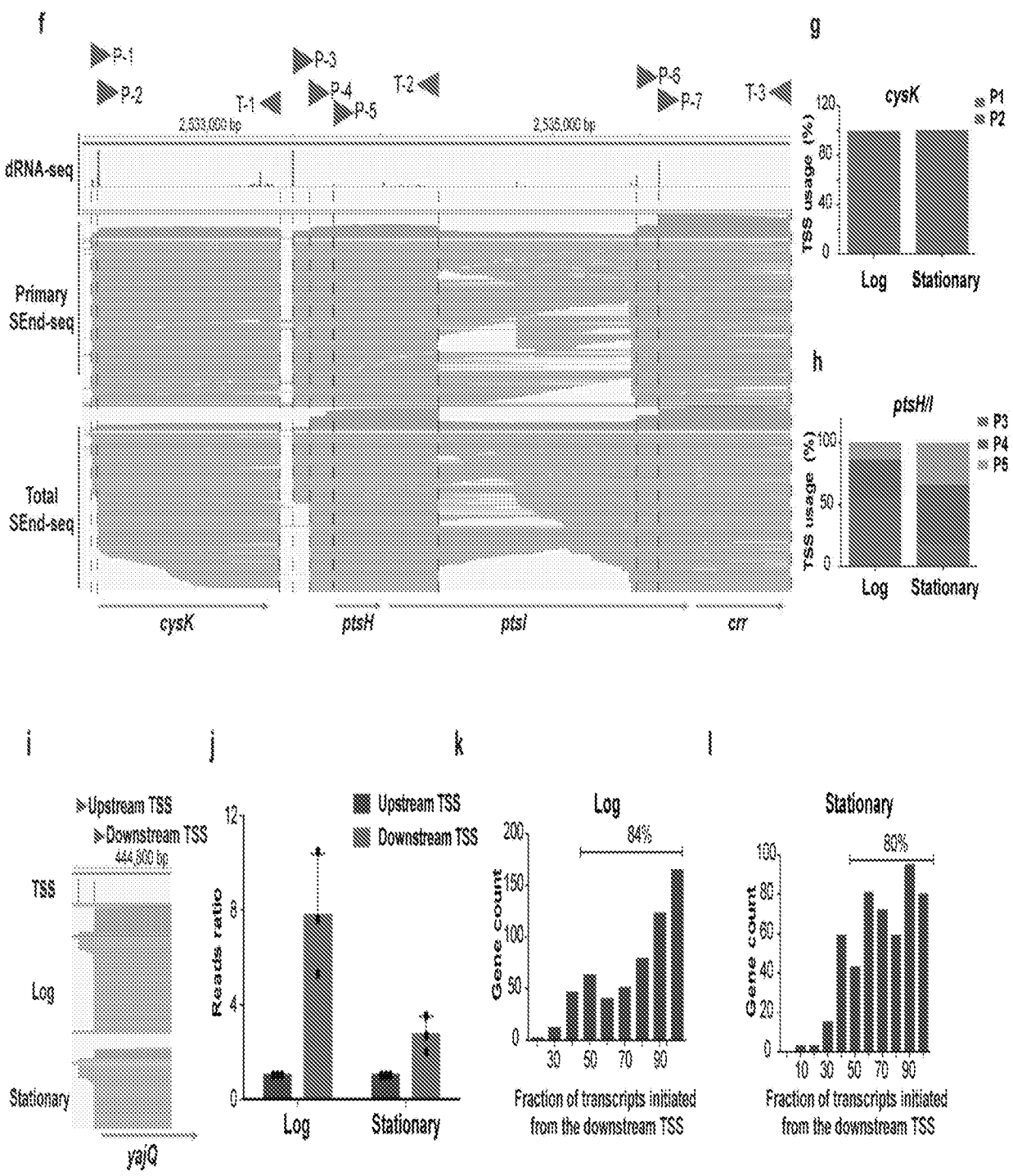


FIGURE 3-1

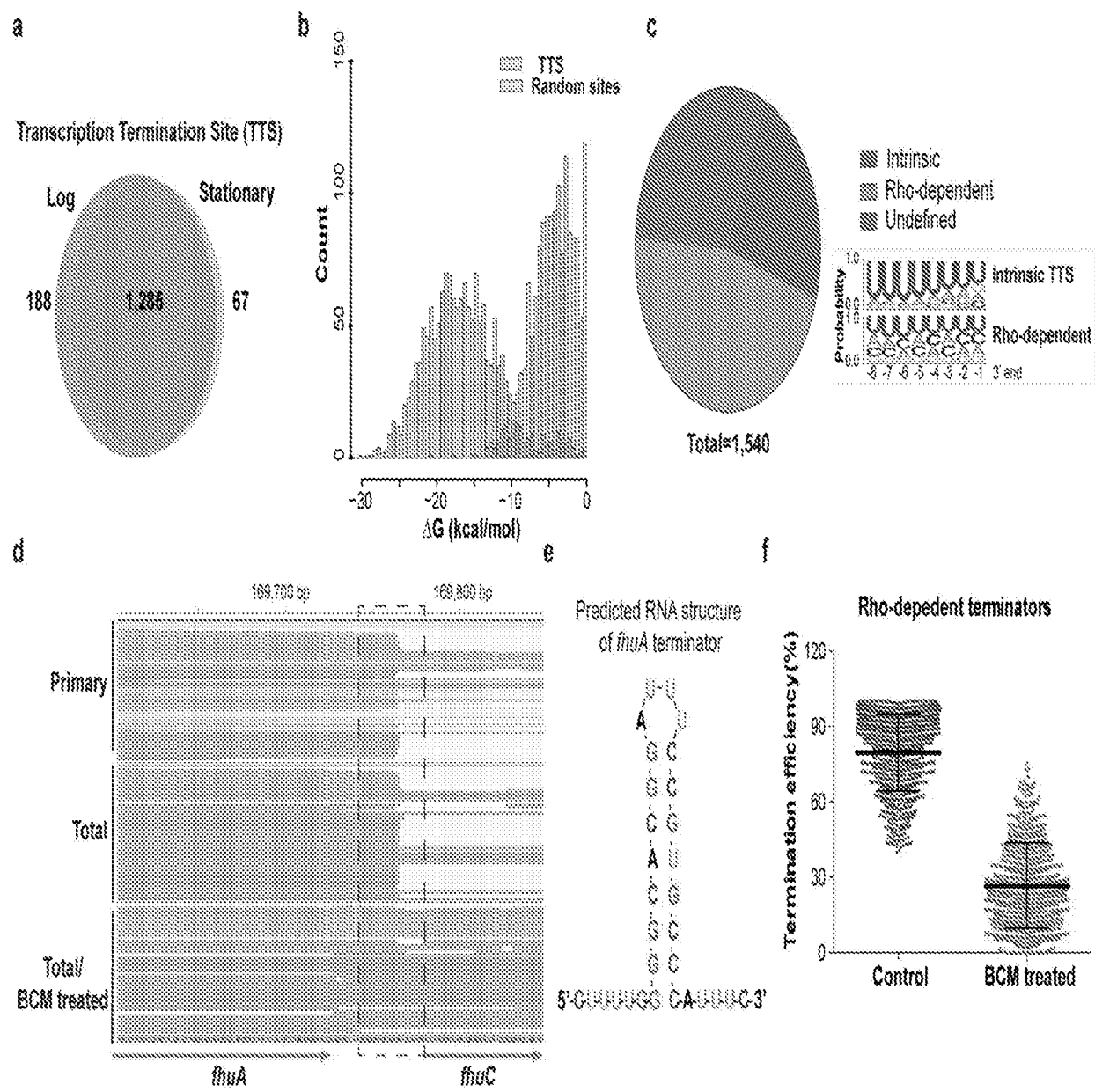


FIGURE 3-2

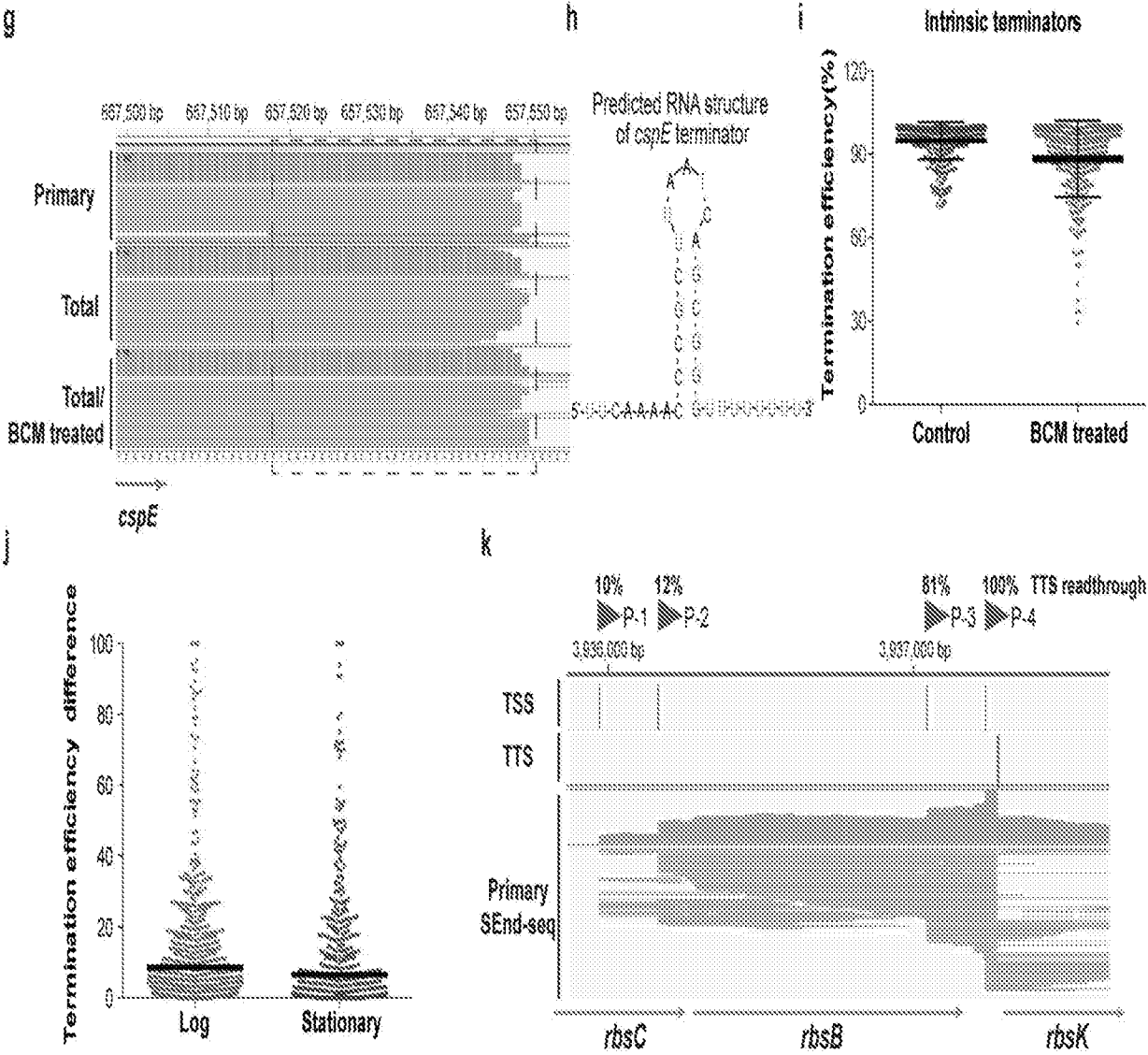


FIGURE 4-1

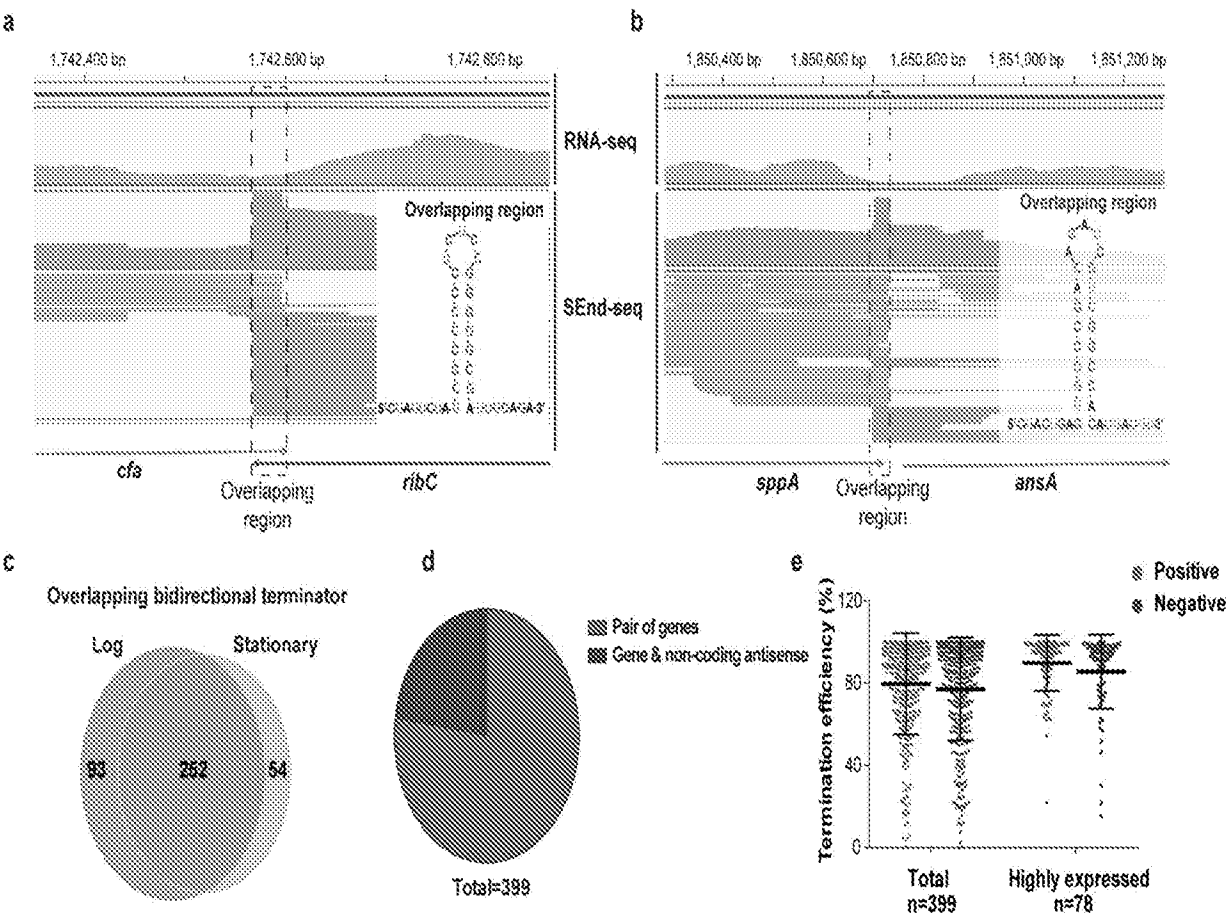


FIGURE 4-2

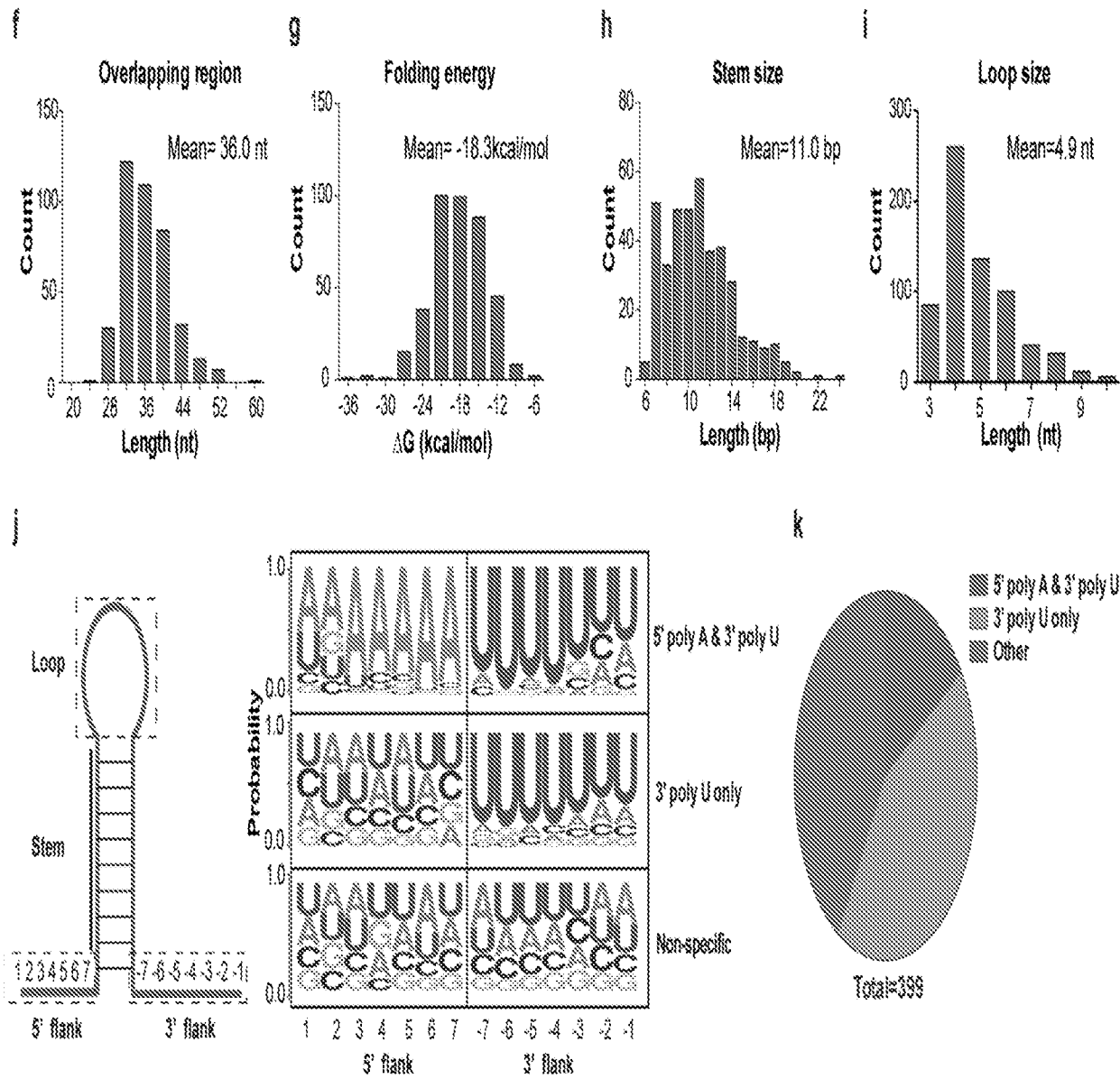


FIGURE 5-1

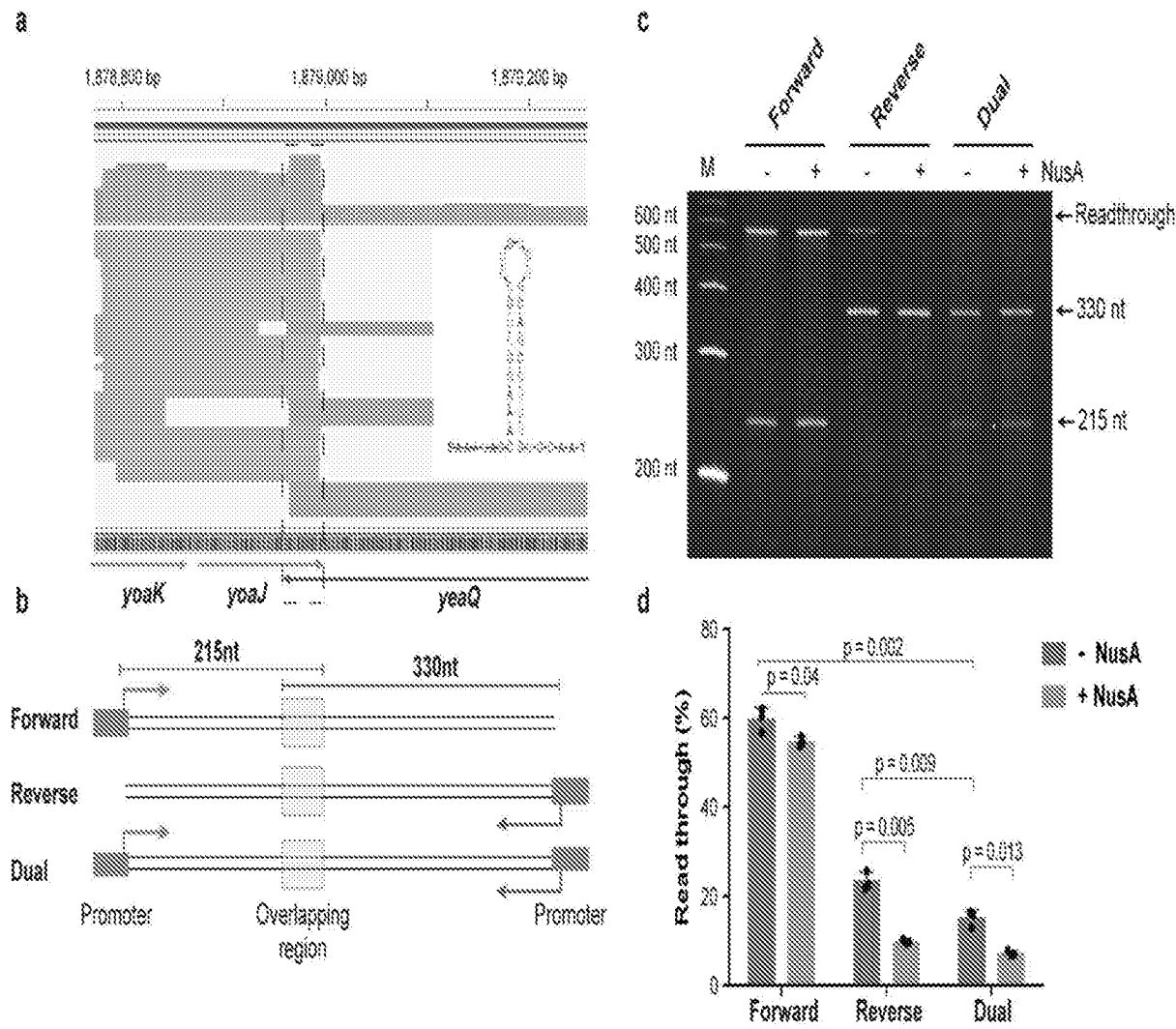


FIGURE 5-2

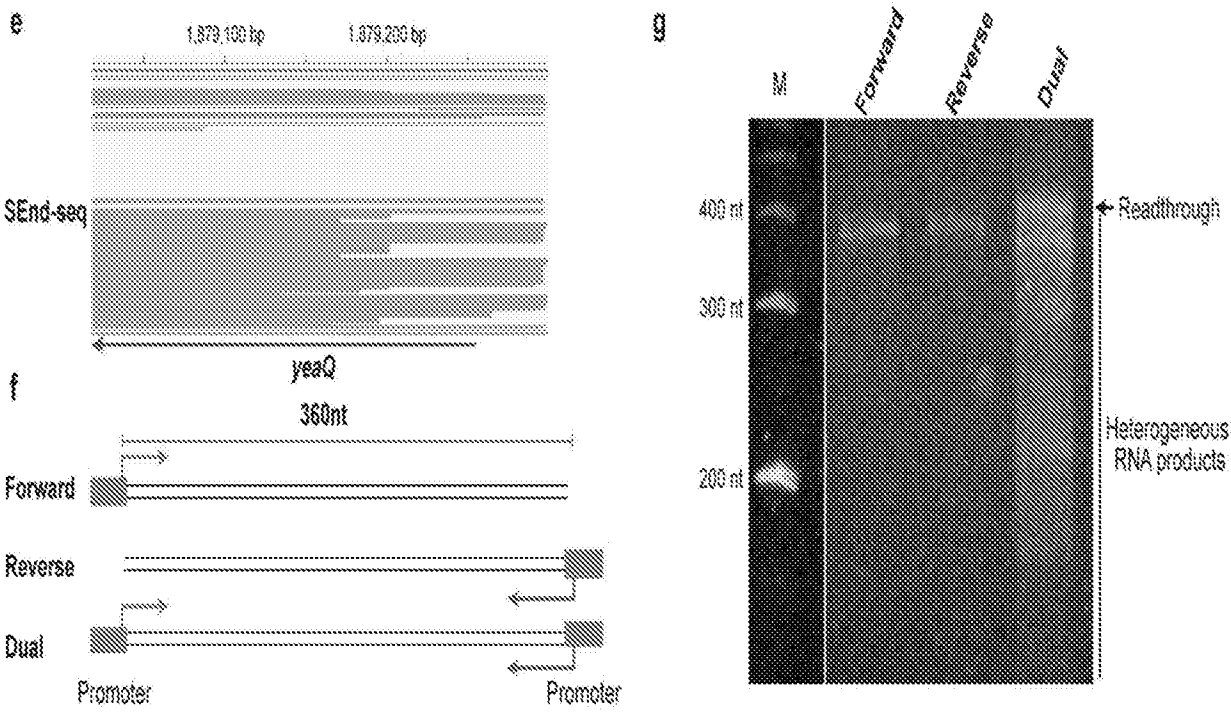


FIGURE 6-1

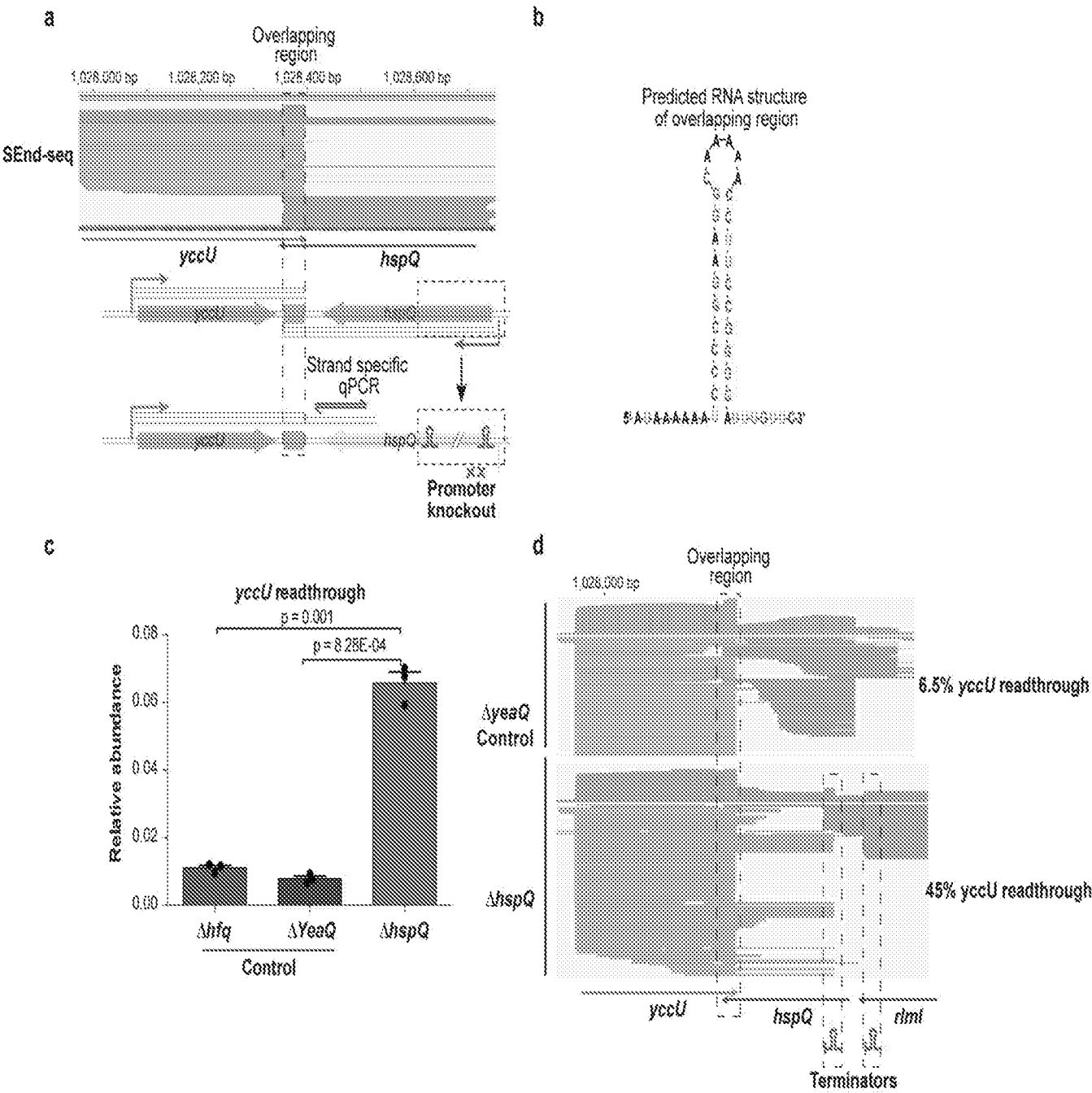


FIGURE 6-2

e

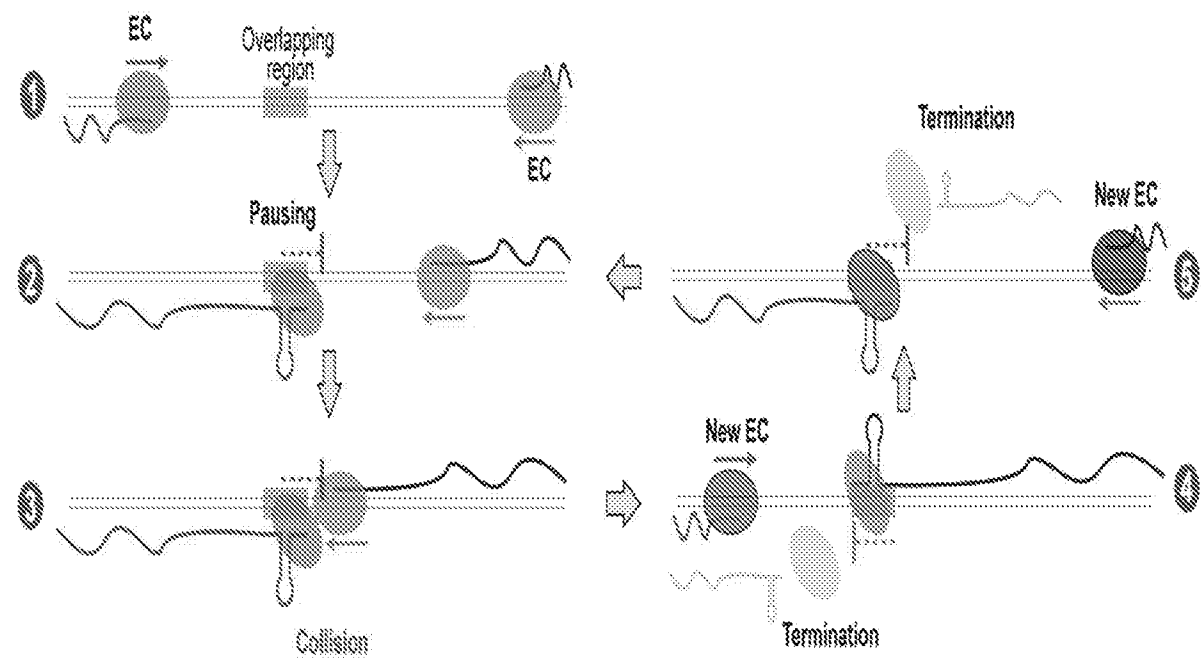


FIGURE 7

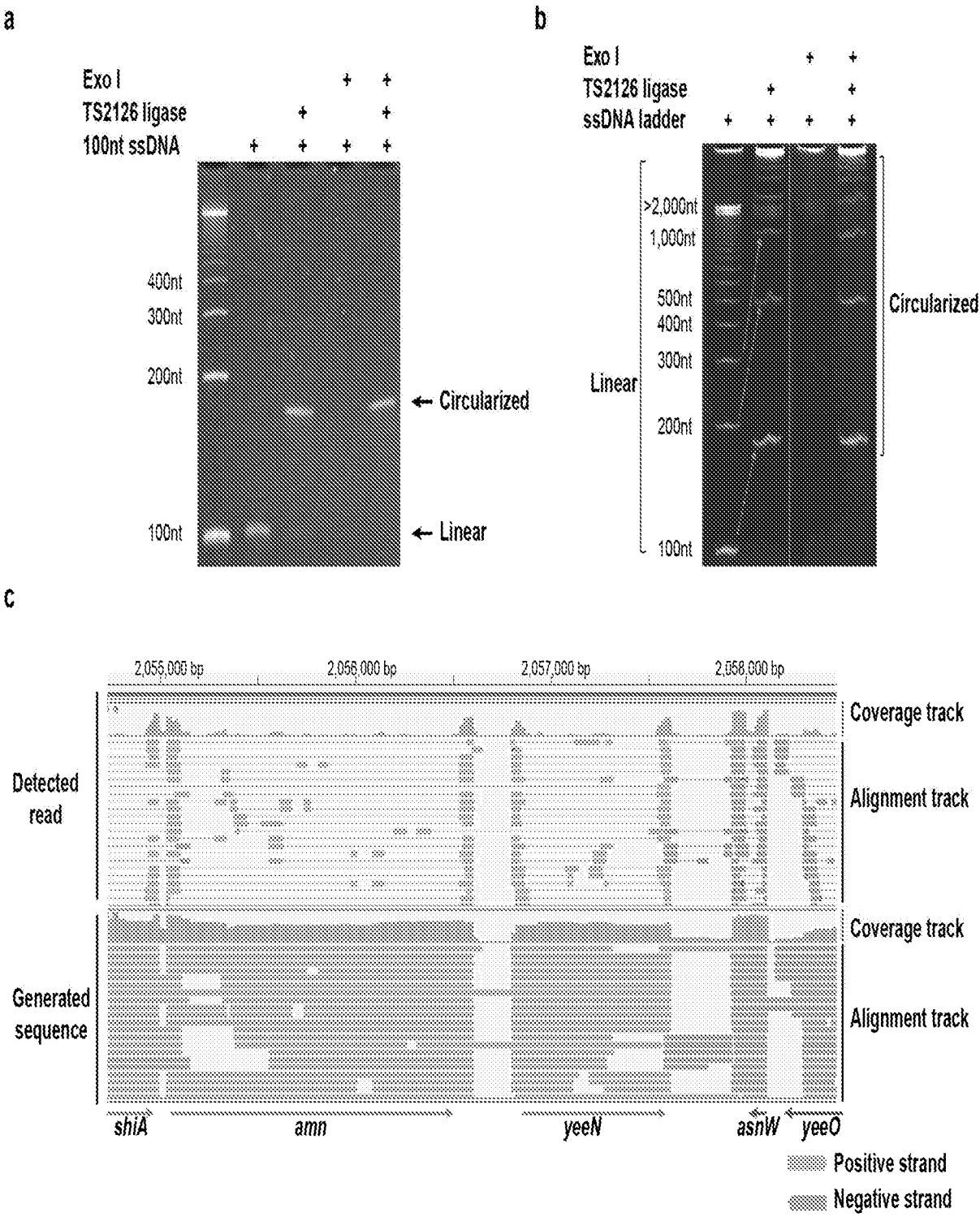


FIGURE 8

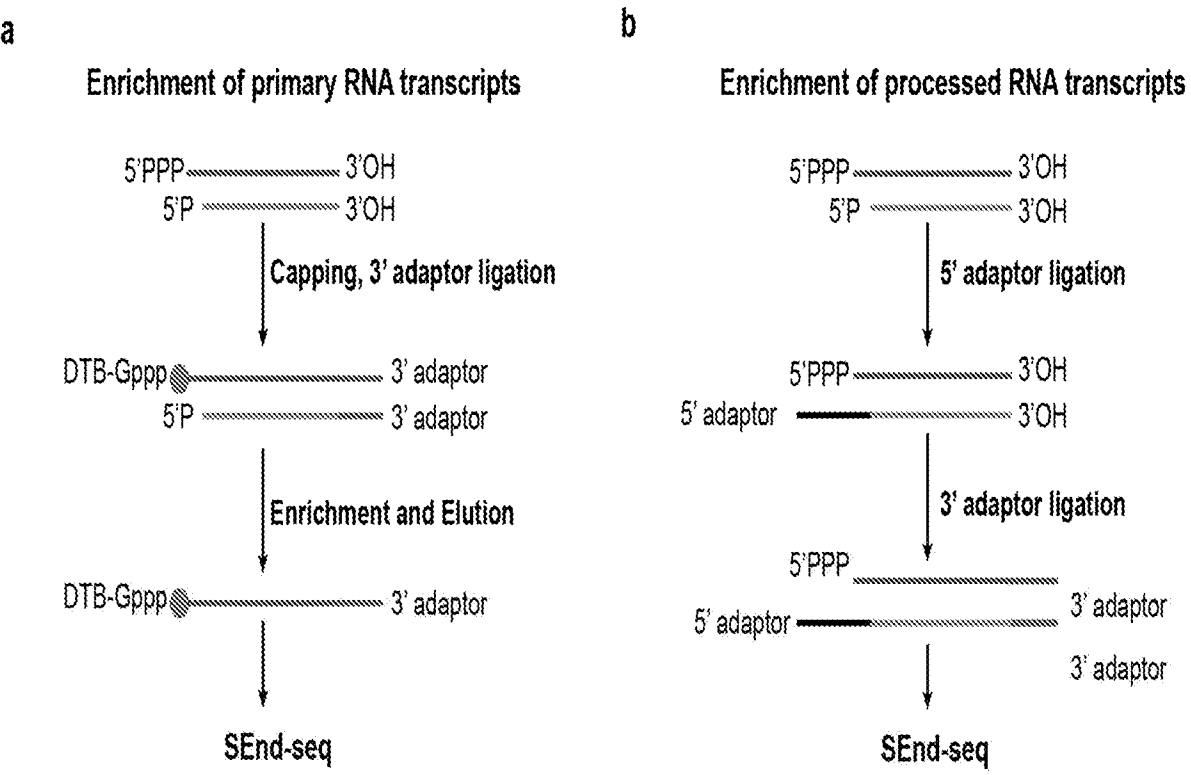


FIGURE 9-1

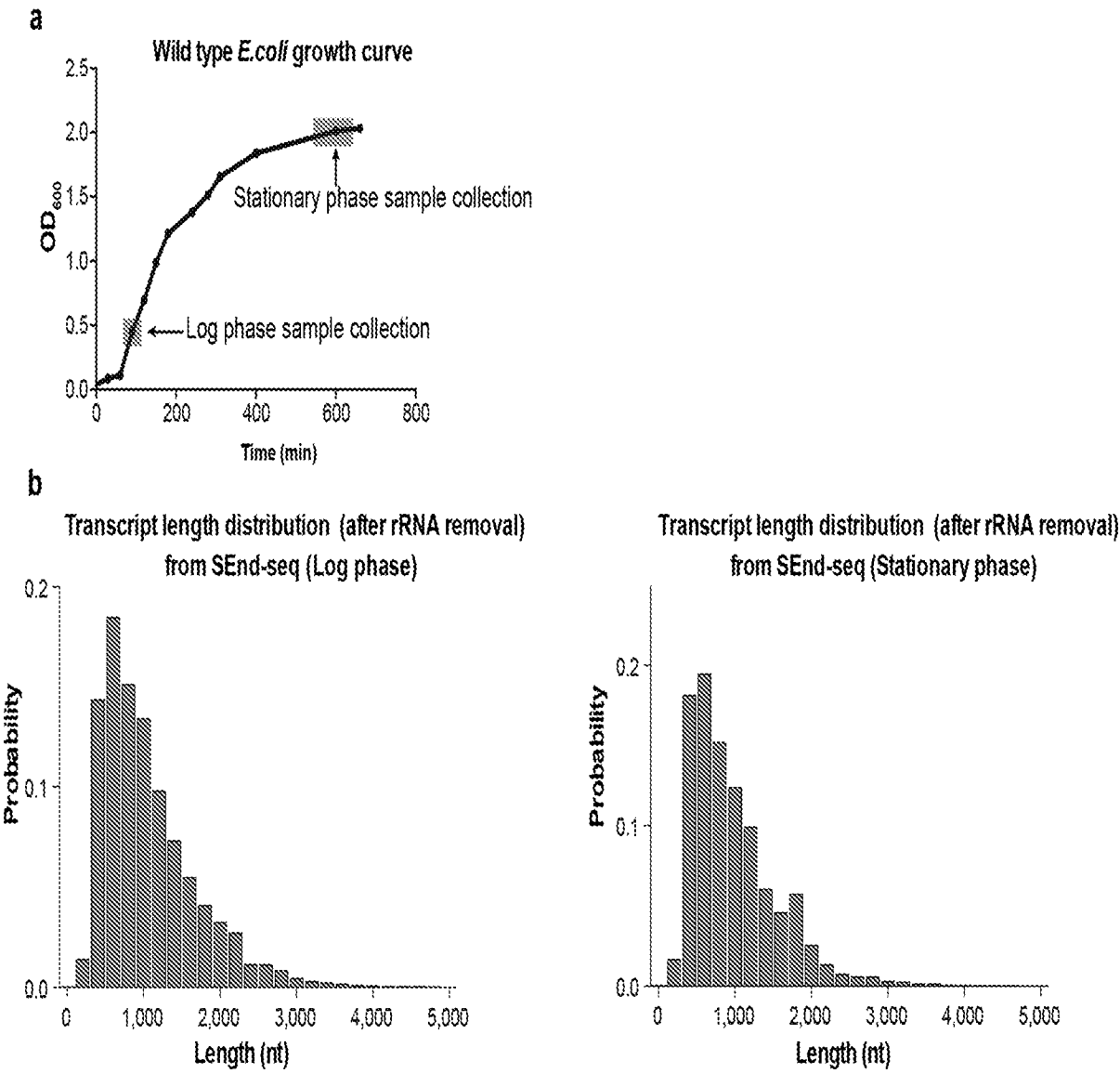


FIGURE 9-2

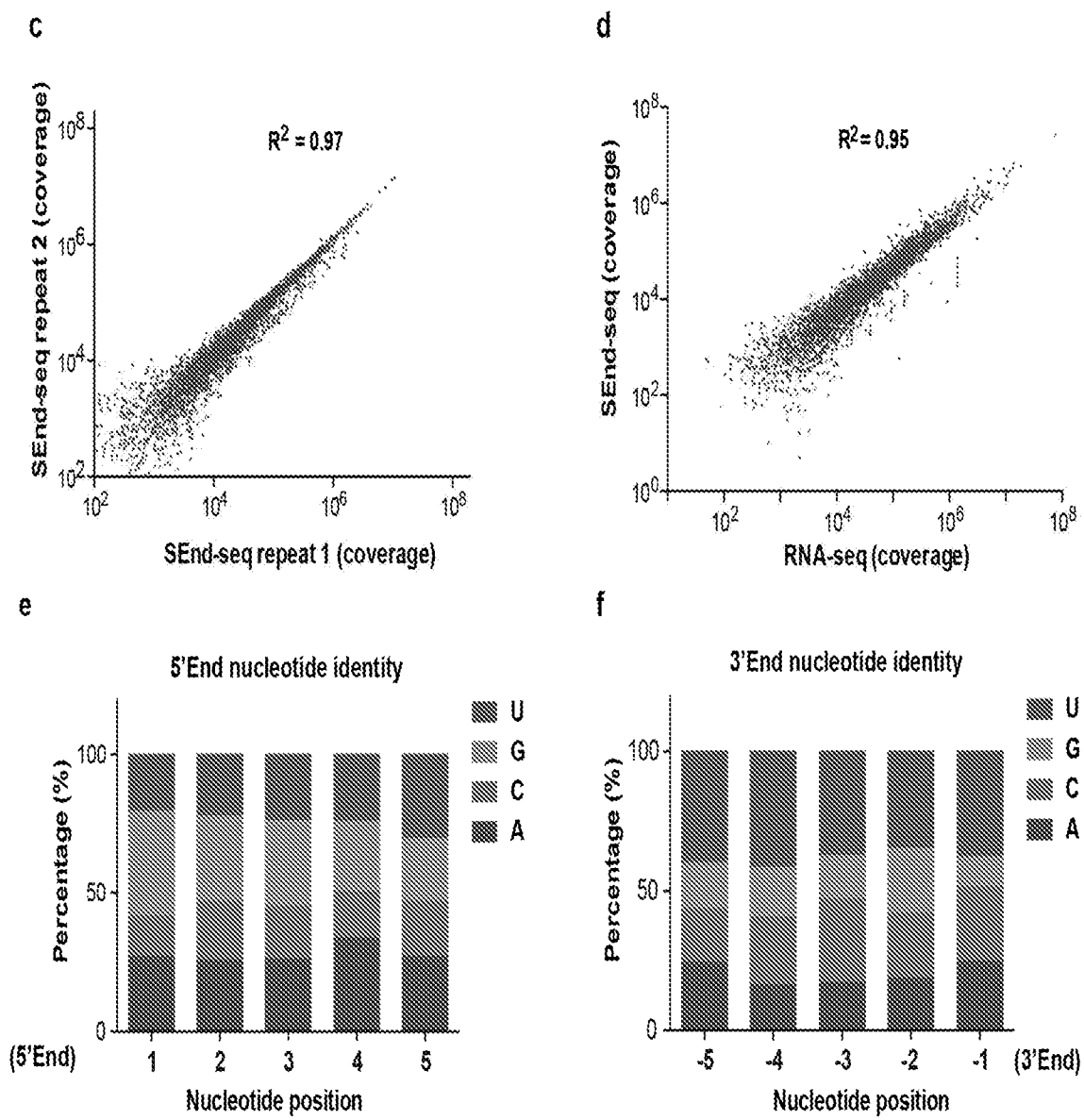


FIGURE 10

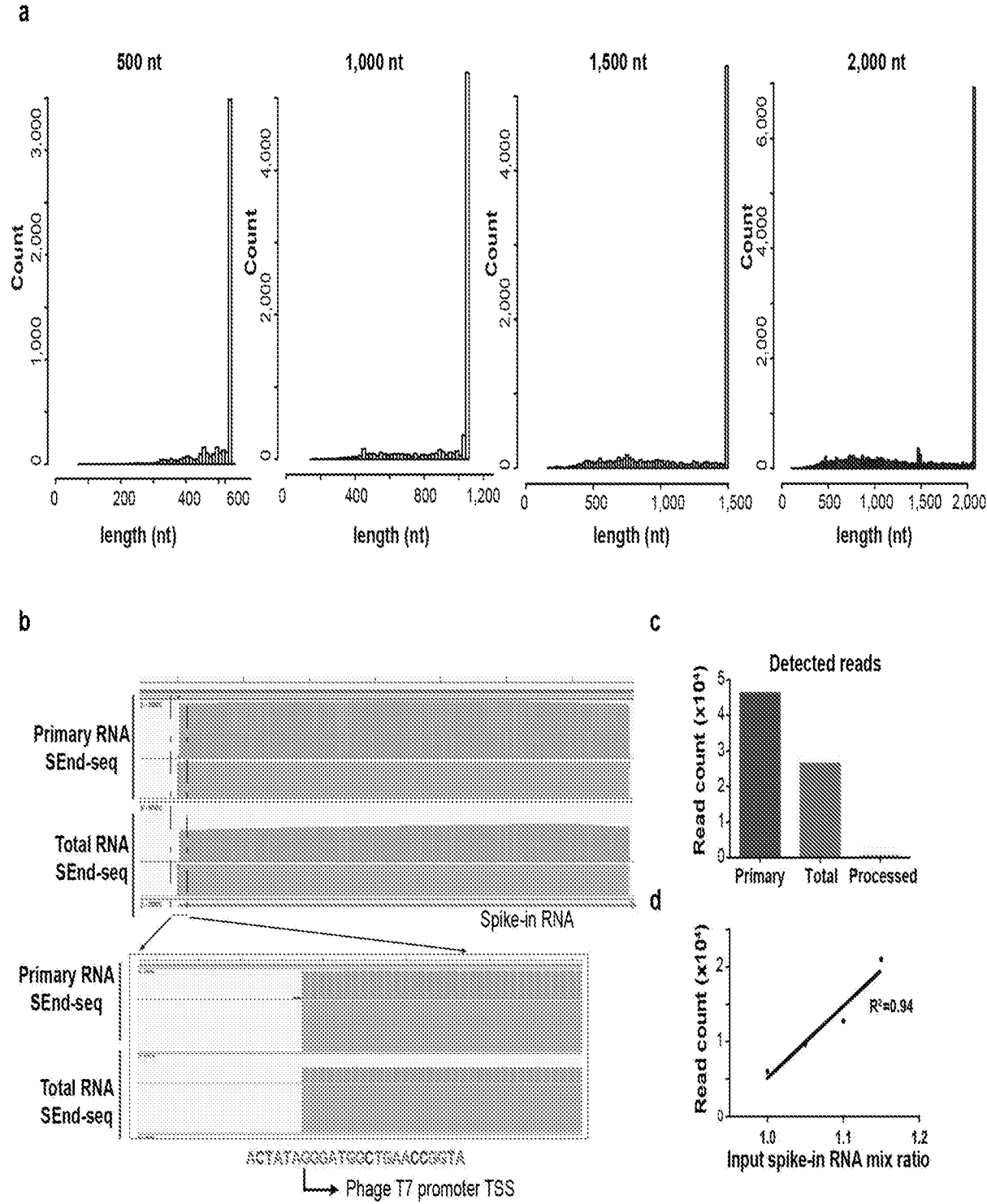


FIGURE 11

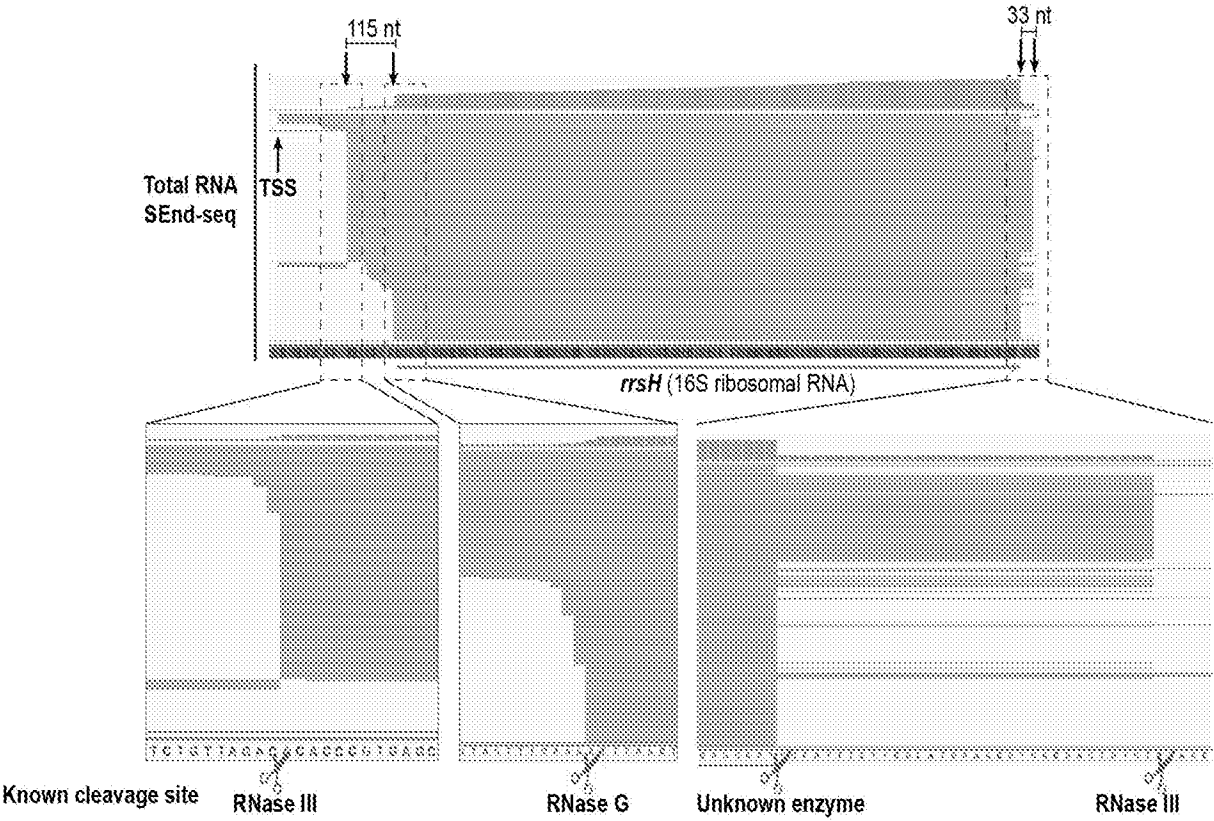


FIGURE 12

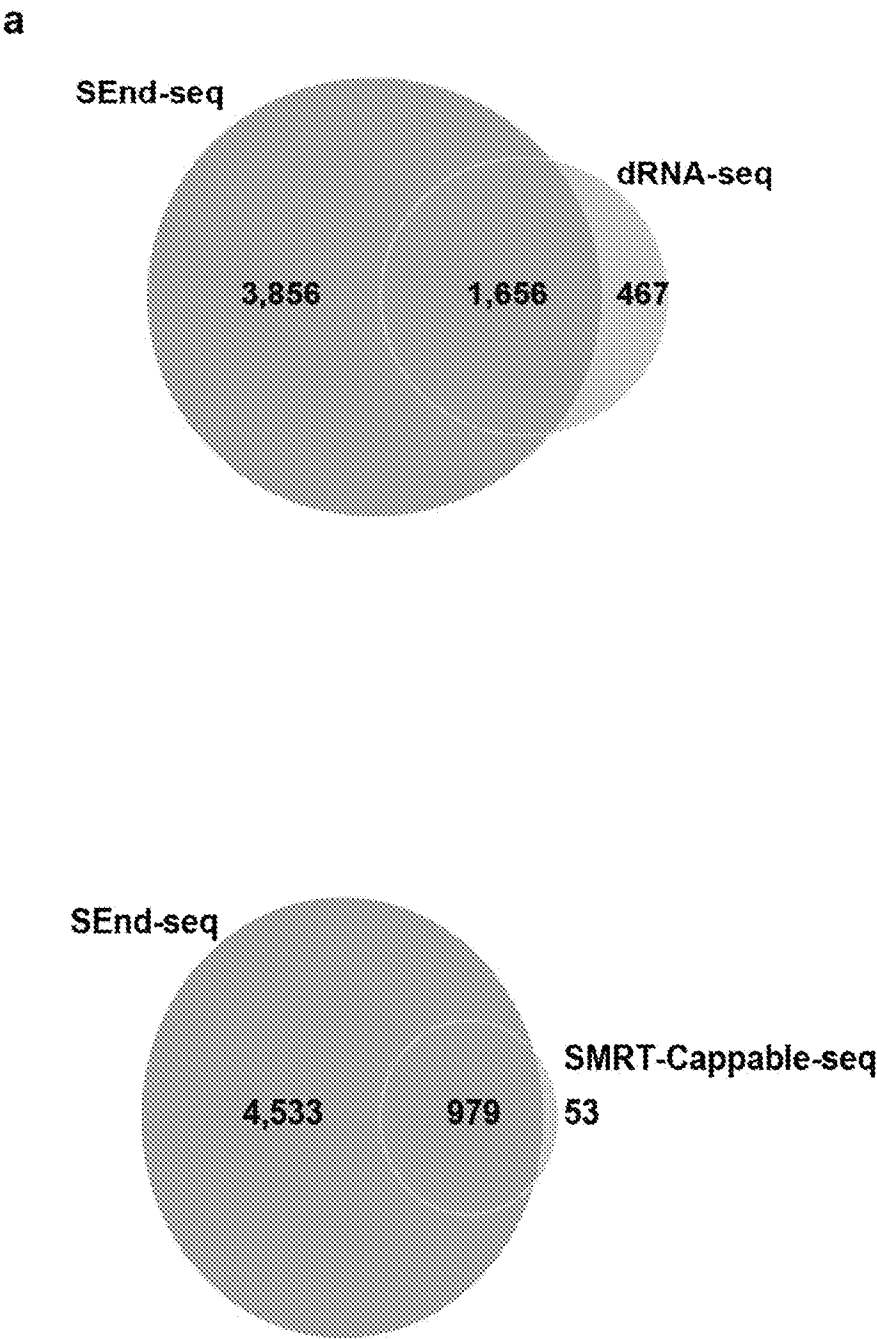


FIGURE 13-1

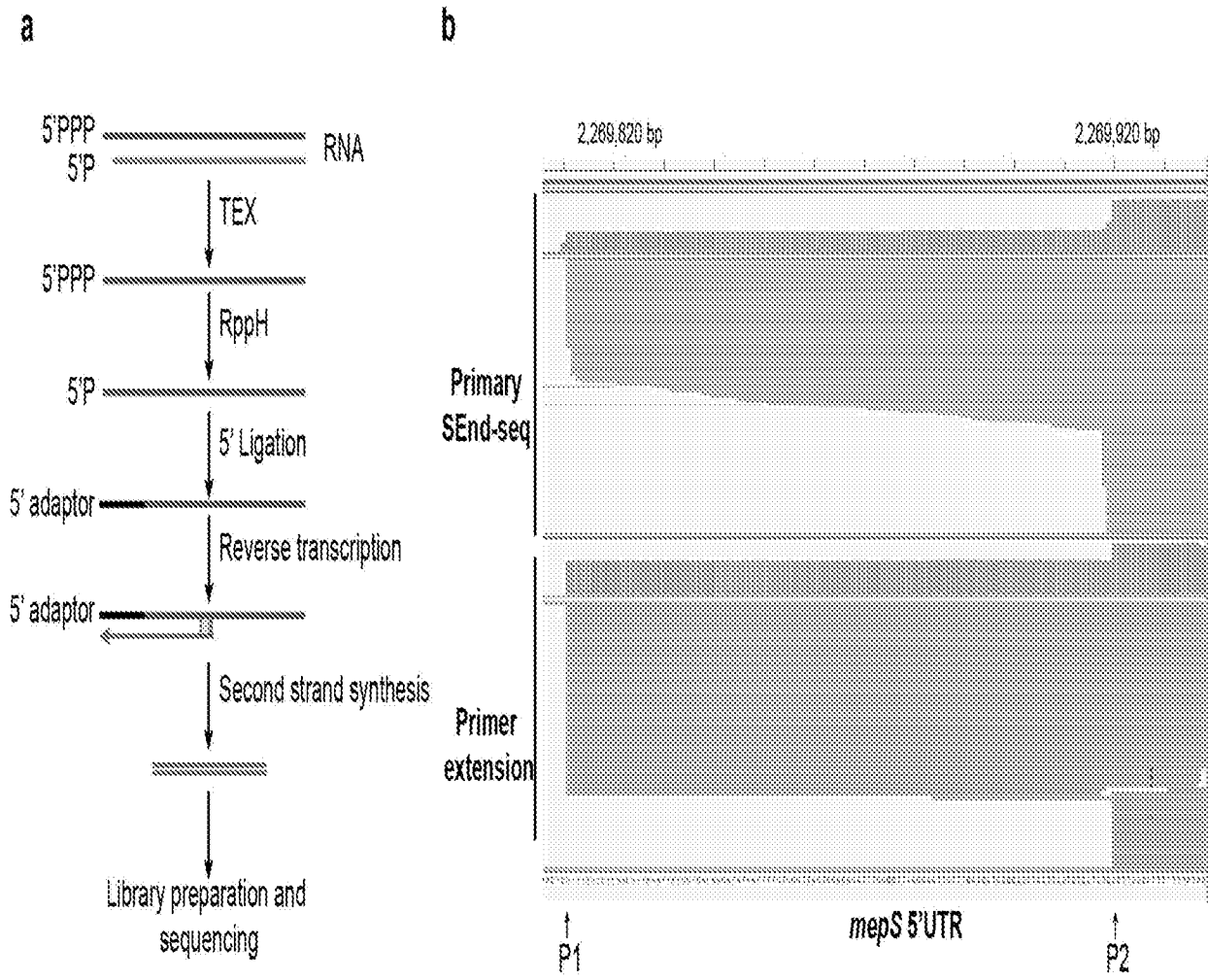


FIGURE 13-2

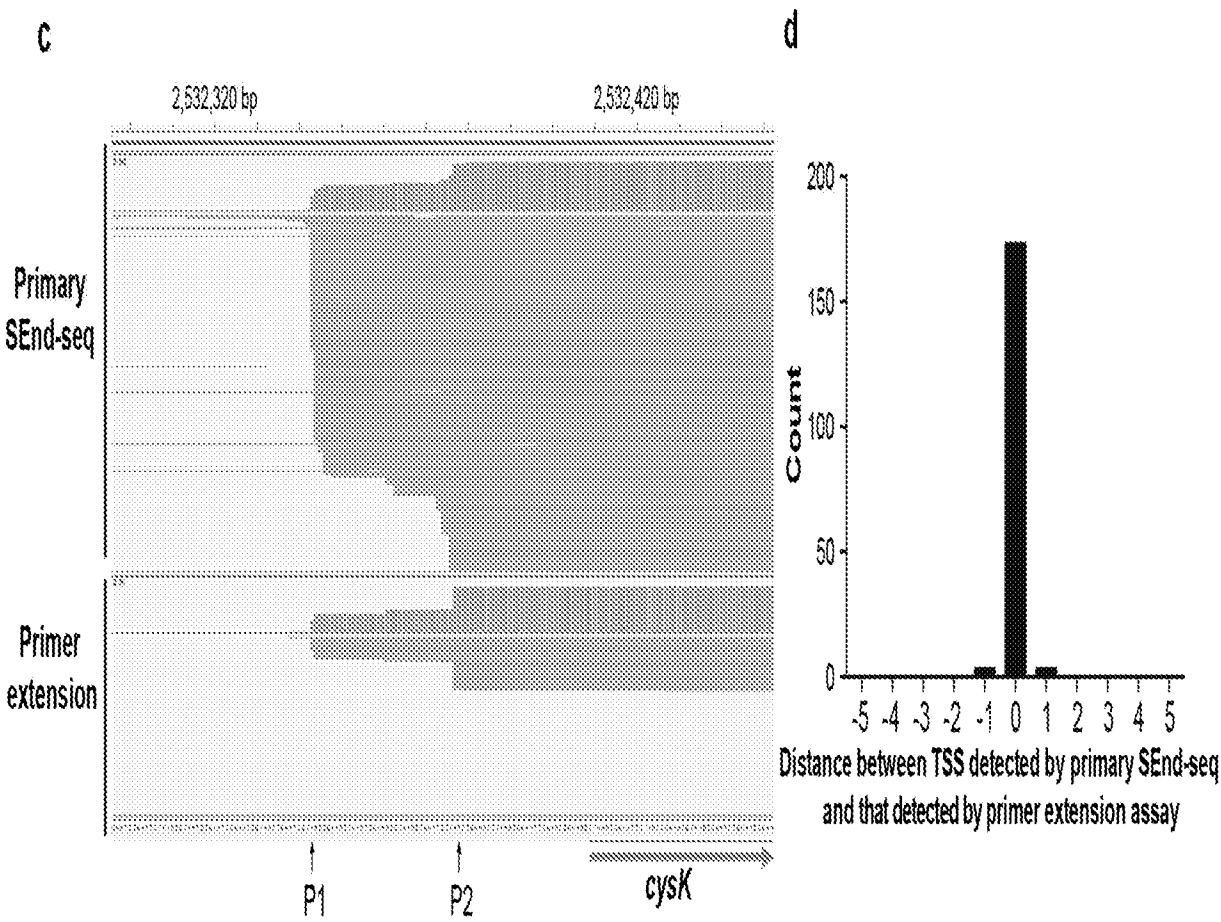


FIGURE 14

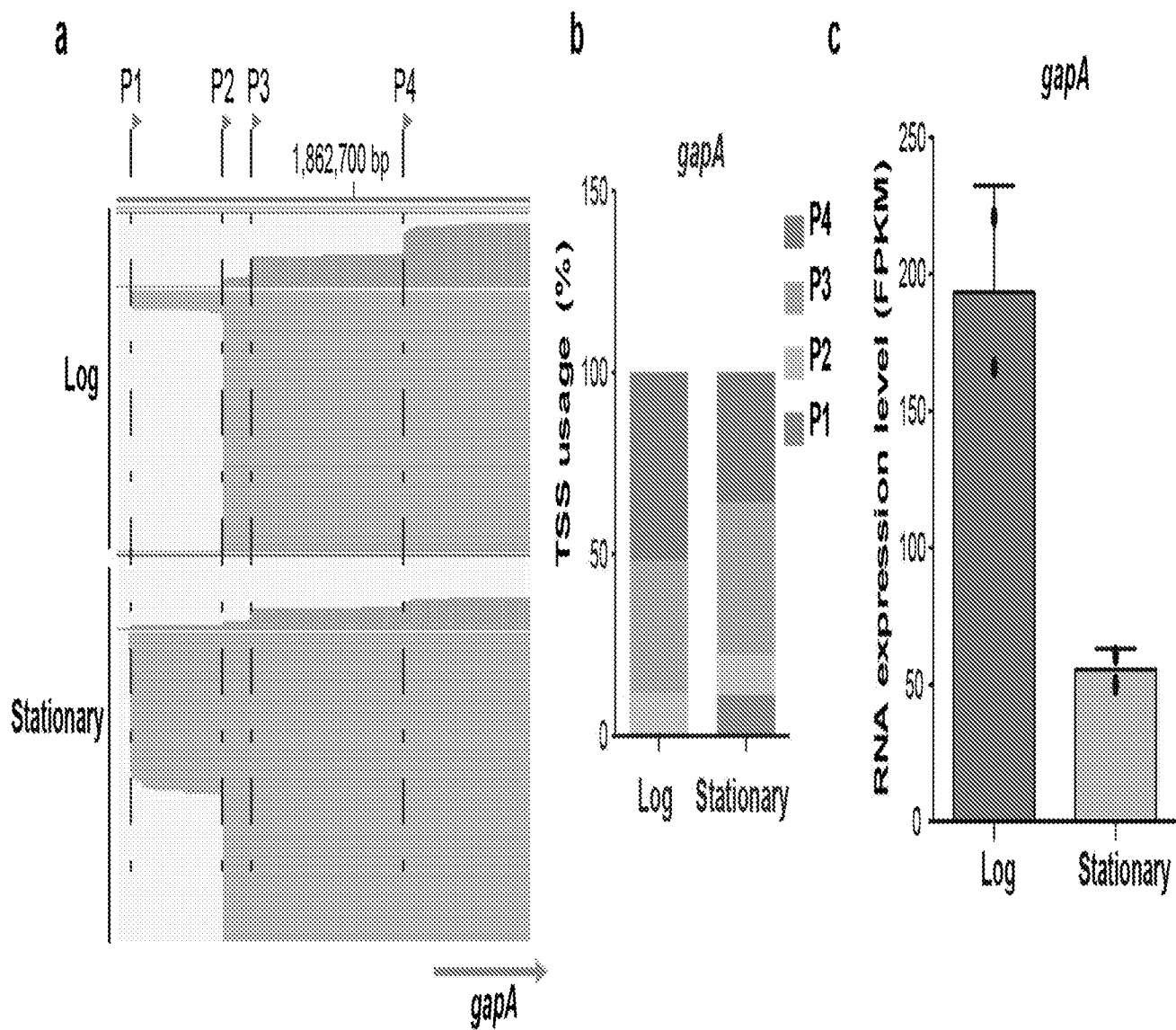


FIGURE 15

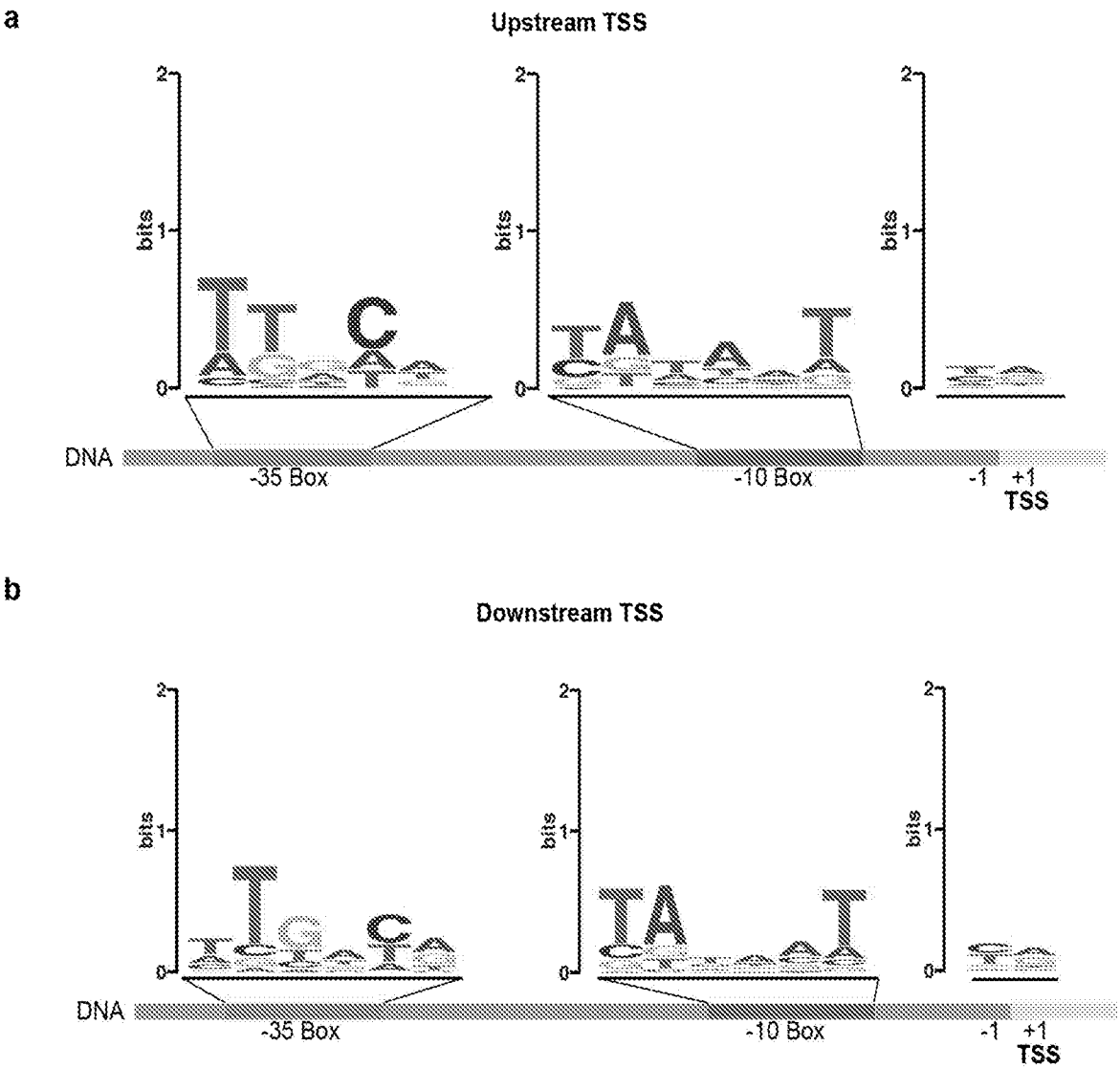


FIGURE 16

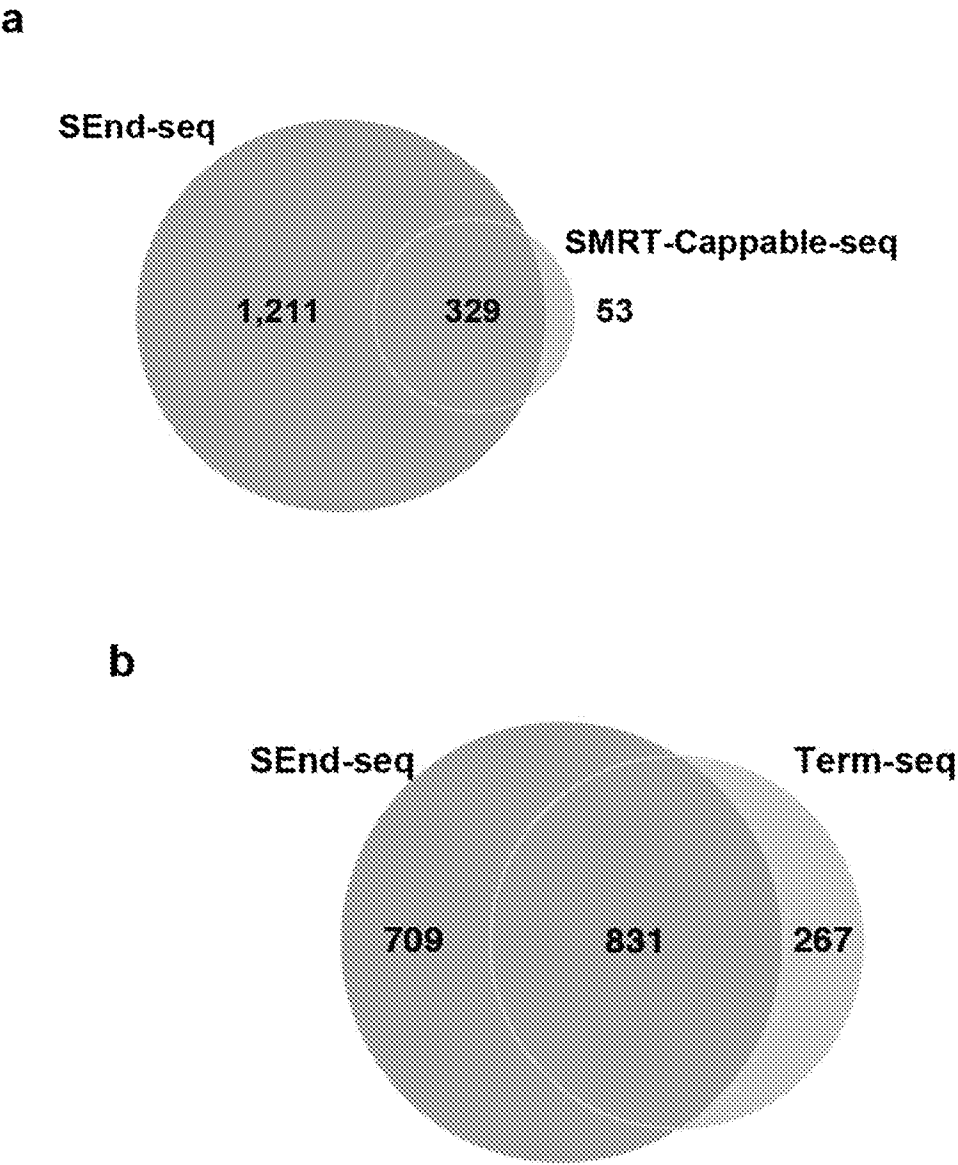


FIGURE 17

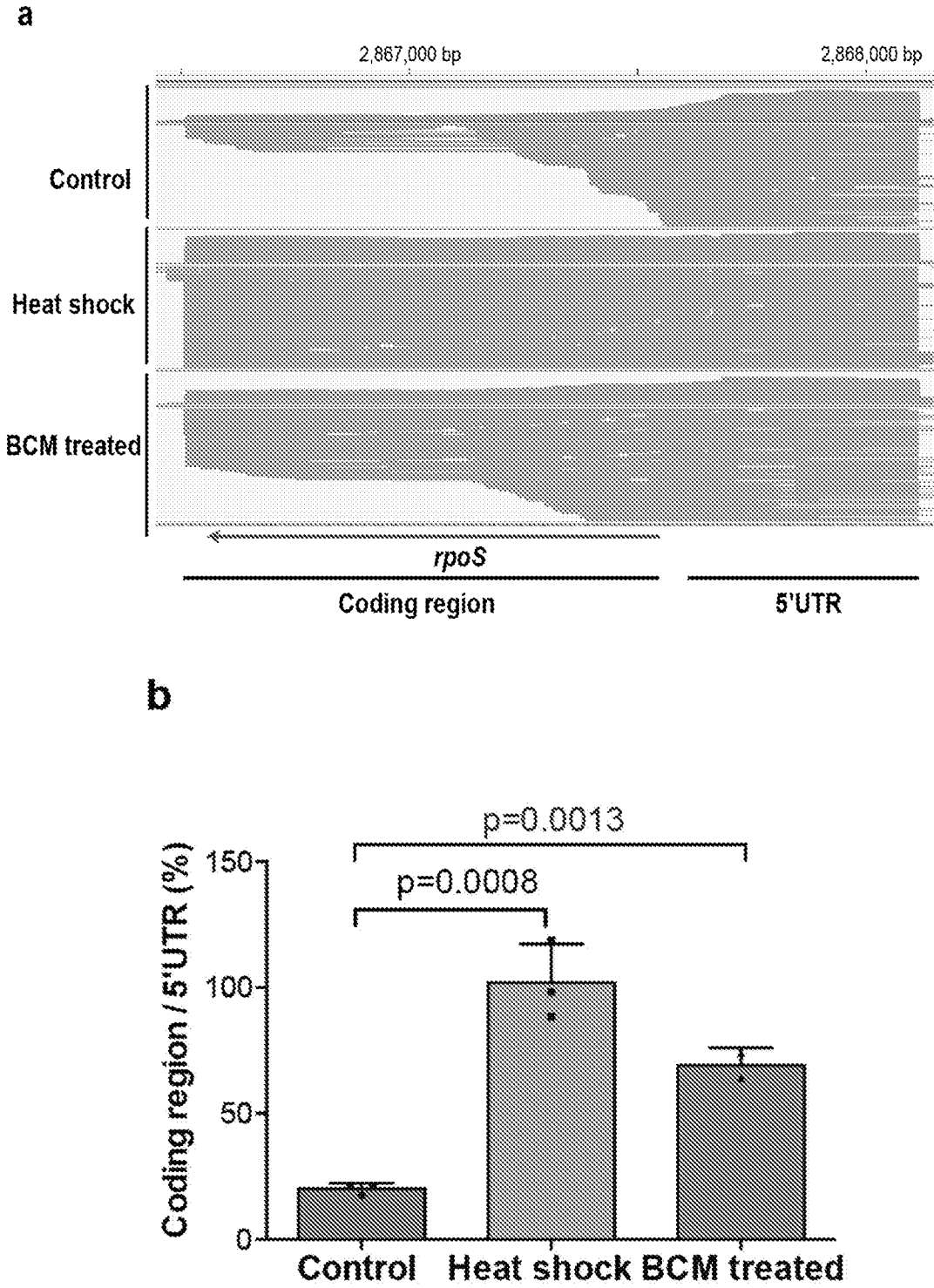


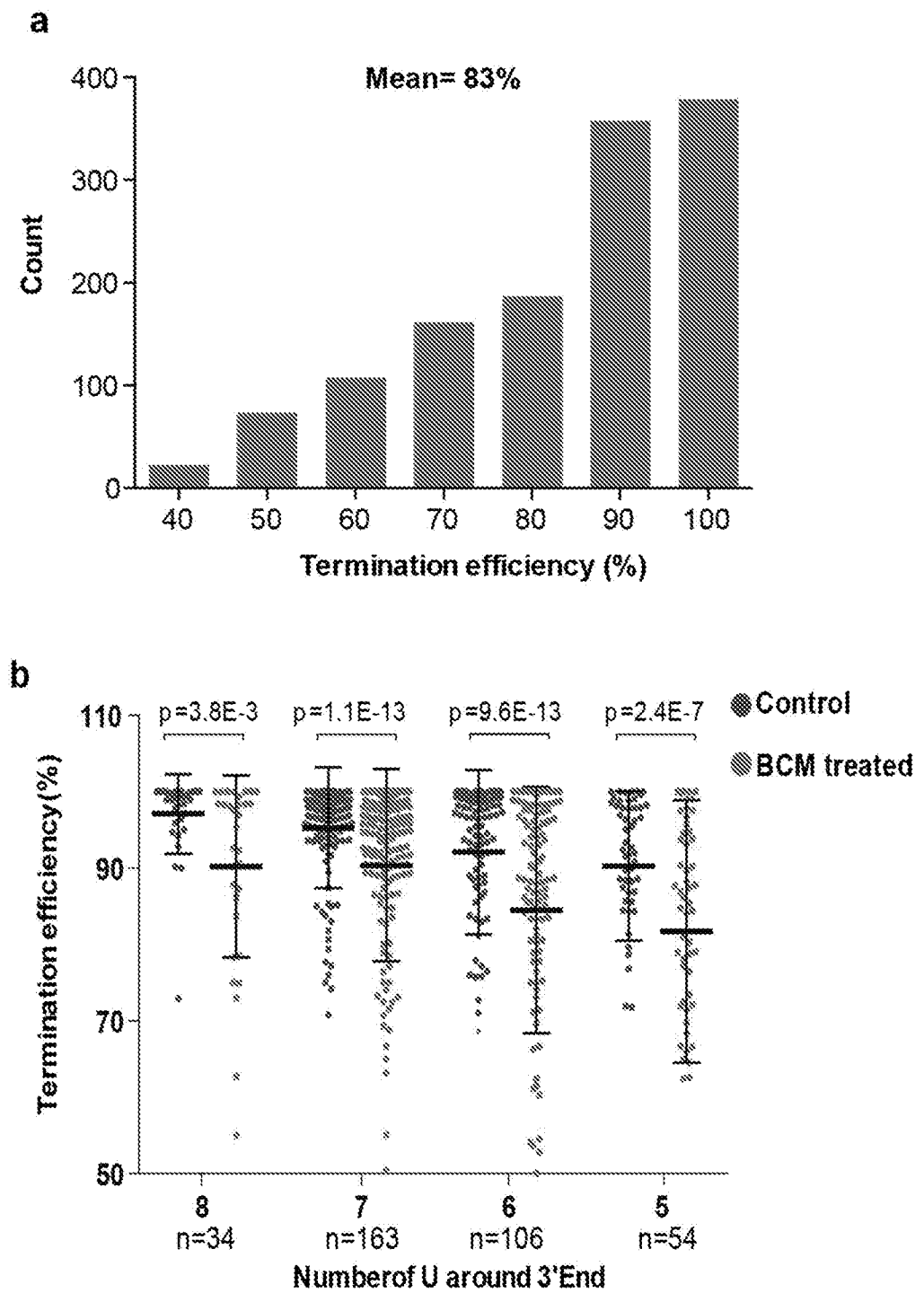
FIGURE 18

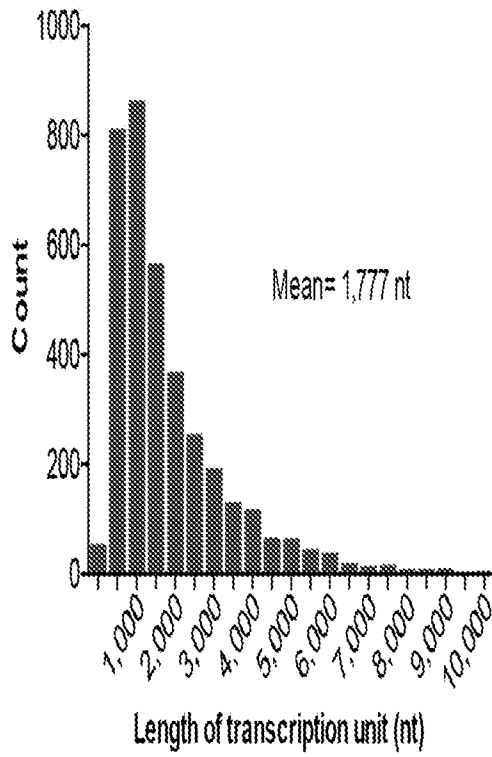
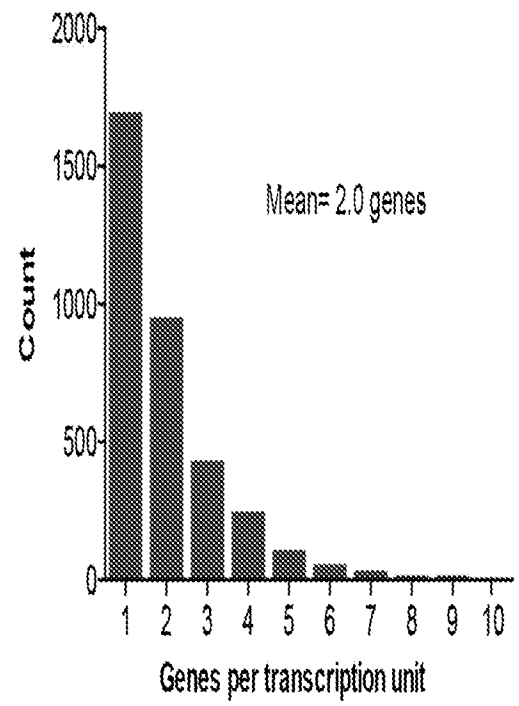
FIGURE 19-1**a****b**

FIGURE 19-2

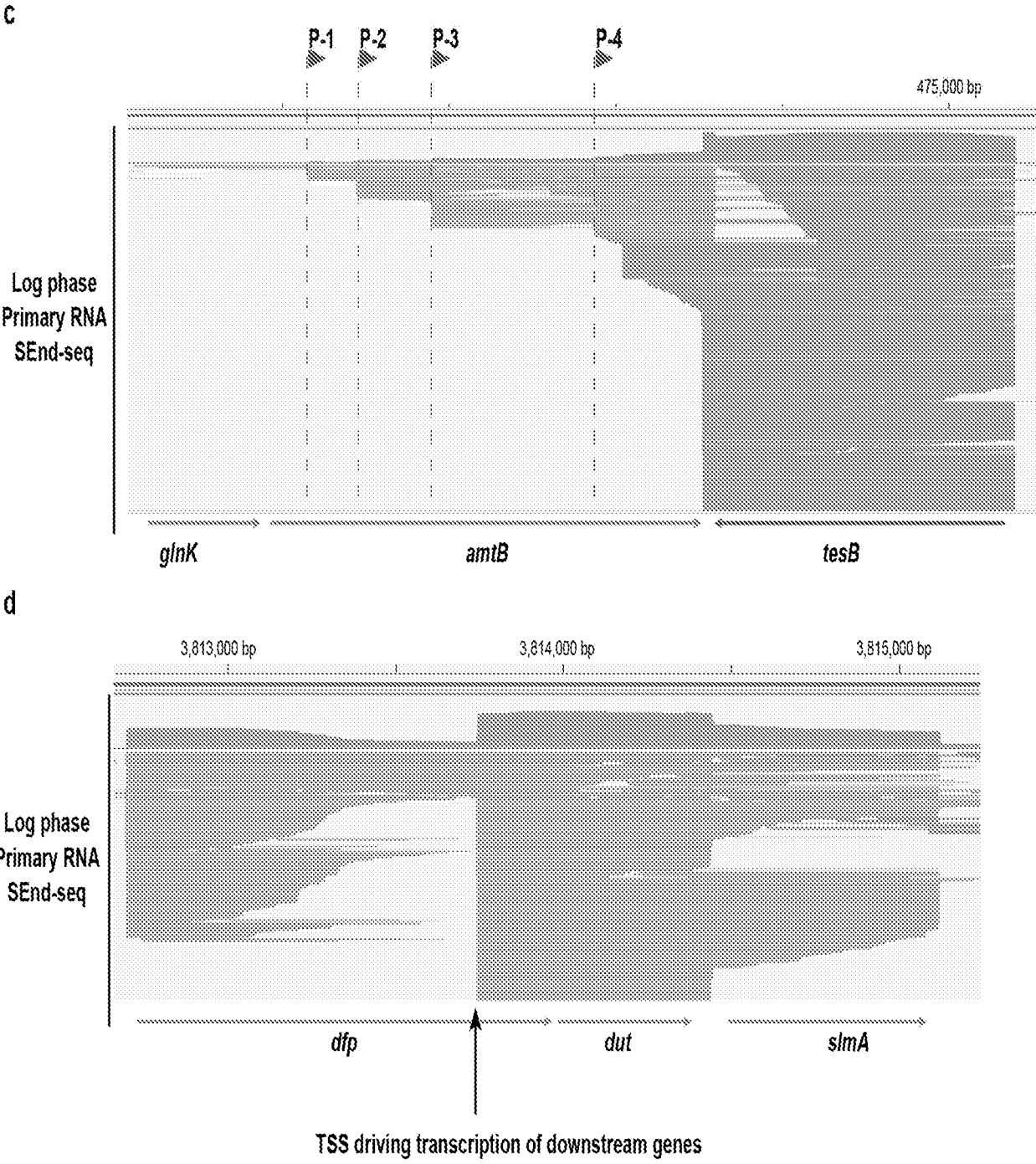


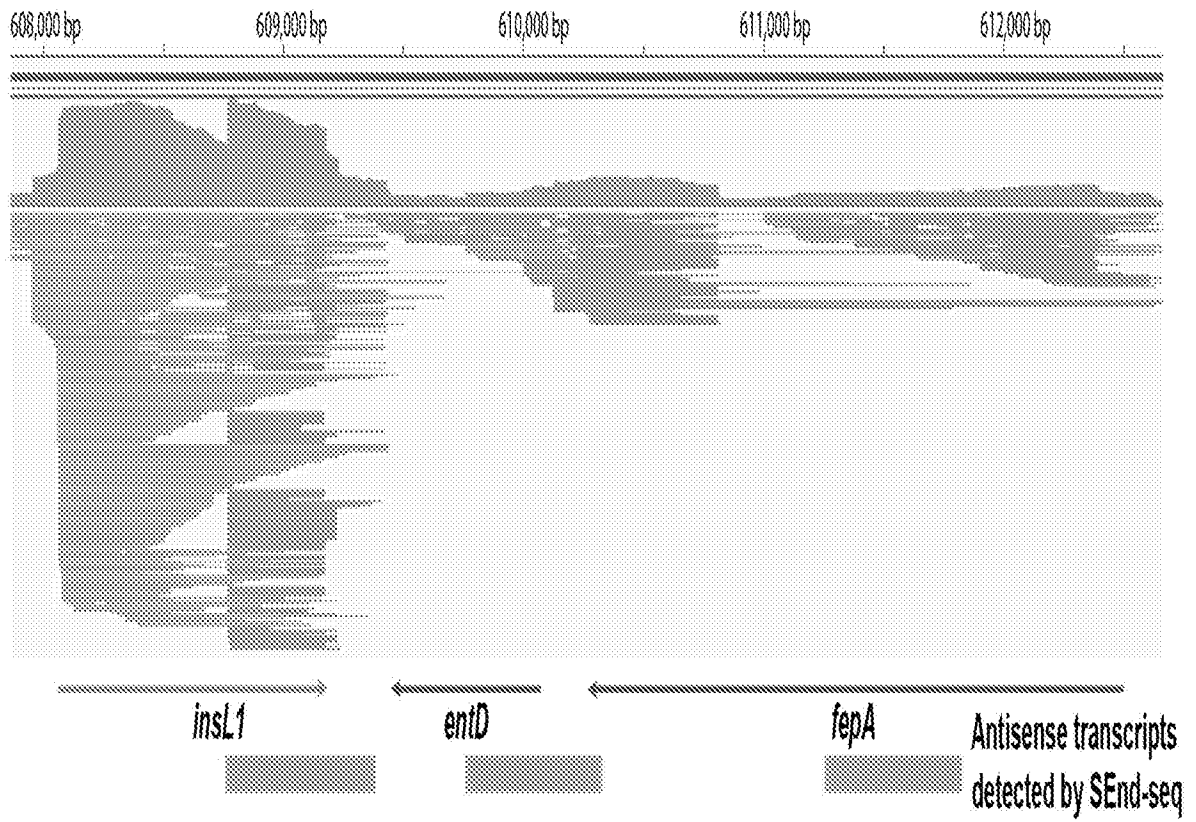
FIGURE 20-1**a**

FIGURE 20-2

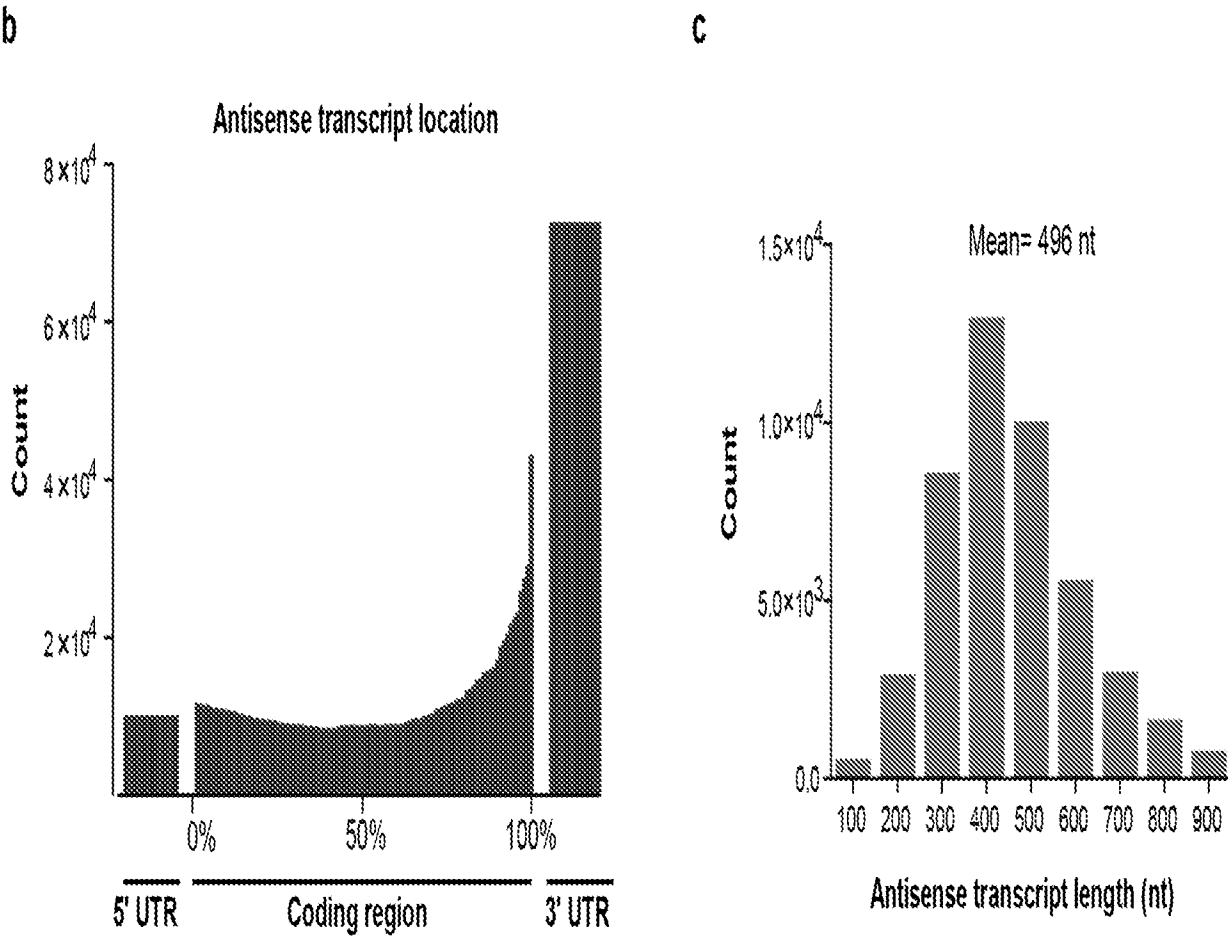


FIGURE 21

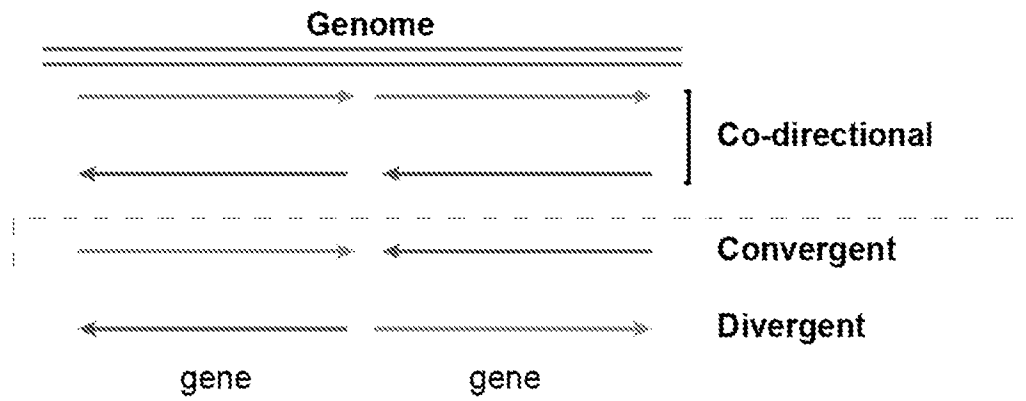
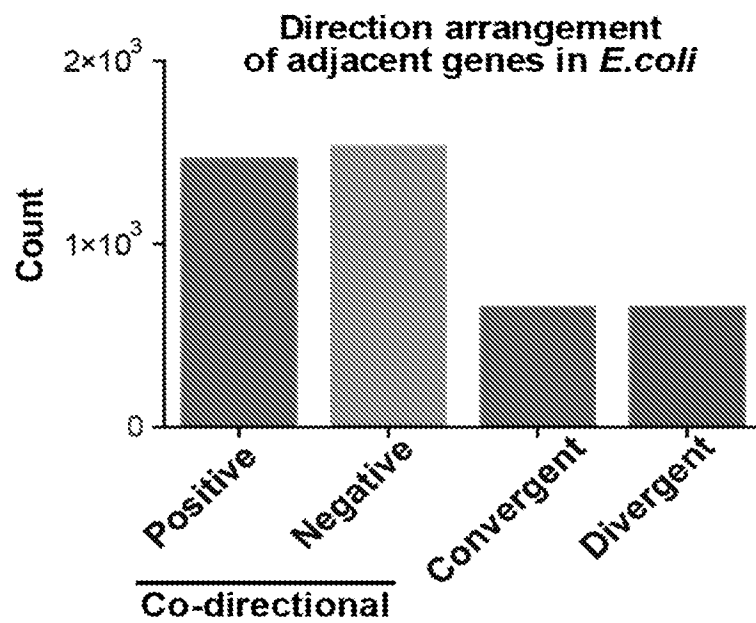
a**b**

FIGURE 22

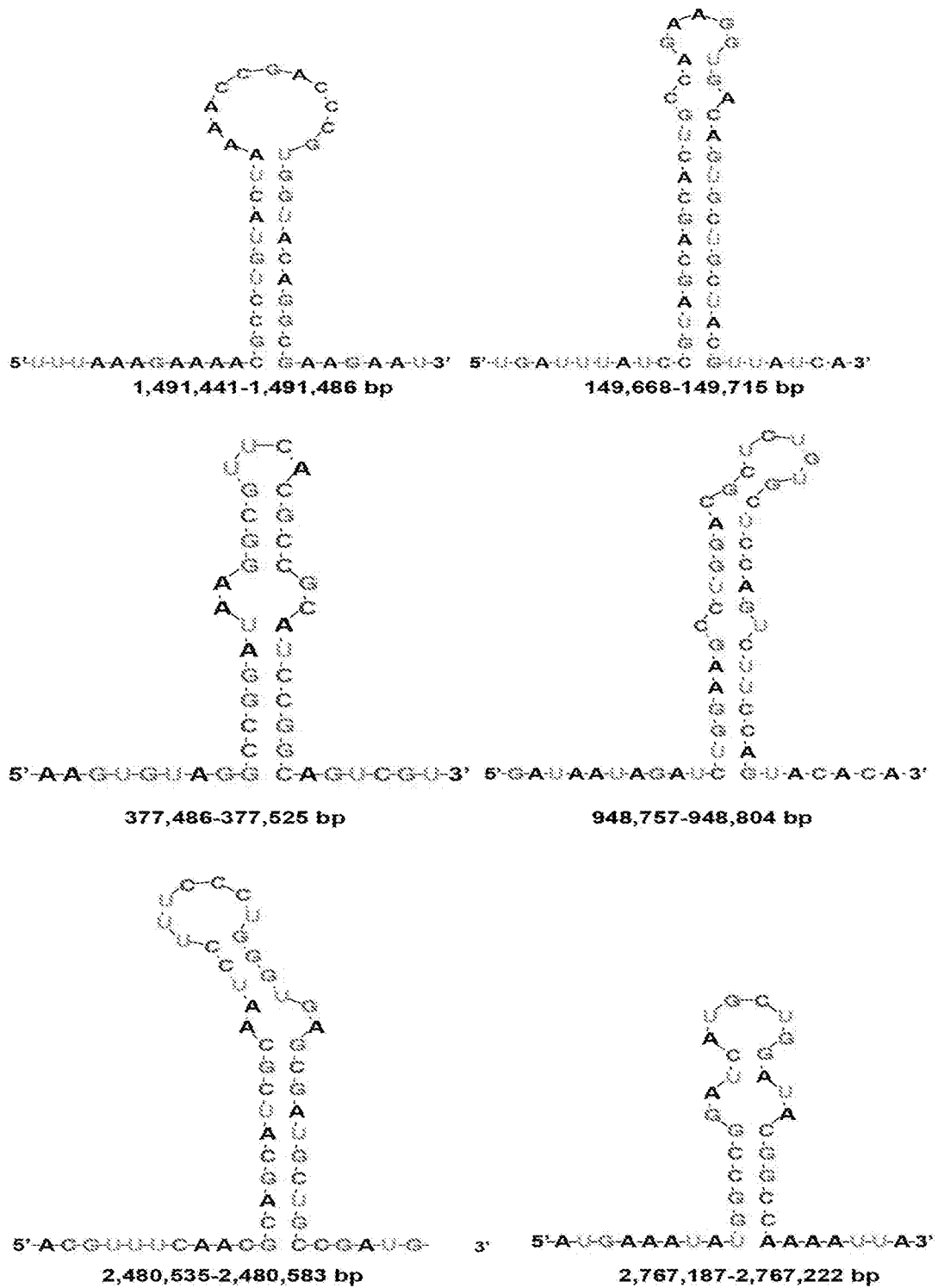
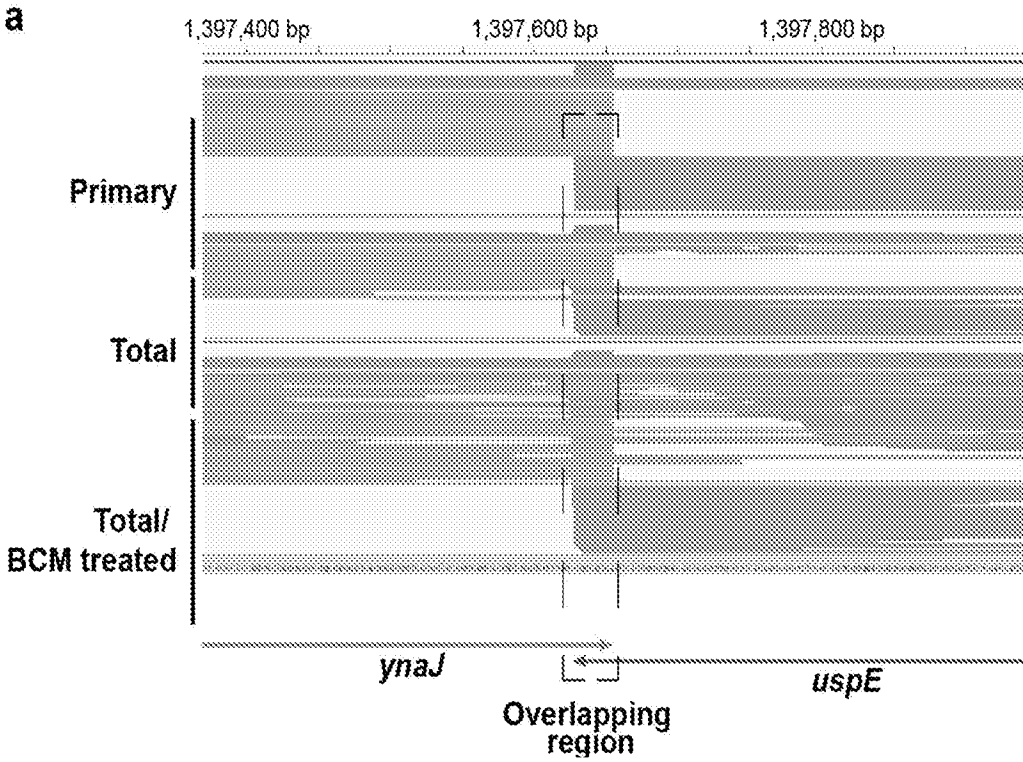


FIGURE 23



b

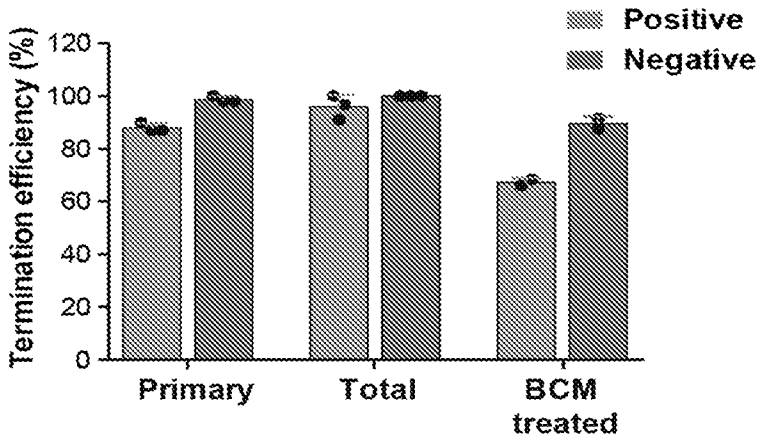


FIGURE 24

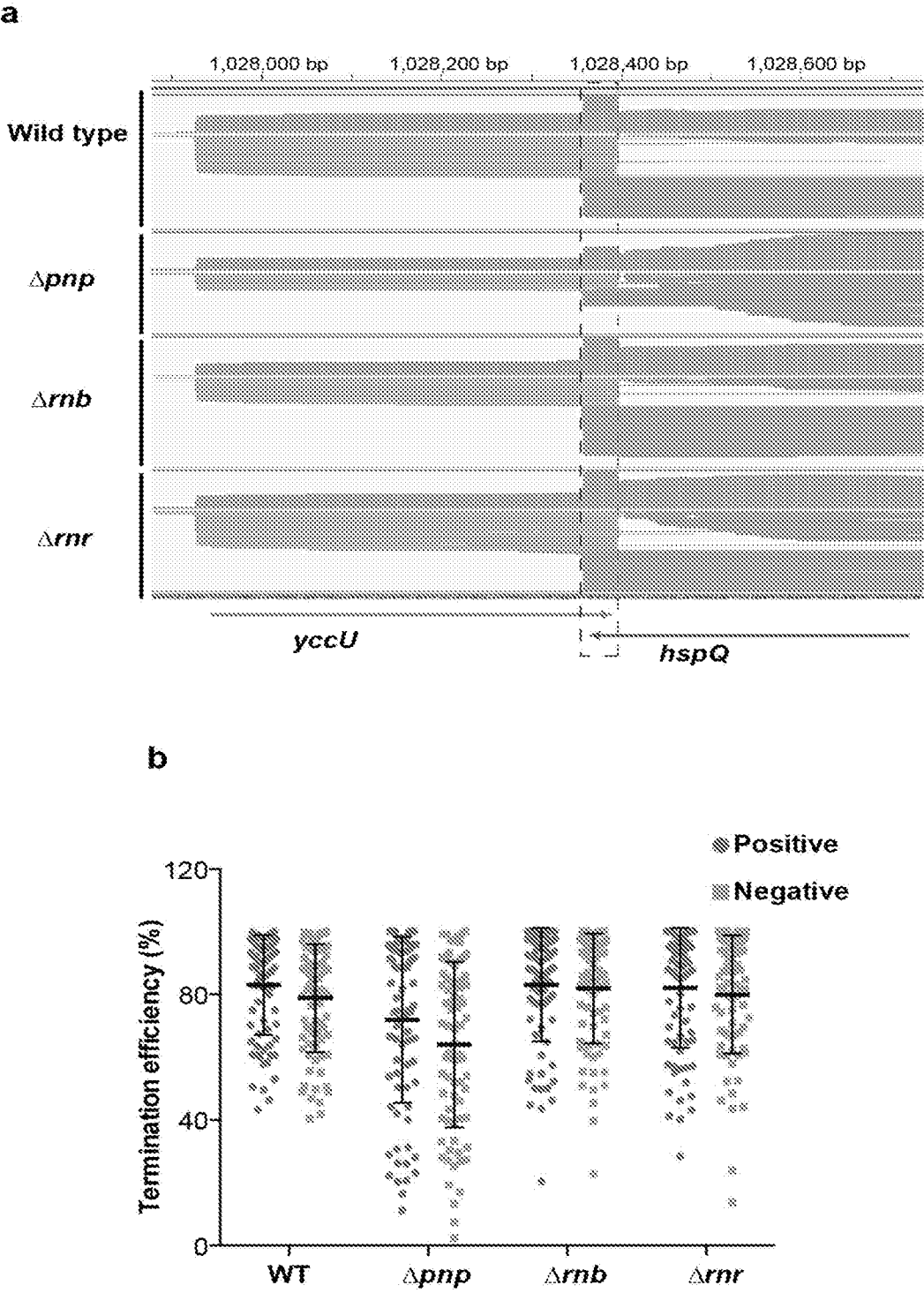


FIGURE 25

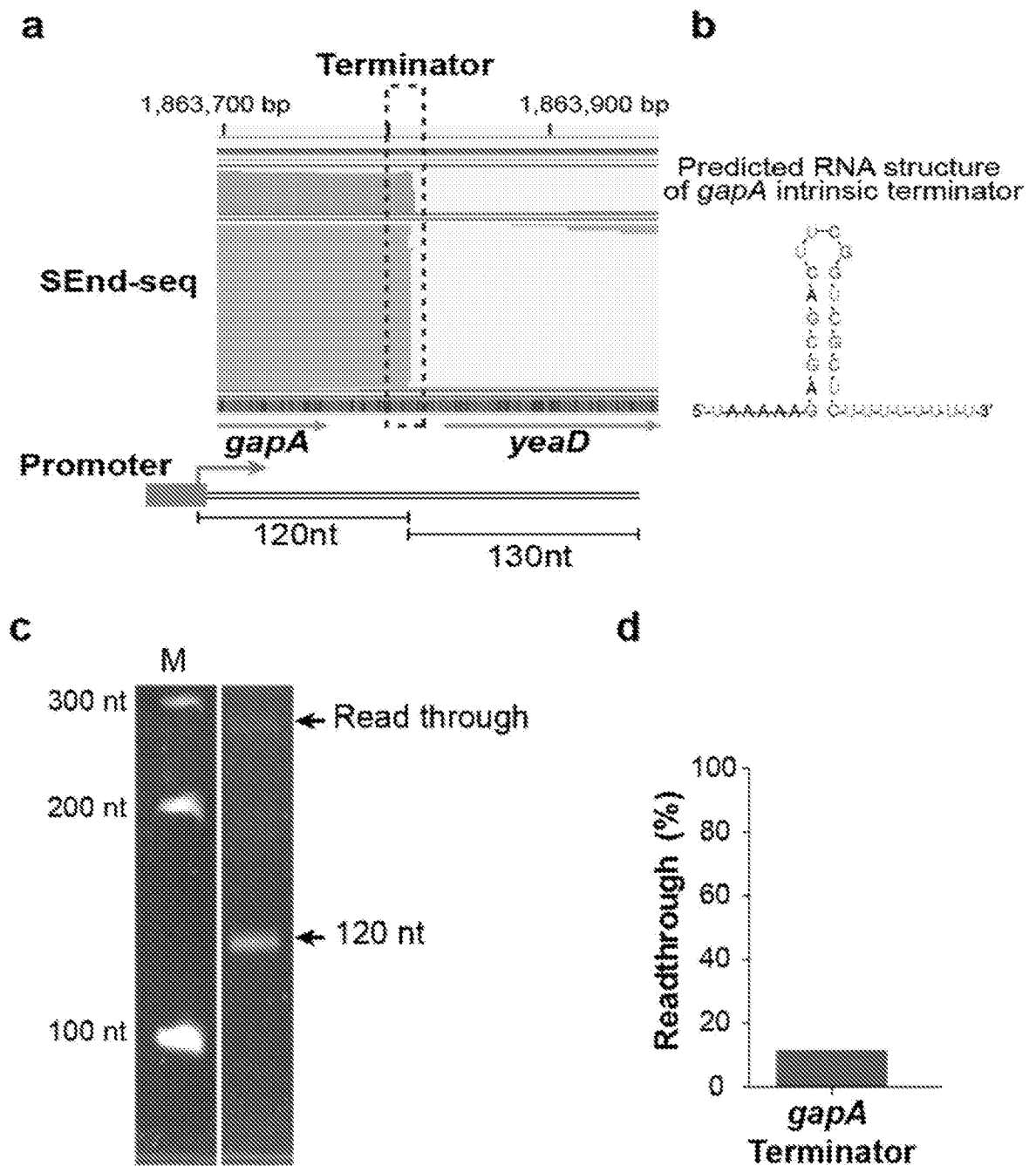


FIGURE 26-1

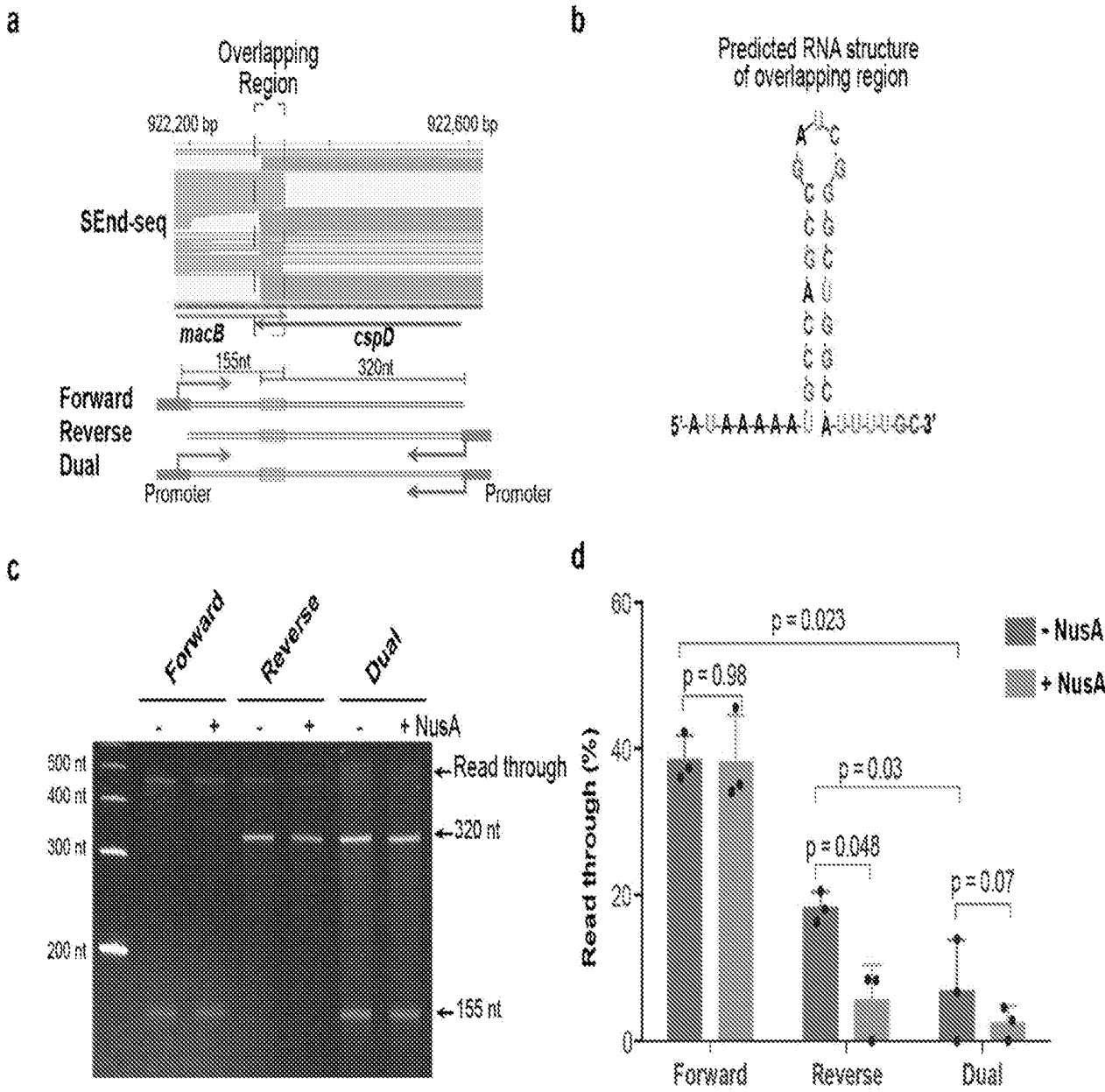


FIGURE 26-2

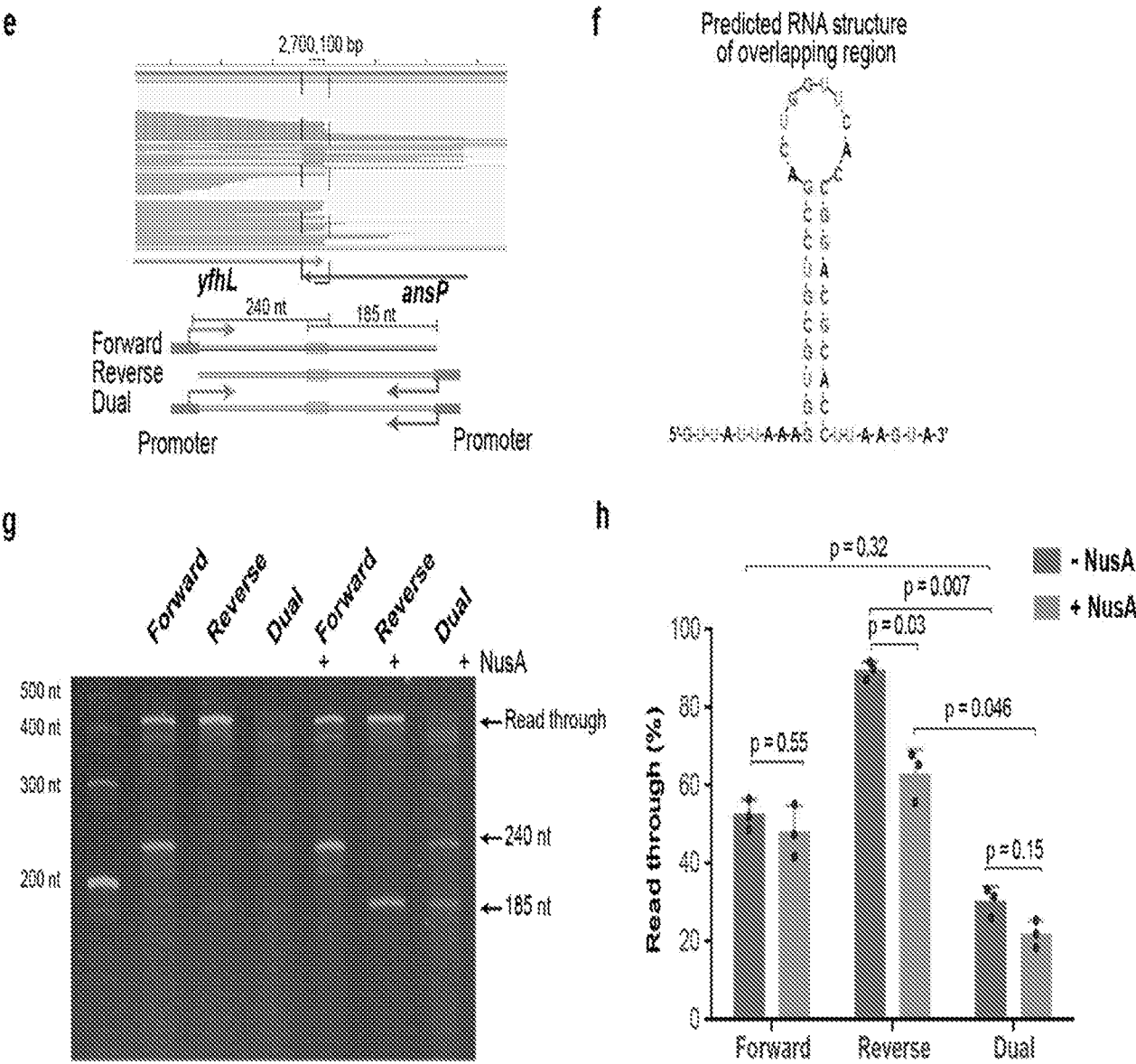


FIGURE 27-1

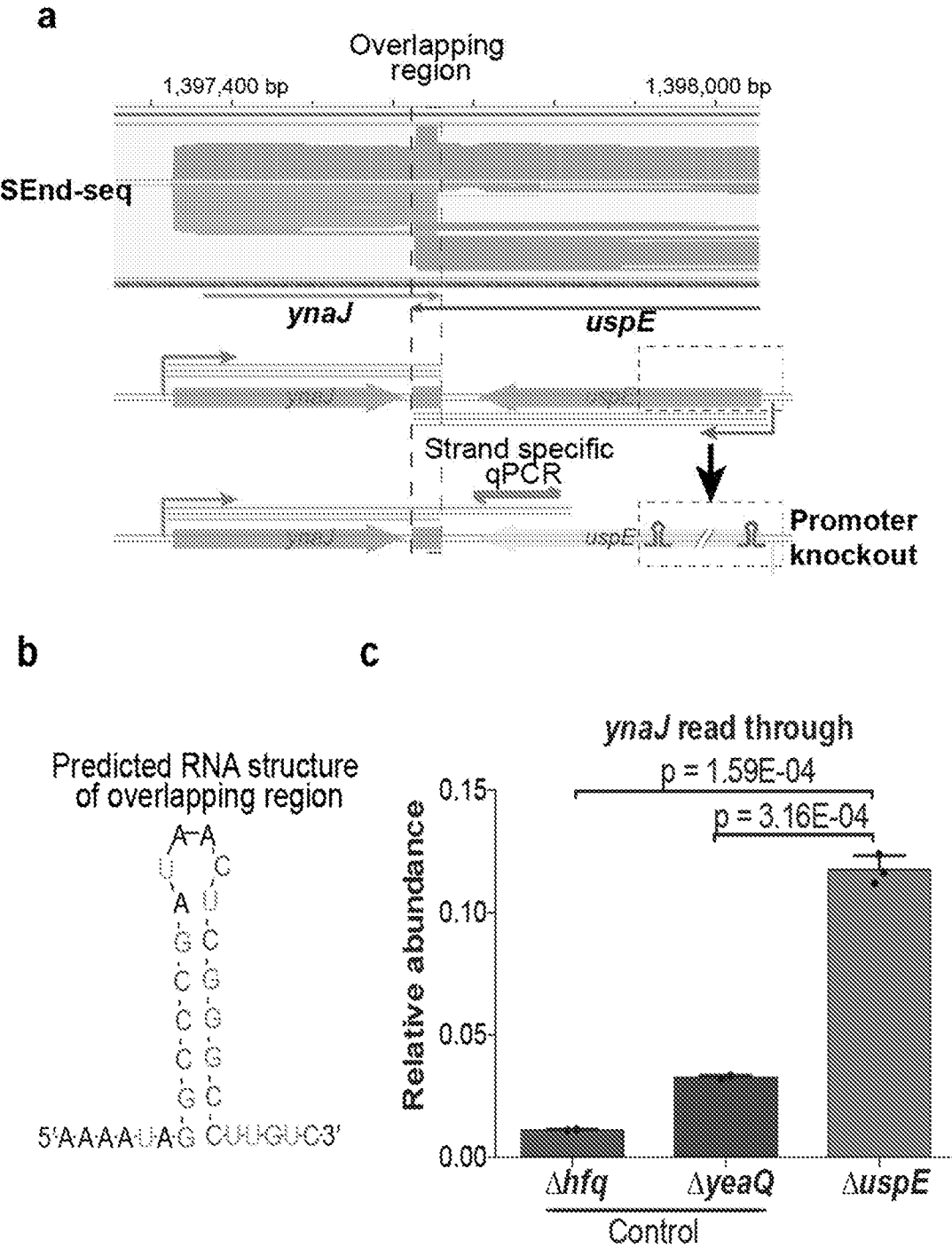
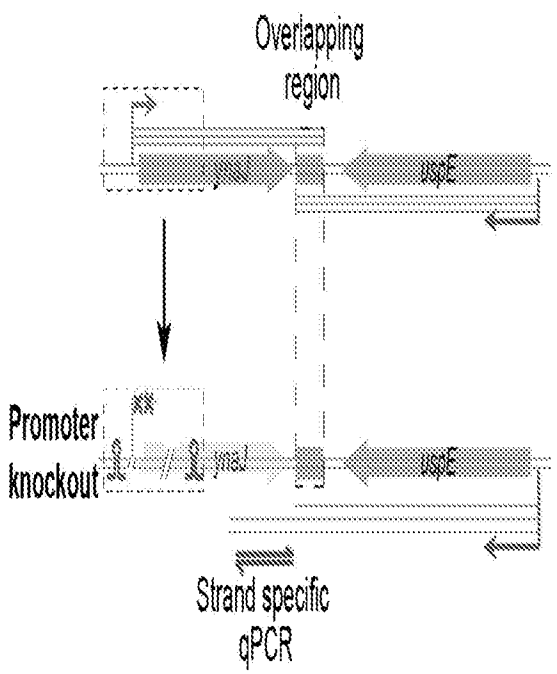


FIGURE 27-2

d



e

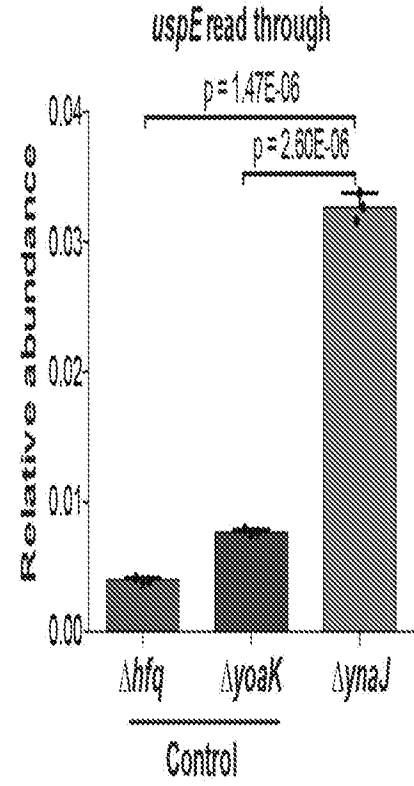


FIGURE 28-1

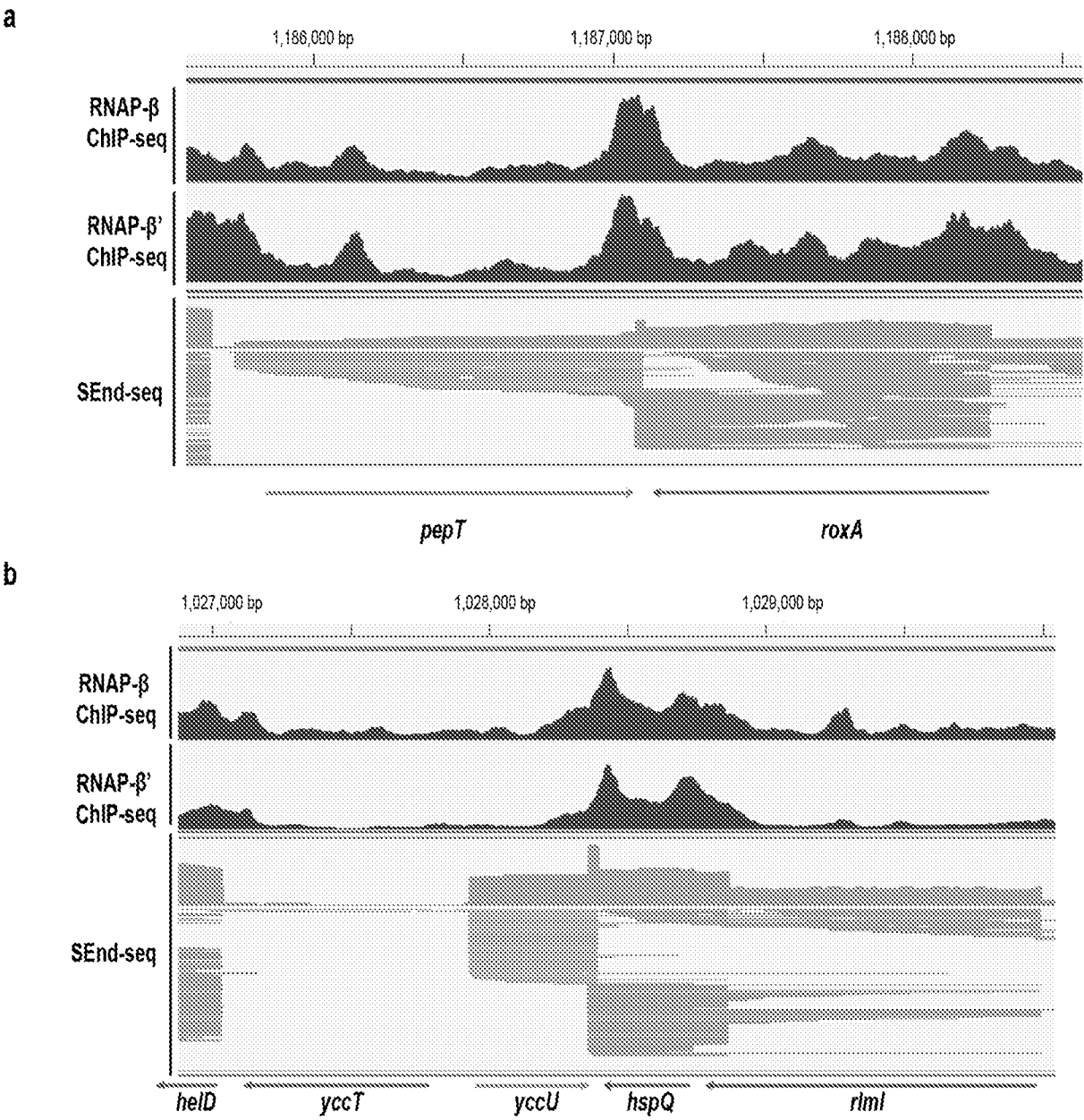


FIGURE 28-2

C

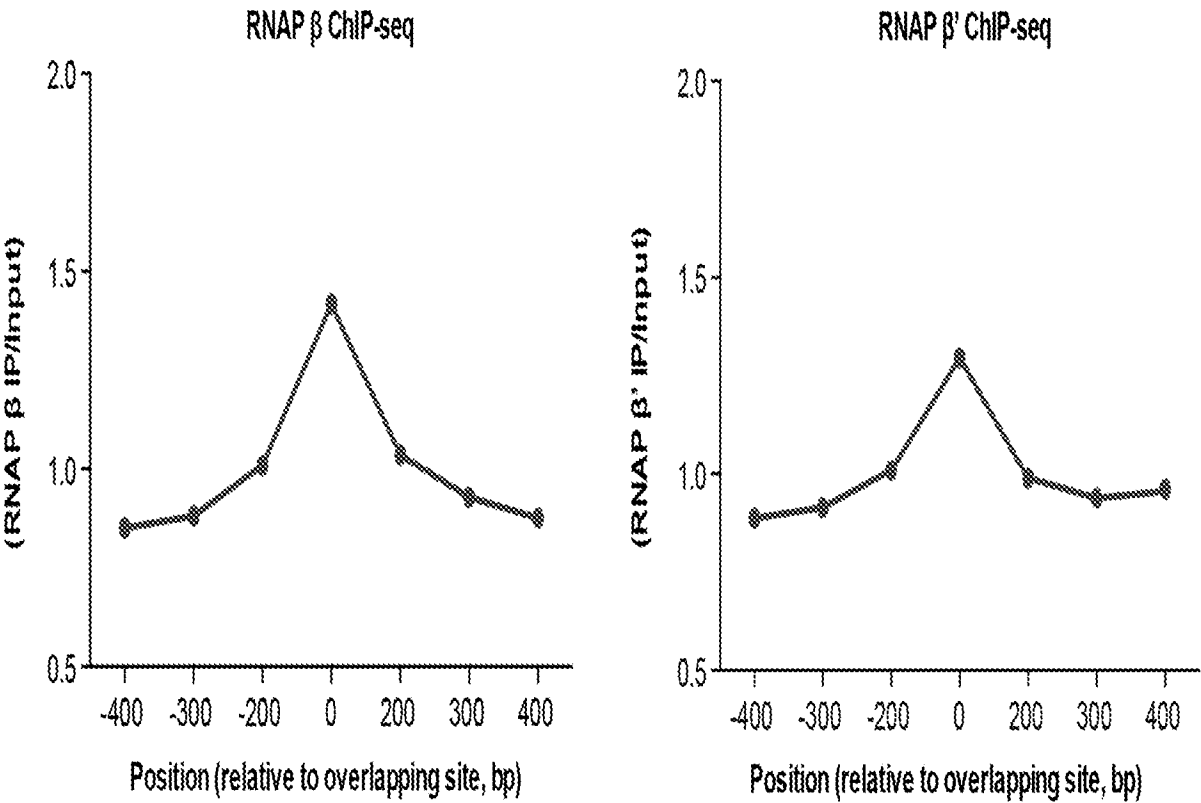


FIGURE 29-1

Figure 5c:

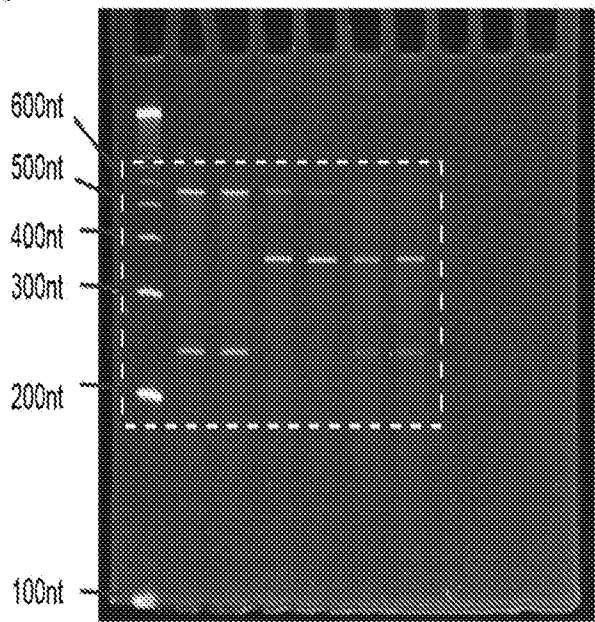
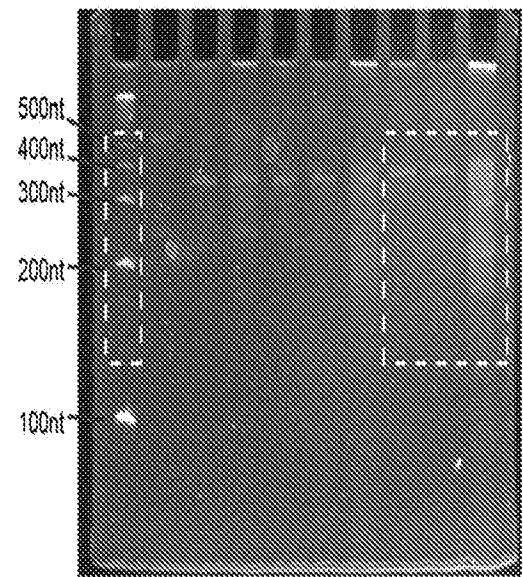
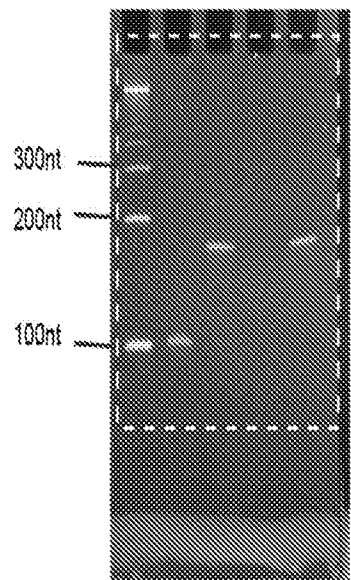


Figure 5g:



Supplementary Figure 1a:



Supplementary Figure 1b:

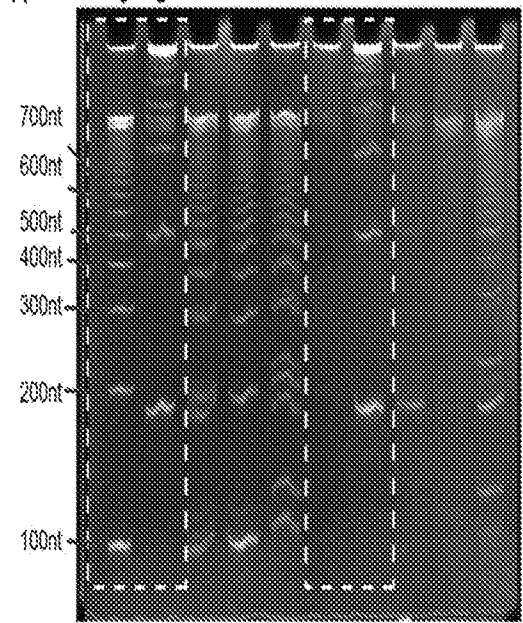
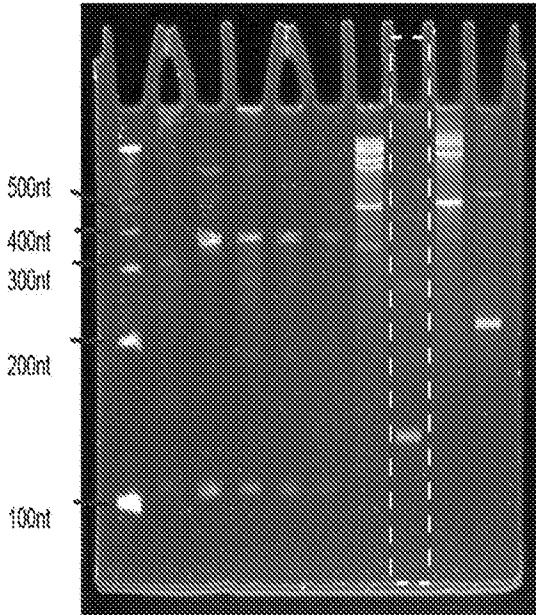
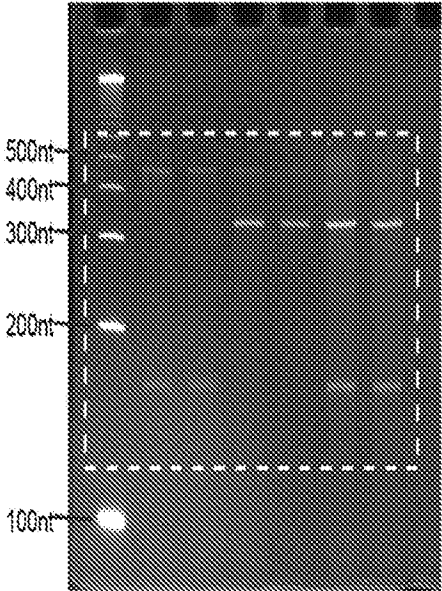


FIGURE 29-2

Supplementary Figure 19c:



Supplementary Figure 20c:



Supplementary Figure 20g:

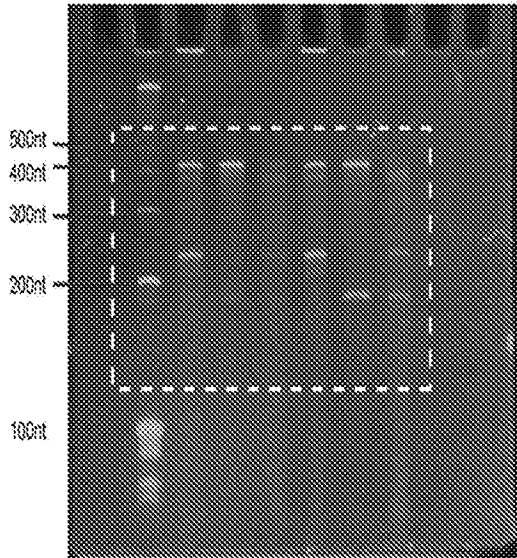


FIGURE 30

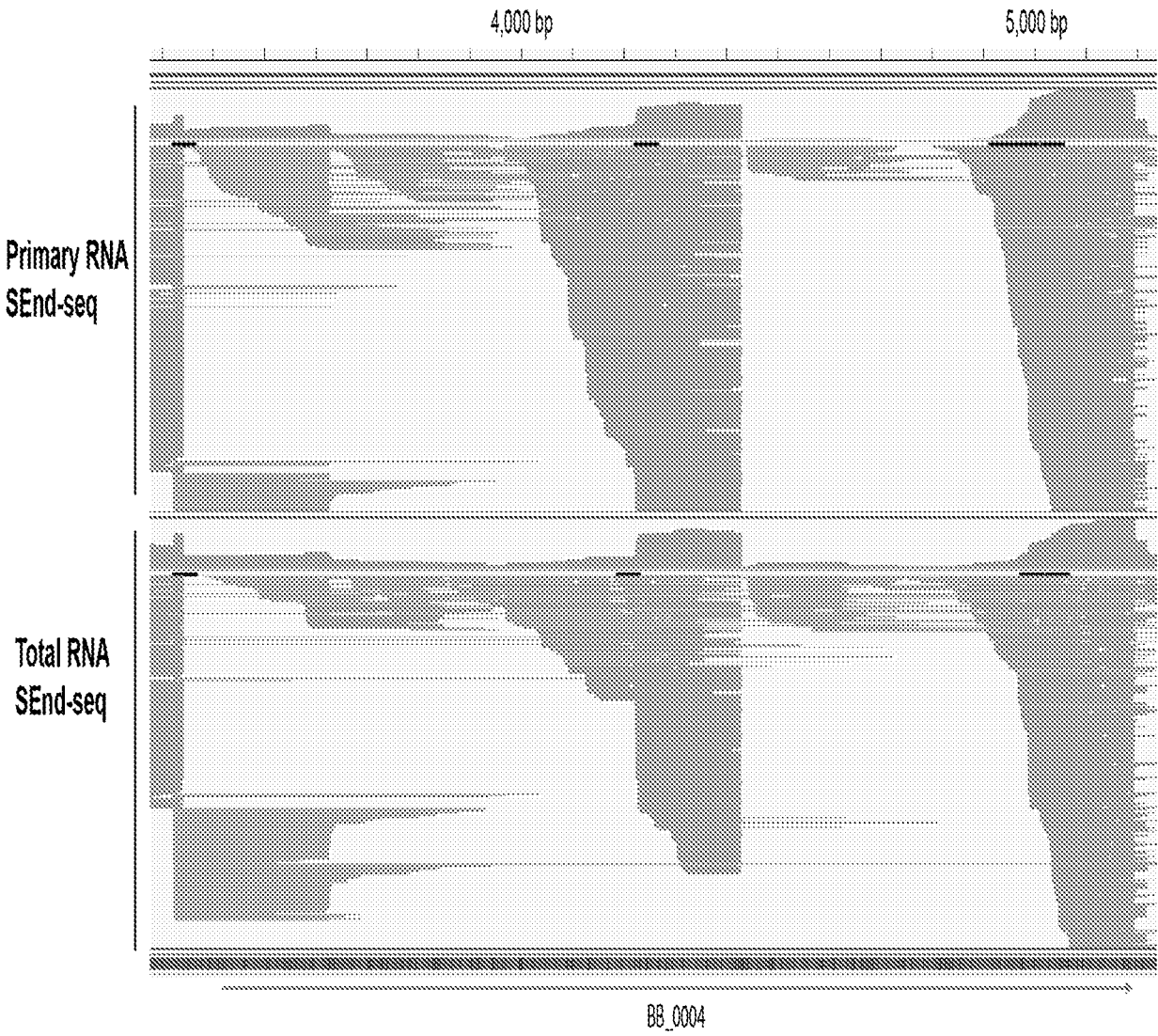


FIGURE 31

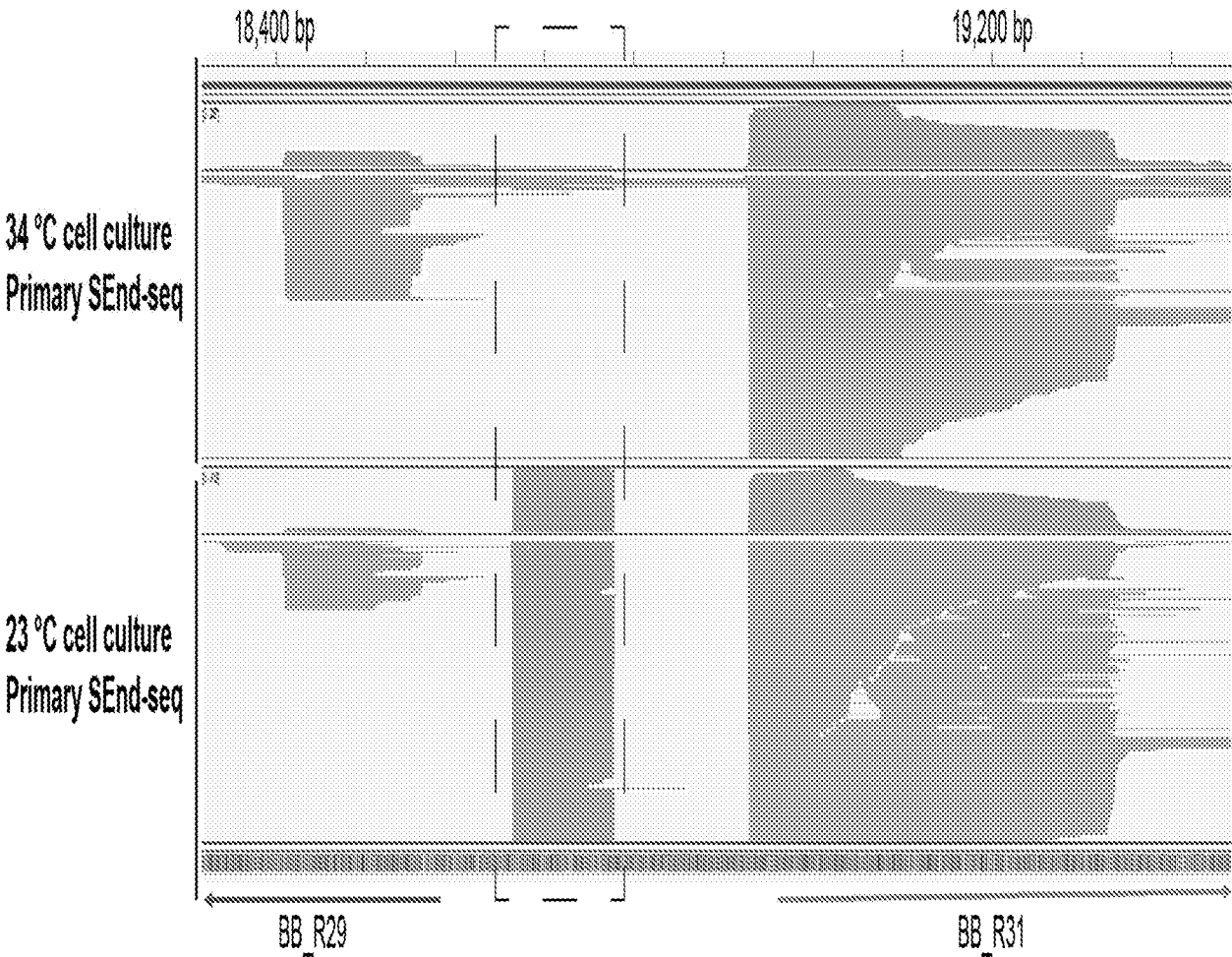


FIGURE 32

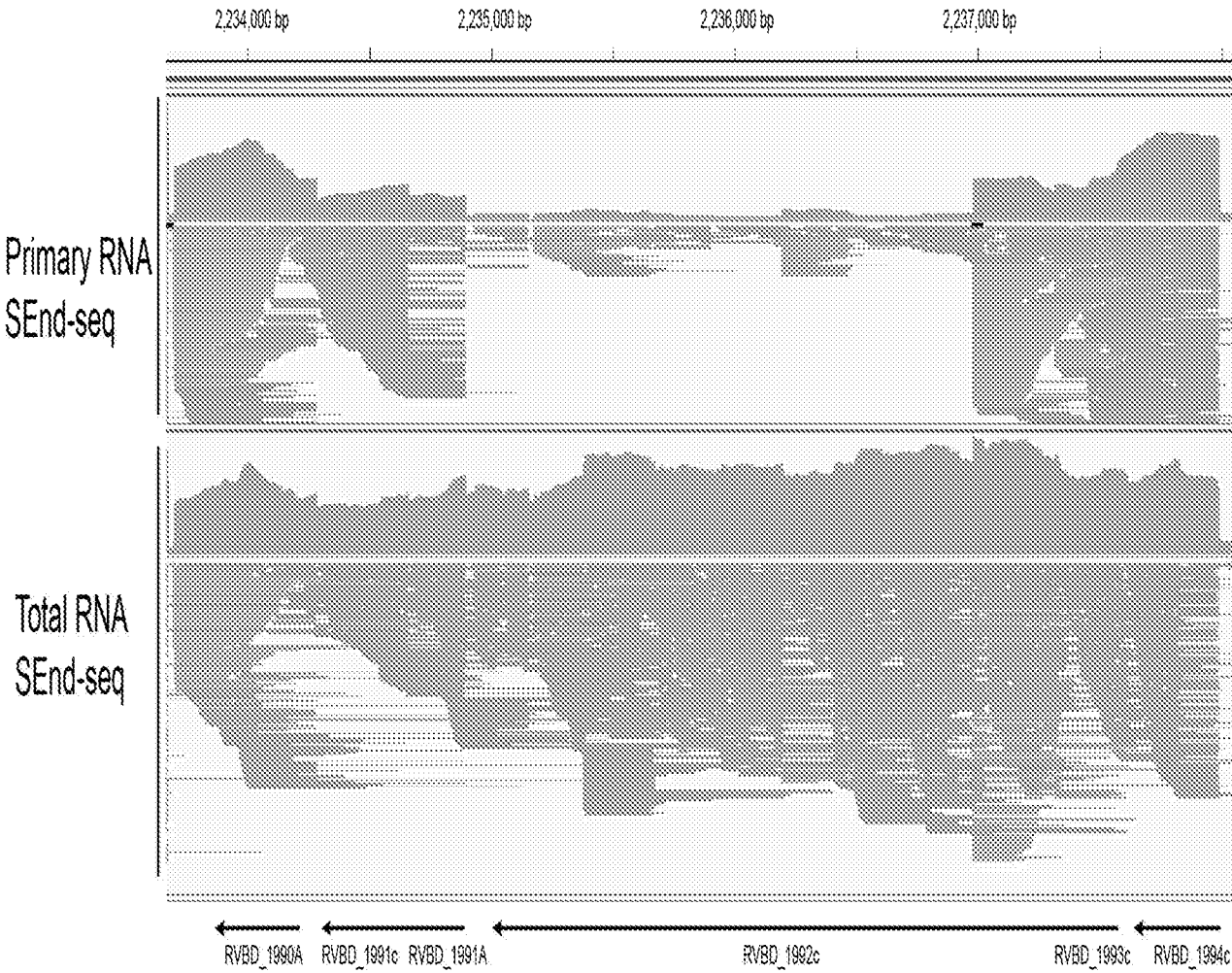
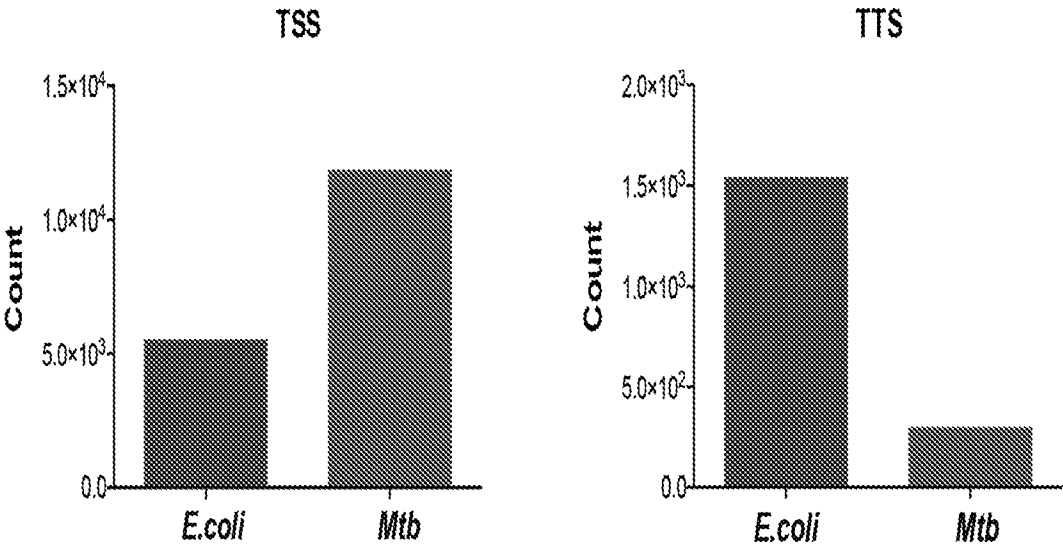


FIGURE 33



INTERNATIONAL SEARCH REPORT

International application No.

PCT/US 20/34127

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:
2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:
3. ☒ Claims Nos.: 7-30, 34-48, 53
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:
This application contains the following inventions or groups of inventions which are not so linked as to form a single general inventive concept under PCT Rule 13.1. In order for all inventions to be searched, the appropriate additional search fees must be paid.

Group I, claims 1-6, directed to a method of full-length transcript profiling of RNA.

Group II, claims 31-33, directed to a method of making a DNA library.

Group III, claims 49-52, directed to a kit for simultaneous 5'- and 3'-end RNA sequence capture.

The inventions listed as Groups I-III do not relate to a single general inventive concept under PCT Rule 13.1 because, under PCT Rule 13.2, they lack the same or corresponding special technical features for the following reasons:

*****Continued in the extra sheet*****

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.
2. ☐ As all searchable claims could be searched without effort justifying additional fees, this Authority did not invite payment of additional fees.
3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:
4. ☒ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:
1-6

Remark on Protest

- ☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.
- ☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.
- ☐ No protest accompanied the payment of additional search fees.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US 20/34127

A. CLASSIFICATION OF SUBJECT MATTER

IPC - C12Q 1/68; C12P 19/34; C40B 40/06; C07H 21/04 (2020.01)

CPC - C12Q 1/68; C12Q 1/6869; C12Q 1/6876; C12N 15/1096; C12Q 1/686; C12Q 1/6844; C12Q 2525/191; C12Q 1/6809; C12N 15/10; C07H 21/00; C12Q 2533/101

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

See Search History document

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

See Search History document

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

See Search History document

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 2015/0099671 A1 (UNIVERSITY OF MASSACHUSETTS) 09 April 2015 (09.04.2015), para [0003], [0009], [0012], [0015], [0016], [0017], [0019], [0032], [0035], [0039], [0046], [0047], [0058], [0064], [0076], and [0078]	1-6
Y	US 2014/0213485 A1 (YALE UNIVERSITY) 31 July 2014 (31.07.2014), Abstract, para [0010], [0013], [0021], [0027], [0082], [0085], [0109], and [0121]	1-6
Y	US 2018/0195061 A1 (NEW ENGLAND BIOLABS, INC.) 12 July 2018 (12.07.2018), Abstract, para [0031], [0045], [0061], [0073], [0079], [0095], [0126], [0127], [0189], [0243], [0245], [0254], and [0324]	2-4, 5/(2-4), 6/[5/(2-4)]
Y	PRADOS et al., TSS-EMOTE, a refined protocol for a more complete and less biased global mapping of transcription start sites in bacterial pathogens. BMC Genomics. 2016, Vol. 17: 849. PDF File: pg 1-16. Abstract; pg 2, col 2, middle para; pg 3, col 2, middle para; and pg 4, Fig 1	6
A	KLERK et al., RNA sequencing: from tag-based profiling to resolving complete transcript structure. Cell Mol Life Sci. 2014, Vol. 71(18), p. 3537-3551. Entire Documentation especially Abstract; pg 3539, Fig 1; pg 3540, Fig 2; and pg 3544, col 2, para 4	1-6
A	LAMA et al., Small RNA-seq: The RNA 5'-end adapter ligation problem and how to circumvent it. J Biol Methods. 2019, Vol. 6(1): e108. Published online 2019 Feb 20. PDF File: pg 1-12. Entire Documentation especially Abstract; and pg 2, Fig 1	1-6

☐ Further documents are listed in the continuation of Box C.

☐ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"D" document cited by the applicant in the international application

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

24 August 2020

Date of mailing of the international search report

16 OCT 2020

Name and mailing address of the ISA/US

Mail Stop PCT, Attn: ISA/US, Commissioner for Patents
P.O. Box 1450, Alexandria, Virginia 22313-1450

Facsimile No. 571-273-8300

Authorized officer

Lee Young

Telephone No. PCT Helpdesk: 571-272-4300

Continuation of:
Box No III (unity of invention is lacking)

Special Technical Feature

Group I includes the special technical feature of sequencing the enriched cDNA fragments to obtain 5'- and 3'-end sequences; and mapping the sequences to a reference genome to provide a full-length transcript profile of genomic RNA, not required by Groups II-III.

Group II includes the special technical feature of making a DNA library, not required by Groups I and III.

Group III includes the special technical feature of kit for simultaneous 5'- and 3'-end RNA sequence capture comprising directions for use of said kit, not required by Groups I-II.

Common Technical Features

The inventions of Groups I-III share the technical features of RNA transcript; isolating RNA (Group III, isolated RNA); ligating an adaptor to 3' end of RNA; converting the 3'-end ligated RNA to cDNA (Group III, initiating transcription of the 3'-end ligated RNA to generate cDNA); circularizing the cDNA (Group III, circularized cDNA); and fragmenting the cDNA (Group III, isolates cDNA fragments); and --Groups I-II further share the technical feature of genomic RNA.

Therefore, in a summary, the common technical features shared by inventions of Groups I-III can be covered by:

A method of transcript profiling of RNA, comprising the steps of:

- (a) isolating genomic RNA;
- (b) ligating an adaptor to 3' end of genomic RNA to provide 3' end ligated genomic RNA;
- (c) converting the 3'-end ligated genomic RNA to cDNA;
- (d) circularizing the cDNA; and
- (e) fragmenting cDNA to provide fragmented cDNA (part of claim 1, comprising part of claim 2)

However, these shared technical features do not represent a contribution over prior art as being obvious over US 2015/0099671 A1 to UNIVERSITY OF MASSACHUSETTS (hereinafter 'Univ_MA'), in view of US 2014/0213485 A1 to YALE UNIVERSITY (hereinafter 'YALE') as follows:

Univ_MA discloses a method of transcript profiling of RNA (para [0003] - 'profiling of the eukaryotic transcriptome in a genome-wide manner and...for the genome-wide analysis and quantification of RNA samples'; para [0076] - 'RNA-Seq ...start by: (1) Reverse transcription (RT) of full length RNAs with primers containing a 3' randomized region), comprising the steps of:

- (a) isolating genomic RNA (para [0012] - 'all of the RNA molecules obtained after the purifying step'; para [0003]; para [0039] - 'single molecule mapping of both the 5' and 3' ends of RNA molecules in a genome-wide manner and from a single sample');
- (b) ligating an adaptor to 3' end of genomic RNA to provide 3' end ligated genomic RNA (para [0009] - 'preparation of RNA molecules from a biological sample for sequencing is provided. ... ligating a DNA adaptor to the 3' end of the RNA molecules'; para [0003] - 'profiling of the eukaryotic transcriptome in a genome-wide manner and...for the genome-wide analysis and quantification of RNA samples');
- (c) converting the 3'-end ligated genomic RNA to cDNA (para [0009] - 'ligating a DNA adaptor to the 3' end of the RNA molecules, ... reverse transcribing the RNA molecules using a unique DNA primer under conditions that optimize the reverse transcription reaction to produce single-stranded cDNA molecules'; para [0003] - 'profiling of the eukaryotic transcriptome in a genome-wide ...for the genome-wide analysis and quantification of RNA samples');
- (d) circularizing the cDNA (para [0009] - 'ligating a DNA adaptor to the 3' end of the RNA molecules, ...reverse transcription reaction to produce single-stranded cDNA molecules; purifying the cDNA molecules ... circularizing the purified cDNA molecules under conditions that optimize the circularization reaction'; para [0058] - 'single-stranded cDNA molecules are circularized (FIG. 16B)... use an intramolecular ligase capable of circularizing DNA'; para [0047] - 'the cDNA molecules are circularized'; para [0078] - 'cDNA circularization is quantitative (see below), we chose to place our barcodes at the 3' end of the 5' adaptor (i.e., between the forward primer sequence and the captured sequences)');

Univ_MA further discloses amplifying the circularized cDNA for sequencing (para [0009] - 'amplifying the circularized cDNA molecules under conditions that optimize the amplification reaction, thereby preparing RNA molecules from a biological sample for sequencing'), and the method can comprise fragmenting mRNA before circularizing cDNA (para [0015] - 'reverse transcribing the fragmented mRNA molecules to produce single-stranded complementary DNA (cDNA) molecules; circularizing the singlestranded cDNA molecules; and amplifying the circularized cDNA molecules, thereby preparing the mRNA molecules in the biological sample for sequencing').

Univ_MA does not specifically teach the method further comprising (e) fragmenting the cDNA. YALE discloses a method for preparing cDNA from RNA for full-length sequencing (Abstract - 'preparation of cDNA libraries that are representative of mRNA or whole RNA expressed by the cell or cells. The cDNA is suitable for sequencing'; para [0010] - 'full-length RNA (cDNA) sequencing for low quantities of cells and single cells'), wherein the cDNA can be circularized and amplified (para [0013] - 'the cDNA is circularized and amplified by a multiple displacement amplification'), and further cDNA can be fragmented for further analysis including adding adapters for sequencing (para[0021] - 'Additional steps can include fragmenting the cDNA library to achieve certain sized fragments suitable for subsequence analysis...adaptor oligonucleotides are ligated to the cDNA to facilitate sequencing'; para [0121] - 'the fragment cDNA separated by size to collect a specific size or size range of fragments for later analysis').

It would have been obvious to one of ordinary skill in the art at the time the invention was made to combine the teachings of Univ_MA and YALE, to obtain a method of transcript profiling of RNA, comprising the steps of: (a) isolating genomic RNA; (b) ligating an adaptor to 3' end of genomic RNA to provide 3' end ligated genomic RNA; (c) converting the 3'-end ligated genomic RNA to cDNA; (d) circularizing the cDNA, based on the teaching of Univ_MA, and further the method further comprising (e) fragmenting the cDNA, based on the combination of YALE and Univ_MA, in order to combine methods and technologies available in the art for facilitating performing transcript profiling of RNA with desired results with an expected success and without undue experimentation.

*****Continued in the next extra sheet*****

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US 20/34127

Continuation of:
The previous extra sheet - Box No III (unity of invention is lacking)

Without a shared special technical feature, the inventions lack unity with one another.

Groups I-III therefore lack unity under PCT Rule 13 because they do not share a same or corresponding special technical feature.

Continuation of item 4: Claims 7-30, 34-48, 53 are not drafted in accordance with the second and third sentences of Rule 6.4 (a). These claims are improper multiple dependent claims.