EmPC-seq: Accurate RNA-sequencing and Bioinformatics Platform to Map RNA Polymerases and Remove Background Error

Yuqing Wang[1, 2, #], Tin Hang Chong[3, #], Ilona Christy Unarta[2, #], Xinzhou Xu[2], Gianmarco D. Suarez[3], Jiguang Wang[2, 4, 7], John T. Lis[5, 6], Xuhui Huang[1, 2, 3, 7] and Peter Pak-Hang Cheung[1, 3, *]

[1]The Hong Kong University of Science and Technology -Shenzhen Research Institute, Shenzhen, China; [2]Bioengineering Graduate Program, Department of Chemical and Biological Engineering, The Hong Kong University of Science and Technology, Hong Kong SAR; [3]Department of Chemistry, State Key Laboratory of Molecular Neuroscience, The Hong Kong University of Science and Technology, Hong Kong SAR; [4]Division of Life Science, The Hong Kong University of Science and Technology, Hong Kong SAR; [5]Department of Molecular Biology and Genetics, Cornell University, Ithaca, USA; [6]The HKUST Jockey Club Institute for Advanced Study (IAS), The Hong Kong University of Science and Technology, Hong Kong SAR; [7]Hong Kong Center for Neurodegenerative Diseases, Hong Kong Science Park, Hong Kong SAR

*For correspondence: ppcheung@ust.hk

[#]Contributed equally to this work

[Abstract] Transcription errors can substantially affect metabolic processes in organisms by altering the epigenome and causing misincorporations in mRNA, which is translated into aberrant mutant proteins. Moreover, within eukaryotic genomes there are specific Transcription Error-Enriched genomic Loci (TEELs) which are transcribed by RNA polymerases with significantly higher error rates and hypothesized to have implications in cancer, aging, and diseases such as Down syndrome and Alzheimer’s. Therefore, research into transcription errors is of growing importance within the field of genetics. Nevertheless, methodological barriers limit the progress in accurately identifying transcription errors. Pro-Seq and NET-Seq can purify nascent RNA and map RNA polymerases along the genome but cannot be used to identify transcriptional mutations. Here we present background Error Model-coupled Precision nuclear run-on Circular-sequencing (EmPC-seq), a method combining a nuclear run-on assay and circular sequencing with a background error model to precisely detect nascent transcription errors and effectively discern TEELs within the genome.

Keywords: Transcriptional mutagenesis, RNA polymerase, Nascent RNA, Deep RNA sequencing, Accurate RNA sequencing

[Background] Transcriptional errors due to ribonucleotide misincorporation are ubiquitous to all living organisms (Carey, 2015). Given that each messenger RNA (mRNA) can be translated 2-4 thousand times (Schwanhausser *et al.*, 2011) and many special RNAs are expressed only once per cell at a given time (Islam *et al.*, 2011; Pelechano *et al.*, 2010), even a single transcription error at a critical residue can make large differences in a specific protein’s expression. In addition, transcriptional errors can accelerate protein aggregation leading to age-related diseases in humans (van Leeuwen *et al.*, 1998).

While transcription errors are conventionally held to have a random distribution across the genome, there is evidence indicating that transcription errors could be enriched at certain structural motifs and specific genomic regions (Imashimizu *et al.*, 2013; van Leeuwen *et al.*, 1998). These Transcription Error-Enriched genomic Loci (TEELs) have notable biological significance in various diseases such as Down syndrome and Alzheimer's and are gaining attention in genetics research (Burns *et al.*, 2010; Saxowsky *et al.*, 2008). Unfortunately, there are major challenges that must be circumvented for the study of transcriptional error due to RNA polymerase unconfounded by RNA-editing processes such as those from post-transcriptional modifications. This requires purification of nascent RNA coupled with a highly accurate RNA sequencing method that can identify TEELs and elucidate transcriptional regulation and dysregulation contributing to transcriptional errors with implications to diseases.

There are several complications which impede the accurate detection of *de novo* transcription errors. The first challenge is eliminating the noise from post-transcriptional modifications, which requires the purification of nascent RNA freshly made by RNA polymerases. Hence, current RNA sequencing (RNA-seq) studies on transcriptional errors often overlook this requirement and therefore overestimate transcription error rates. The second challenge is rectifying the systematic noise from Next Generation Sequencing (NGS). NGS on average misreads approximately one base in every 1,000 (Minoche *et al.*, 2011), and this is further compounded by the fact that reverse transcriptase (required for generating cDNA for NGS) misincorporates one base in every 10,000 (Ji and Loeb, 1992). The third challenge is computationally discerning TEELs from background noise. Even with accurate sequencing data, it is still difficult to computationally identify TEELs amongst background errors which are stochastically introduced by RNA polymerases (de Mercoyrol *et al.*, 1992). Here, we present our background Error Model-coupled Precision nuclear run-on Circular-sequencing (EmPC-seq) method (Figure 1) to overcome these three main challenges. EmPC-seq consists of three core components: (1) a nuclear run-on assay to capture nascent RNA before post-transcriptional modifications (Mahat *et al.*, 2016), (2) a circular-resequencing step that generates cDNA via rolling-cycle reverse transcription of circularized nascent RNA molecules (Acevedo and Andino, 2014) to improve sequencing accuracy by generating tandem cDNA repeats of the same circularized RNA molecule by rolling circle amplification so that the RNA molecule can be sequenced multiple times. (3) We also developed a background error model algorithmic analysis to remove stochastic background noise by simulating *de novo* sequencing data and subsequent error to serve as a control group carrying background alterations from sequencing noise, non-uniform sequencing depth, and alignment artifacts (Cheung *et al.*, 2020). EmPC-seq aims to detect nascent transcriptional errors and elucidate their origins that may have implications to diseases.

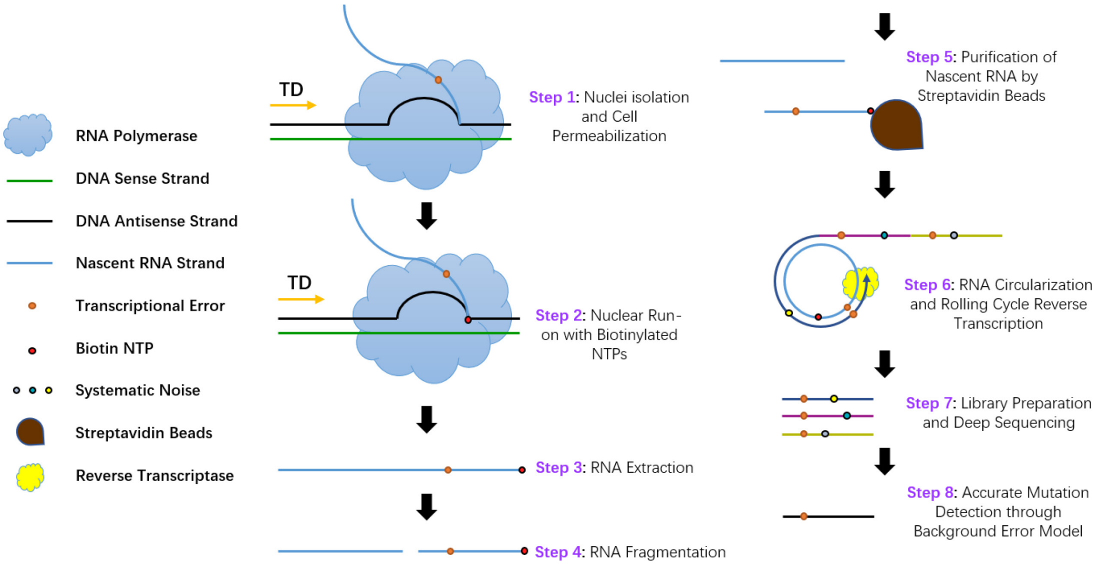
RNA Polymerase
DNA Sense Strand
DNA Antisense Strand
Nascent RNA Strand
Transcriptional Error
Biotin NTP
Systematic Noise
Streptavidin Beads
Reverse Transcriptase
TD
Step 1: Nuclei isolation and Cell Permeabilization
TD
Step 2: Nuclear Run-on with Biotinylated NTPs
Step 3: RNA Extraction
Step 4: RNA Fragmentation
Step 5: Purification of Nascent RNA by Streptavidin Beads
Step 6: RNA Circularization and Rolling Cycle Reverse Transcription
Step 7: Library Preparation and Deep Sequencing
Step 8: Accurate Mutation Detection through Background Error Model

Figure 1. Schematic of EmPC-seq. Real transcription errors are represented using orange dots. Dots in other colors represent systematic noise, including enzymatic errors and sequencing errors. (Step 1) Yeast cell is permeabilized. (Step 2) *In vivo* transcription is halted by adding all 4 kinds of biotinylated NTPs during the Nuclear Run-on assay. (Step 3) Yeast total RNA is extracted and purified via ethanol precipitation. (Step 4) RNA is fragmented with base hydrolysis into short (60-100nt) RNAs. (Step 5) Biotin-labeled nascent RNA is enriched through Streptavidin bead purification. (Step 6) Re-purified nascent RNAs are circularized by RNA ligase and processed into tandem copy cDNAs through rolling circle reverse transcription. (Step 7) Library DNA is prepared with a kit and then submitted for Next Generation Sequencing. (Step 8) Transcription errors are accurately detected by combining consensus sequence results with our background error model. This schematic is adapted from Cheung *et al.* (2020).

Materials and Reagents

1. 1.5 ml tubes
2. Pipette tips
3. Cuvettes
4. 0.22 µm filter
5. W303 yeast cells (GenBank Number: JRIU00000000)
6. 1 M Dithiothreitol (DTT, ThermoFisher, catalog number: P2325)
7. Diethyl pyrocarbonate (Sigma, catalog number: 40718)
8. UltraPure^{TM} DNase/RNase-Free Distilled Water (ThermoFisher, catalog number: 10977015)
9. Yeast Extract (Sigma, catalog number: Y1625)
10. Peptone (Sigma, catalog number: P0556)
11. D-(+)-Glucose (Sigma, catalog number: G8270)
12. Adenine (Sigma, catalog number: A8626)
13. N-Lauroylsarcosine sodium salt, Sarkosyl (Sigma, catalog number: L9150)

14. Trizma® hydrochloride (Sigma, catalog number: T5941)
15. Potassium chloride (Sigma, catalog number: P9333)
16. Magnesium chloride (Sigma, catalog number: M8266)
17. Biotinylated Nucleotides (Jena Bioscience, catalog number: NU series)
18. RNase Inhibitor, Murine (NEB, catalog number: M0314L)
19. Diethyl pyrocarbonate, DEPC (Sigma, catalog number: 40718)
20. Liquified Phenol (Sigma, catalog number: P9346)
21. Sodium acetate (Sigma, catalog number: S2889)
22. Ethylenediaminetetraacetic acid, EDTA (Sigma, catalog number: EDS)
23. Sodium dodecyl sulfate (Sigma, catalog number: L3771)
24. Chloroform (Sigma, catalog number: C2432)
25. GlycoBlue™ Coprecipitant (ThermoFisher, catalog number: AM9515)
26. Ethyl alcohol, Pure (Sigma, catalog number: E7023)
27. Sodium hydroxide (Sigma, catalog number: S8045)
28. Triton™ X-100 (Sigma, catalog number: X100)
29. Monarch® RNA Cleanup Kit (NEB, catalog number: T2030L)
30. Dynabeads™ M-280 Streptavidin (ThermoFisher, catalog number: 60210)
31. Sodium chloride (Sigma, catalog number: S7653)
32. TRIzol™ Reagent (ThermoFisher, catalog number: 15596018)
33. Ambion™ T4 RNA Ligase (ThermoFisher, catalog number: AM2141)
34. T4 Polynucleotide Kinase (NEB, catalog number: M0201S)
35. Polyethylene glycol 8000 (Sigma, catalog number: 1546605)
36. Adenosine 5'-Triphosphate, ATP (NEB, catalog number: P0756S)
37. dNTP Mix (ThermoFisher, catalog number: 18427088)
38. Random Hexamer Primer (ThermoFisher, catalog number: SO142)
39. SuperScript™ III First-Strand Synthesis System (ThermoFisher, catalog number: 18080051)
40. NEBNext® Ultra™ II Directional RNA Second Strand Synthesis Module (NEB, catalog number: E7550L)
41. MinElute PCR Purification Kit Print (QIAGEN, catalog number: 28004)
42. NEBNext® Ultra™ II DNA Library Prep Kit for Illumina® (NEB, catalog number: E7645L)
43. Qubit™ dsDNA HS Assay Kit (ThermoFisher, catalog number: Q32851)
44. YEPD medium (see Recipes)
45. 2.5× Transcription buffer (see Recipes)
46. AES Buffer (see Recipes)
47. Beads washing buffer (see Recipes)
48. Binding washing buffer (see Recipes)
49. Low Salt washing buffer (see Recipes)
50. High Salt washing buffer (see Recipes)
51. DEPC-H_2O (see Recipes)

52. 1 M sodium acetate solution (see Recipes)

Equipment

1. Eppendorf® Research® Plus Pipettes (Eppendorf, catalog number: EP series)
2. MaxQ™ 6000 Incubated/Refrigerated Stackable Shakers (ThermoFisher, catalog number: SHKE6000)
3. Eppendorf BioPhotometer® (Eppendorf, model: D30)
4. Megafuge® (Heraeus, model: 1.0R)
5. 5 Liter General Purpose Water Bath (PolyScience, catalog Number: WBE05A11B)
6. NEBNext® Magnetic Separation Rack (NEB, catalog number: S1515S)
7. Roto-Shake Genie® (Zymo Research, catalog number: S5008)
8. ProFlex PCR System (ThermoFisher, catalog number: 4484075)
9. 5200 Fragment Analyzer System (Agilent, catalog number: M5310AA)

Software

1. ProSize (Agilent, https://explore.agilent.com/Software-Download-Fragment-Analyzer-Prosize)
2. Python (version 2.7.12, https://www.python.org/)
3. Cython (version 0.23.4, https://cython.org/)
4. NumPy (version 1.11.0, https://numpy.org/)
5. SciPy (version 0.17.0, https://www.scipy.org/)
6. Burrows-Wheeler Aligner (version 0.7.17-r1188, http://bio-bwa.sourceforge.net/)
7. samtools (version 1.9, http://www.htslib.org/)
8. pysam (version 0.15.0, https://pysam.readthedocs.io/en/latest/installation.html)
9. matplotlib (version 2.2.2, https://matplotlib.org/)

Procedure

A. Prepare yeast cells for nuclear run-on assay

1. Inoculate a single yeast colony in 5 ml YEPD medium.
2. Incubate overnight in a 30 °C incubator with the rotation speed set as 200 rpm.
3. Measure the optical density (OD_{600nm}) of each cell culture.
4. Dilute and re-inoculate the yeast to 10 ml of YEPD medium with an OD_{600} of 0.2.
5. Incubate the yeast cells in a 30 °C incubator until they reach an OD_{600} of 0.4-0.6.

B. Yeast nuclear run-on assay

1. Centrifuge the yeast cells at 2,000 × *g* for 5 min at 4 °C and draw off the supernatant.
2. Wash the cell pellet with 5 ml of ice-cold DEPC-H_2O by pipetting up and down.

3. Re-centrifuge the yeast cells at 2,000 × *g* for 5 min at 4 °C and draw off the supernatant.
4. Resuspend the cell pellet with 4.75 ml of ice-cold DEPC-H_2O.
5. Add 250 µl of 10% Sarkosyl solution and gently mix the solution.
6. Hold on ice for 20 min.
7. Centrifuge the yeast cells at 400 × *g* for 5 min at 4 °C and draw off the supernatant.
8. Resuspend the cell pellet with 120 µl of 2.5× Transcription Buffer, 6 µl of 0.1 M DTT, 3.5 µl of each kind of 1 mM biotin-NTP, and 3 µl of RNase Inhibitor to a total volume as 143 µl.
9. Add 142 µl DEPC-H_2O and 15 µl of 10% Sarkosyl and gently mix the solution.
10. Incubate the mixture at 30 °C for 5 min with gentle pipetting up and down at the half-time point (2.5 min).

C. Nascent RNA extraction

1. Centrifuge the reaction mixture at 400 × *g* for 5 min at 4 °C and draw off the supernatant.
2. Quickly resuspend the pellet in 500 µl of Liquified phenol.
3. Add 500 µl of AES Buffer and pipette up and down to rupture cells.
4. Incubate the mixture at 65 °C for 5 min with vortex once every minute.
5. Incubate the mixture on ice for 5 min.
6. Add 200 µl of chloroform to it and vortex for 30 s.
7. Incubate the mixture at room temperature for 2 min.
8. Centrifuge it at 14,000 × *g* for 5 min at 4 °C.
9. Draw off the aqueous layer and put it into another new tubes (be careful to avoid the interface).
10. Add 1 M sodium acetate solution to the aqueous layer to a final concentration of 200 mM.
11. Split the aqueous RNA solution mixed with sodium acetate into two 1.5 ml tubes.
12. Add 1 µl of GlycoBlue^{TM} Coprecipitant and 3× volume of 100% ethanol into the two 1.5 ml tube containing aqueous layer in Step C11 (this mixture can be stored overnight).
13. Centrifuge the mixture at 14,000 × *g* for 30 min at 4 °C and draw off the supernatant (be careful to avoid disturbing the blue pellet).
14. Wash the pellet with freshly prepared RNase-free 75% ethanol.
15. Centrifuge it at 14,000 × *g* for 5 min at 4 °C and draw off the supernatant.
16. Let the pellet dry for 5 min and resuspend it in 20 µl of DEPC-H_2O.

D. RNA fragmentation by Base Hydrolysis

1. Heat denature the RNA solution at 65 °C for 40 s and place it on ice.
2. Add 5µl of ice cold 1 N NaOH and incubate it on ice for 10 min.
3. Add 25 µl of 1 M Tris-HCl (pH 6.8).
4. Purify the RNA with Monarch® RNA Cleanup Kit (10 µg) with elution volume of 20 µl.
5. Add 1 µl RNase inhibitor.

E. Biotin-labeled nascent RNA pull-out

1. Wash 30 µl of Streptavidin M280 beads with 500 µl of beads washing buffer by adding and mixing the washing buffer with the beads, setting on a magnet for 1 min and drawing off the supernatant.
2. Wash the beads twice with 500 µl of 100 mM NaCl solution, the operation is the same as Step E1.
3. Resuspend the beads in 50 µl of binding washing buffer.
4. Mix the purified RNA in section D to 50 µl with binding washing buffer.
5. Mix the RNA solution with the resuspended beads.
6. Incubate the mixture at room temperature on a rotator for 20 min.
7. Place the mixture on a magnet for 1 min and draw off the supernatant.
8. Resuspend the beads in 500 µl of high salt washing buffer.
9. Place the beads on a magnet for 1 min and draw off the supernatant.
10. Repeat Steps E8-E10.
11. Wash the beads twice with 500 µl of binding washing buffer.
12. Wash the beads twice with 500 µl of low salt washing buffer.
13. Resuspend the beads in 300 µl of TRIzol^{TM} solution.
14. Incubate the mixture on ice for 3 min.
15. Add 60 µl of chloroform to the mixture and vortex it thoroughly for at least 20 s.
16. Centrifuge the mixture at 20,000 × *g* for 5 min at 4 °C.
17. Purify the RNA in the aqueous layer (be careful to avoid the interface) with Monarch® RNA Cleanup Kit (10 µg) with an elution volume of 20 µl.

F. Cyclization of RNA sample

1. Heat denature the purified RNA at 65 °C for 1 min and place it on ice.
2. Add the following reagents to 19 µl of the RNA solution: 4 µl of 10× T4 ligase I reaction buffer, 2 µl of T4 Ligase I enzyme, 2 µl of PNK enzyme, 1 µl of RNase inhibitor, 8 µl of 50% PEG8000 and 4 µl of 10 mM ATP solution.
3. Incubate the mixture at 25 °C for 2 h or at 16 °C overnight.
4. Purify the RNA in the mixture with Monarch® RNA Cleanup Kit (10 µg) with elution volume of 20 µl.

G. Rolling-cycle reverse transcription

1. Add the following reagents to 9 µl of the RNA solution: 4 µl of 10mM dNTPs solution, 4 µl of 50 ng/µl Random Hexamers and 3 µl of RNase-free water.
2. Heat denature the reaction mix at 65 °C for 1 min and place it on ice for more than 2 min.
3. Add the following materials into the reaction mix: 8 µl of 5× First-strand synthesis buffer, 4 µl of 0.1 mM DTT solution, 4 µl of SuperScript III enzyme and 8µl of water

Note: The buffers and enzymes used here are from SuperScript^{TM} III First-Strand Synthesis System Kit.

4. Incubate the mixture at 25 °C for 10 min and then incubate at 42 °C for 20 min.
5. Purify the RNA in the mixture with Monarch® RNA Cleanup Kit (10 µg) with an elution volume of 20 µl.
6. Add 20 µl of RNase-free water to the RNA solution.

H. Second strand synthesis

1. Add 38 µl of RNA solution with the following materials: 8 µl of 10× Second strand synthesis buffer, 4 µl of Second strand synthesis enzyme mix and 30 µl of RNase-free water
   Note: The buffers and enzymes used here are from NEBNext® Ultra^{TM} II Directional RNA Second Strand Synthesis Module Kit.
2. Incubate the reaction mix at 16 °C for 2 h with the thermocycler lid set at 50 °C.
3. Purify the DNA in the mixture with the MinElute PCR Purification Kit.

I. Library preparation and submitting for NGS

1. Measure the concentration of cDNA in step H3 by using Qubit^{TM} dsDNA HS Assay Kit.
2. Prepare the sequencing library with the NEBNext® Ultra^{TM} II DNA Library Prep Kit for Illumina® according to the manufacturer's instructions by end preparation, adaptor ligation, size selection and PCR enrichment.
3. Submit the cDNA library to MiSeq Illumina platform with read length as single-end 300 base pairs for performing Next Generation Sequencing.

Data analysis

1. Use Ubuntu 16.04 LTS (Xenial Xerus) for running the script.
2. Download the script from GitHub (https://github.com/ustsam/Em-PC_seq).
3. Unzip the file.
4. Make sure all software in software section is installed.
5. Open a terminal and enter the script directory
6. Compile the code by typing "python setup_newreloc.py build_ext –inplace" in a terminal.
7. Call the function using the command in the terminal:

```
"./run_noQsfilter.sh {PATH to the output directory} {PATH to the reference file} {PATH to the script directory} DUMMY 2 ${twice of the max readlength} ${PATH of the data file in gzipped form}."
```

8. data.sam.gz can be found in the output directory. "data.sam.gz" contains all the transcripts after the consensus generation step. The file is in compressed sam format. To decompress the file,

Please cite this article as: Wang *et. al.*, (2021). EmPC-seq: Accurate RNA-sequencing and Bioinformatics Platform to Map RNA Polymerases and Remove Background Error,*Bio-protocol* 11 (4): e3921. DOI: 10.21769/BioProtoc.3921.

one can type the following command in a linux terminal:

```
gunzip data.sam.gz
```

9. Run the command in the output directory to perform data analysis:

```
“bash data_analysis.sh {PATH to the output directory} {PATH to the reference file} {PATH to the script directory} 1 {maximum depth per site} {minimum base quality} {minimum mapping quality} {number of simulation fastq files generated}.”
```

Ambiguity is defined as the number of ways to map a transcript to the reference genome (Cheung *et al.*, 2020). We used the strictest threshold, which is ambiguity=1, minimum mapping quality=30, and minimum base quality=30.

10. Run the following command to reproduce the analysis (assuming the script and output directory, as well as the reference fasta file are in the current directory):

```
“bash data_analysis.sh ./ ./rDNA1.fasta ./ 1 500000 30 30 100.”
```

11. Run the following command to preprocess the output from “data_analysis.sh” script for plotting:

```
“bash plot.sh {PATH to the output directory} {PATH to the reference file} {PATH to the script directory} {ambiguity threshold} {maximum depth per site} {minimum base quality} {minimum mapping quality} 1.”
```

12. The output of plot.sh files figures (.png format) are the following:
    a. “Distribution_NumberOfWaysToMap.png”: the distribution of the number of ways the transcripts can be mapped to the reference genome (ambiguity). The y-axis is the number of transcripts and the x-axis is ambiguity. An example is shown below (Figure 2).

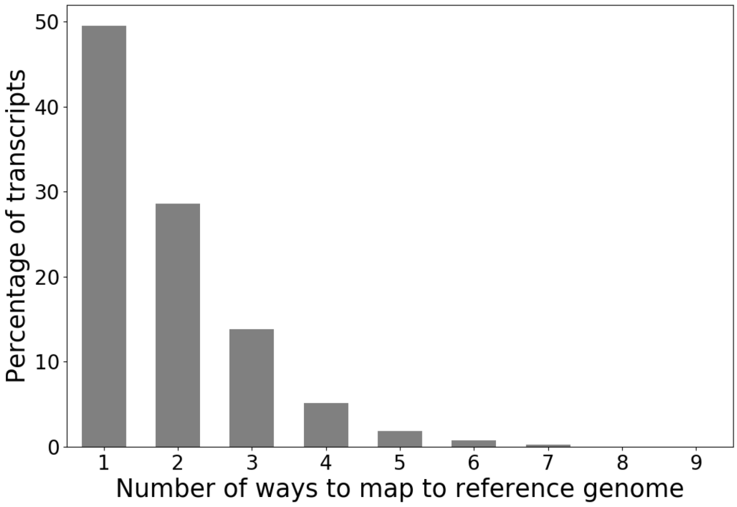

Figure 2. An example of a figure of the percentage of transcripts with a particular ambiguity. The ambiguity is the number of ways a transcript can be mapped to the reference genome. Ambiguity arises from the lack of information about the start of transcription due to the circularization step during the experiment.

b. "MutationTypeSpectrum.png": The mutational frequency for each type of mutation in the RNA transcript. The mutational frequency is the number of errors divided by the coverage of the corresponding reference base. For example: number of A → C errors divided by coverage of base A. An example is shown below (Figure 3).

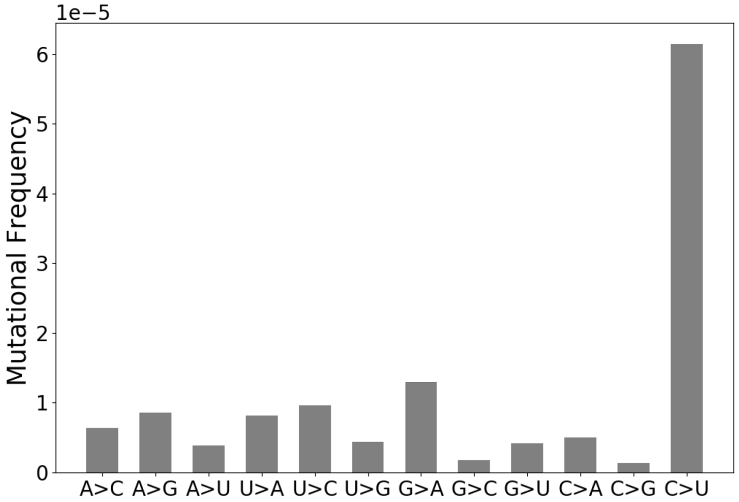

Figure 3. An example of a figure of the mutational spectrum. The mutational frequency of each type of substitution on the RNA chain is shown.

c. "Muta_Frequency_inChrom_***.png": The mutational frequency along the sites in the chromosome for the experimental and simulation data. "***" is the name of the chromosome. The transcription error-enriched genomic loci (TEEL) is shown as red dots. An example is shown below (Figure 4).

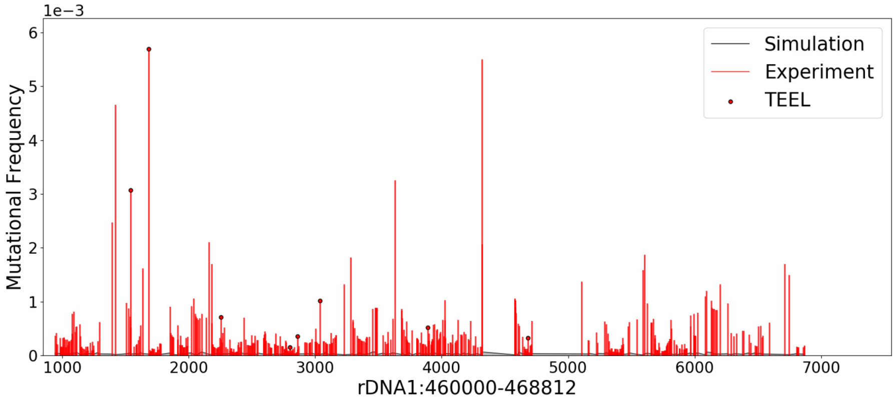

Figure 4. An example of a figure of the mutational frequency across the genomic sites. The mutational frequency for each site in the chromosome is shown as the red lines. The background error obtained by simulation is shown as the gray line. The sites that are identified as TEEL are shown as red dots.

d. "ErrorRate_per_PositionInTranscripts.png": The error rate at each position in the transcript. The Position 0 corresponds to the 3' end of the transcript. An example is shown below (Figure 5).

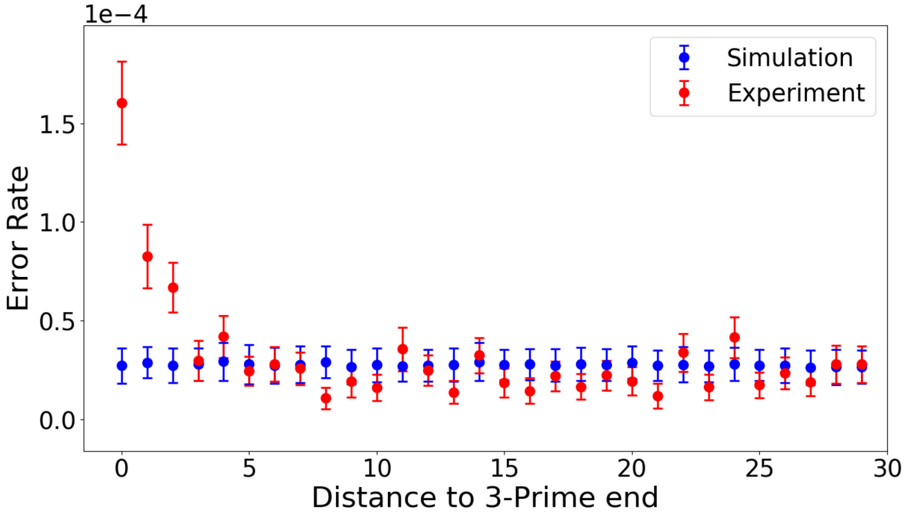

Figure 5. An example of a figure of the error rate at each positional site in the transcript. The error rate for each position from the experimental data is shown as red dots. The error rate for each position from the simulation data is shown as blue dots. The x-axis is the position in the transcript, where position 0 corresponds to the 3' end of the transcript. The increased error rate could be correlated by pausing of the RNA Polymerase, yet the order of event could not be determined in this experiment. Two scenarios could fit the observed data, either RNA Polymerase pauses to fix the error or RNA Polymerase pauses due to the misincorporation.

13. If one would prefer to use other software tools, the output text files to plot the figures are:
    a. "Distribution_of_Ambiguity.txt" can be used to plot the figure described in step 12a.
    b. "MutationTypeSpectrum.txt" can be used to plot the figure described in step 12b.
    c. "MutationalFrequency_Exp_chrom_***.txt" can be used to plot the figure described in step 12c and "MutationalFrequency_Sim_chrom_***.txt" are the files containing the Mutational Frequencies per site in a chromosome for experimental and simulation, respectively. "MutationalFrequency_TEEL_chrom_***.txt" contains the mutational frequency of the sites considered as TEEL. The first column is the position in the chromosome and the second column is the mutational frequency.
    d. "MutationalFrequency_PerPositionInTranscript.txt" can be used to plot the figure described in step 12d. The first column is the position in the transcript. The second and third column is the average and standard deviation of the simulation data. The fourth and fifth column is the average and standard deviation of the experimental data. The average and standard deviation of experimental data are calculated from error rate binomial distribution estimated by maximum likelihood.

Notes

1. RNA digestion: There are several alternative methods of RNA fragmentation since the type of RNA and the expected final size impact the sequencing technology to be used. For instance, ultra-sonication is widely used to fragment DNA for NGS sequencing libraries and can be optimized for RNA fragmentation as well, however, the size of the product is around 100-200 nt and careful consideration is needed when trying to obtain more than 2 tandem repeats during rolling-circle reverse transcription. Meanwhile, another biological digestion method using RNase III enzyme can fragment RNA molecules into smaller sizes (60-120 nt) but the recovery rate is relatively low (around 10%).
2. RNA electrophoresis: We suggest collecting RNA samples at each step for troubleshooting. RNA electrophoresis can assay the size and amount of RNA. For small amounts of RNA larger than 200 bp but smaller than 6,000 bp, a Fragment Analyzer (Agilent) can be used since it needs no more than 2 ng of RNA and can generate a comprehensive size/amount spectrum of the RNA samples.
3. RNA purification: Monarch® RNA Cleanup Kit is used to replace ethanol precipitation as it is quicker and easier to operate with multiple samples. Ethanol precipitation can still be used if RNA cleanup kit is not available but improper phase separation of aqueous layer may result in contamination with TRIzol™.
4. Library Preparation: In this protocol we use the NEBNext® UltraTM II DNA Library Prep Kit for Illumina® to prepare the cDNA library for sequencing. There are also other alternative kits for library preparation such as Ovation® Ultralow V2 DNA-Seq Library Preparation Kit and KAPA HTP/LTP Library Preparation Kits. While the steps of preparing cDNA libraries using different

library preparation kits are similar, there are several points that are needed to be mentioned. Firstly, the dilution of the adaptor and the PCR enrichment cycles are different among these kits according to the specific mass of cDNA used for library preparation. For example, NEBNext® Ultra™ II DNA Library Prep Kit for Illumina® needs 25-fold dilution of adaptor and a 10-cycles PCR if the sample mass is around 1 ng. Secondly, different library preparation kits have designated index primers, it should be considered that each kit might have their own requirement about the index combination based on the number of cDNA samples. For instance, NEBNext® Ultra™ II DNA Library Prep Kit for Illumina® suggests several unique combinations when the number of cDNA samples is less than 7.

5. DEPC Operation: Solid DEPC is toxic and harmful. Thus, there should be proper eye-shield, face-shield, full-face respirator, gloves and chemical hood when performing any operation with solid DEPC.
6. Data analysis: Due to circularization, the information about the transcriptional direction and starting point of the transcripts are unknown. The scripts prepared are specifically used to treat rRNA transcripts, where they assume that all the transcripts are transcribed in the negative direction, *i.e.*, the starting position is the 3' end of the transcript. The consensus generation step (steps 7 and 8) is general for all raw data and does not depend on the transcription direction. However, starting from step 9, we assume negative direction of transcription, thus changes are required to generalize the script for other transcripts. The following files will need to be modified: "data_analysis.sh", "pysam_make_pileup.py", "simulation.py", "binomial_distribution.py", "plotting.py" and "plotting_preprocess.py". In addition, ambiguity occurs due to re-localization during data treatment. Ambiguity is defined as the number of ways a transcript can be mapped to the reference genome (Cheung *et al.*, 2020). The script assumes the strictest requirement, where only the transcripts with ambiguity equals to one are considered in analysis.

Recipes

1. 1 L YEPD medium
   a. Add 10 g of Yeast Extract
   b. Add 20 g of peptone
   c. Add 850 ml dH_2O
   d. Dissolve 20 g glucose in 100 ml dH_2O in a new bottle
   e. Dissolve 40 mg adenine in 50 ml dH_2O in a new bottle
   f. Autoclave above reagents at 120 °C for 15 min
   g. Add glucose and adenine solutions into the pre-medium after it cools down
2. 2.5× Transcription buffer
   a. Add Tris-HCl (pH=7.7) to the final concentration as 50 mM
   b. Add KCl to the final concentration as 500 mM
   c. Add $MgCl_2$ to the final concentration as 12.5 mM

   d. Add DEPC-H_2O to a total volume as 50 ml

3. AES Buffer
   a. Add Sodium Acetate (pH=5.3) to the final concentration as 50 mM
   b. Add EDTA to the final concentration as 10 mM
   c. Add SDS to the final concentration as 1% (w/v)
   d. Add DEPC-H_2O to a total volume as 50 ml
4. Beads washing buffer
   a. Add NaOH to the final concentration as 0.1 N
   b. Add NaCl to the final concentration as 50 mM
   c. Add DEPC-H_2O to a total volume as 50 ml
5. Binding washing buffer
   a. Add Tris-HCl (pH = 7.4) to the final concentration as 10 mM
   b. Add NaCl to the final concentration as 300 mM
   c. Add Triton^{TM} X-100 to the final concentration as 0.1% (v/v)
   d. Add DEPC-H_2O to a total volume as 50 ml
6. Low Salt washing buffer
   a. Add Tris-HCl (pH = 7.4) to the final concentration as 5 mM
   b. Add Triton^{TM} X-100 to the final concentration as 0.1% (v/v)
   c. Add DEPC-H_2O to a total volume as 50 ml
7. High Salt washing buffer
   a. Add Tris-HCl (pH = 7.4) to the final concentration as 50 mM
   b. Add NaCl to the final concentration as 2 M
   c. Add Triton^{TM} X-100 to the final concentration as 0.5% (v/v)
   d. Add DEPC-H_2O to a total volume as 50 ml
8. DEPC-H_2O
   a. Add 0.1% (v/v) DEPC to RNase-free Water
   b. Mix overnight and then autoclave
   c. Filter-sterilize the solution with a 0.22 µm filter
   d. Store in room temperature for up to one year
9. 1 M sodium acetate solution
   a. Add 4.1 g solid sodium acetate
   b. Add DEPC-H_2O to a total volume as 50 ml

Acknowledgments

This work was supported by Shenzhen Science and Technology Innovation Committee (JCYJ20170413173837121), Innovation and Technology Commission (ITCPD/17-9), the Hong Kong Research Grant Council (HKUST C6009-15G, AoE/P-705/16, T13-605/18-W) to X.H; GRF 16301319 to P.P.C; National Institute of Health GM25232 to J.L. NSFC (31922088); Hong Kong ITC

(ITCPD/17-9, ITS/480/18FP); Hong Kong RGC (26102719, N_HKUST606/17, C6002-17GF, C7065-18GF, R4017-18) to J.W. B.J. was supported by the Institute of Advanced Studies (IAS) postdoctoral fellowship.

This protocol was derived from Cheung *et al.* (2020)

Competing interests

All authors declare no conflicts of interests.

References

1. Acevedo, A. and Andino, R. (2014). Library preparation for highly accurate population sequencing of RNA viruses. *Nat Protoc* 9(7): 1760-1769.
2. Burns, J. A., Dreij, K., Cartularo, L. and Scicchitano, D. A. (2010). O6-methylguanine induces altered proteins at the level of transcription in human cells. *Nucleic Acids Res* 38(22): 8178-8187.
3. Carey, L. B. (2015). RNA polymerase errors cause splicing defects and can be regulated by differential expression of RNA polymerase subunits. *Elife* 4: e09954.
4. Cheung, P. P., Jiang, B., Booth, G. T., Chong, T. H., Unarta, I. C., Wang, Y., Suarez, G. D., Wang, J., Lis, J. T. and Huang, X. (2020). Identifying Transcription Error-Enriched Genomic Loci Using Nuclear Run-on Circular-Sequencing Coupled with Background Error Modeling. *J Mol Biol* 432(13): 3933-3949.
5. de Mercoyrol, L., Corda, Y., Job, C. and Job, D. (1992). Accuracy of wheat-germ RNA polymerase II. General enzymatic properties and effect of template conformational transition from right-handed B-DNA to left-handed Z-DNA. *Eur J Biochem* 206(1): 49-58.
6. Imashimizu, M., Oshima, T., Lubkowska, L. and Kashlev, M. (2013). Direct assessment of transcription fidelity by high-resolution RNA sequencing. *Nucleic Acids Res* 41(19): 9090-9104.
7. Islam, S., Kjallquist, U., Moliner, A., Zajac, P., Fan, J. B., Lonnerberg, P. and Linnarsson, S. (2011). Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res* 21(7): 1160-1167.
8. Ji, J. P. and Loeb, L. A. (1992). Fidelity of HIV-1 reverse transcriptase copying RNA *in vitro*. *Biochemistry* 31(4): 954-958.
9. Mahat, D. B., Kwak, H., Booth, G. T., Jonkers, I. H., Danko, C. G., Patel, R. K., Waters, C. T., Munson, K., Core, L. J., and Lis, J. T. (2016). Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). *Nat Protoc* 11: 1455-1476.
10. Minoche, A. E., Dohm, J. C. and Himmelbauer, H. (2011). Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. *Genome Biol* 12(11): R112.

Please cite this article as: Wang *et. al.*, (2021). EmPC-seq: Accurate RNA-sequencing and Bioinformatics Platform to Map RNA Polymerases and Remove Background Error,*Bio-protocol* 11 (4): e3921. DOI: 10.21769/BioProtoc.3921.

11. Pelechano, V., Chavez, S. and Perez-Ortin, J. E. (2010). A complete set of nascent transcription rates for yeast genes. *PLoS One* 5(11): e15442.
12. Saxowsky, T. T., Meadows, K. L., Klungland, A. and Doetsch, P. W. (2008). 8-Oxoguanine-mediated transcriptional mutagenesis causes Ras activation in mammalian cells. *Proc Natl Acad Sci U S A* 105(48): 18877-18882.
13. Schwanhausser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W. and Selbach, M. (2011). Global quantification of mammalian gene expression control. *Nature* 473(7347): 337-342.
14. van Leeuwen, F. W., de Kleijn, D. P., van den Hurk, H. H., Neubauer, A., Sonnemans, M. A., Sluijs, J. A., Koycu, S., Ramdjielal, R. D., Salehi, A., Martens, G. J., Grosveld, F. G., Peter, J., Burbach, H. and Hol, E. M. (1998). Frameshift mutants of beta amyloid precursor protein and ubiquitin-B in Alzheimer's and Down patients. *Science* 279(5348): 242-247.