**LOTTE-seq (Long hairpin oligonucleotide based tRNA high-throughput sequencing): S****pecific selection of tRNAs with 3’-CCA end for high-throughput sequencing**

Lieselotte Erbera\*, Anne Hoffmannb\*, Jörg Fallmannb, Heike Betata, Peter F. Stadlerb,c,d,e,f,g, Mario Mörla

Contact:

aInstitute for Biochemistry, Leipzig University, Leipzig, Germany, bBioinformatics Group, Department of Computer Science and Interdisciplinary Center for Bioinformatics, Leipzig University, Leipzig, Germany, cGerman Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Competence Center for Scalable Data Services and Solutions, and Leipzig Research Center for Civilization Diseases, Leipzig University, Leipzig, Germany, dMax Planck Institute for Mathematics in the Sciences, Leipzig, Germany, eFacultad de Ciencias, Universidad Nacional de Colombia, Sede Botoga, Colombia, fSanta Fe Institute, Santa Fe, USA, gDepartment of Theoretical Chemistry of the University of Vienna, Vienna, Austria

\*To whom correspondence should be addressed. Tel.: +49(0)341 97 36 911; E-mail: mario.moerl@uni-leipzig.de;

\*These authors contributed equally to this work

**SUPPLEMENTARY MATERIAL**

**MATERIALS AND Methods**

**Bioinformatics analysis pipeline**

To separate the sequenced reads into individual FAstQ files, samples were demultiplexed using Illumina’s bcl2fastq Conversion Software (v2.20, https://support.illumina.com/downloads/bcl2fastq-conversion-software-v2-20.html), allowing 0 mismatches in the barcode sequences. Adapters of raw reads were trimmed using Cutadapt v1.16 1, retaining only reads which surpass a quality cutoff of 25, a maximum allowed error rate of 0.15 and a read length from 8 to 95 nts after trimming of adapter and low quality bases. For standard pre- and post-trimming quality control, FASTQC v0.11.4 2 was applied.

Genomes of *D. discoideum* (dicty 2.7, assembly GCA\_000004695), *G. stearothermophilus* (strain ATCC 12980, assembly GCA\_001277805.1), *E. coli* (strain K-12 substr. MG1655, assembly GCA\_000005845.2), *S. oleracea* (assembly GCA\_002007265.1, KY768855.1), *S. cerevisiae* (strain BY4741, GCA\_000766575.2) and *H.  sapiens* (hg38, GCA\_000001405.27) were downloaded from NCBI (Database resources of the National Center for Biotechnology Information 2016).

Nuclear, pseudo and mitochondrial tRNAs were annotated with tRNAscan-SE v2.0 3 for each genome using the default parameter settings (Table 1). For *S. cerevisiae*, only 16 of 24 mt-tRNAs could be annotated via tRNAscan-SE. The missing mt-tRNAs were added from the YeastMine database 4.

The analysis was prepared on the basis of the best practice workflow for accurate mapping of tRNA reads 5. All annotated tRNAs were masked in the genome using BEDtools v2.27.1 6. To reduce an artificial increase in multiple mapping reads, sequence regions identical to the annotated tRNAs were also masked. To simulate the 5’-leader sequence and reduce softclipping artefacts, tRNA sequences were elongated with a 30 nt long 5’-genomic flanking region. Further, introns were removed and a 3’ CCACCA tail appended. The modified tRNA library was added as extra ‘chromosomes’ to the masked genome. The customized genome is referred to as the artificial genome.

Reads were mapped to the artificial genomes using Segemehl v0.2.0-418 7. Due to the similarity of some short tRNA reads to, among others, repetitive regions in the genome, we considered an excessively high value of 100,000 mappings per seed. We allowed 3 mismatches in seed regions, requesting a minimal accuracy of 80% and increased the e-value of seed extensions to 500 to obtain highly modified tRNA reads.

To analyze the specificity of the different RNA-Seq methods, the number of the unique tRNA mapped reads showing a 3’-CCA, -CC, -C as well as the absence of a 3’-CCA end were counted. Multiply mapped reads were counted as fraction of their number of hits or filtered to obtain uniquely mapped counts.

To compare our LOTTE-seq data with a procedure highly similar to the Pang approach we analyzed RNA-seq data of *G. stearothermophilus* (GSM2617017: SRR5535061, SRR5535061, SRR5535063 8 following the workflow described above.

**SUPPLEMENTARY FIGURES**

****

**Figure S1. Ligation of the UMI-containing hairpin adapter to tRNAs.** **(A)** The UMI-containing version of the DNA hairpin adapter. It carries a UMI sequence (cyan), indicated by the N stretch. Opposite of the UMI, a stretch of A residues is inserted to reduce the possibility of misfolding of the adapter. Such UMI sequences allow for the identification of amplification biases. **(B)** The adapter with a UMI region of 10 randomized positions (N10) was tested for hybridization and ligation to the tRNA 3’-CCA-end. **(C)** In the ligation reaction, the UMI adapter shows a highly selective and efficient ligation to radioactively labeled tRNA transcript with complete CCA-end, comparable to that of the original DNA hairpin adapter. Hence, UMI sequences can be easily implemented into LOTTE-Seq, without loss of efficiency or accuracy. The introduction of a UMI sequence in the 3’-adapter allows to detect also biases in the ligation or reverse transcription reaction.



**Figure S2. Sensitivity of library preparation using the CCA-specific DNA hairpin adapter.** The DNA hairpin adapter was fused to varying amounts of *E. coli* total RNA, ranging from 2µg to 50 ng. After cDNA synthesis, ligation of the second adapter and PCR amplification, reaction products were separated by agarose gel electrophoresis and stained with ethidium bromide. 250 ng of total RNA preparation are sufficient for the amplification of tRNA-derived cDNAs. The size range of 160 to 200 bp indicates the amplification of full-length as well as partial cDNAs resulting from premature RT stops. This indicates that the hairpin adapter ligation used in LOTTE-Seq represents a highly sensitive approach for the analysis of tRNA pools in a total RNA preparation, without prior isolation of a tRNA fraction.

****

**Figure S3. Read termination positions identified by LOTTE-Seq.** The read termination positions for the tested tRNA pools are an indication for modified positions in the tRNA sequences. In the eukaryotic samples, the main read termination occurs at position 58, which is usually occupied by a highly conserved methylated adenosine (m1A). Read terminations at positions 9 and 22 of eukaryotic tRNAs correspond to further conserved m1A positions. G26 (eukaryotes) and G46 (*E. coli*) might represent methylated guanosine residues that are frequently found at these positions. Additional but unassigned terminations are found at position 37 (which is also very frequently modified) and 47.



**Figure S4. Comparison of LOTTE-seq, optimized TruSeq sRNA protocol and standard Illumina TruSeq sRNA procedure. (A)** Two replicates of MiSeq-based sequence analyses of HEK293T, *S. oleracea*, *S. cerevisiae*, *D. discoideum*, *E. coli* and *G. stearothermophilus* tRNA pools were investigated. The percentage of tRNAs with a 3’-CCA-end (blue), with 3’-ends other than a CCA (orange) and non-tRNA reads (grey) are depicted. In all investigations, LOTTE-seq shows the highest content of tRNAs with CCA-end, ranging from 90% (*S. oleracea*) to 100% (*D. discoideum*). For the optimized TruSeq sRNA procedure, tRNAs with 3’-CCA-end range from 41% (*S. cerevisiae*) to 80% (*S. oleracea*). The standard Illumina TruSeq sRNA procedure showed the lowest tRNA content with only about 1 – 6% of the reads corresponding to tRNAs with CCA-end and 0.5 – 10% of the reads representing tRNAs with 3’-ends other than CCA. Hence, LOTTE-seq is superior to the other methods for the characterization of mature tRNAs.

**SUPPLEMENTARY TABLES:**

**Table S1: Numbers of annotated tRNAs.** Numbers of annotated nuclear, mitochondrial and pseudo tRNAs for each genome used within this study.

|  |  |  |  |
| --- | --- | --- | --- |
| Species | # nuclear tRNAs | # pseudo tRNAs | # mitochondrial tRNAs |
| *D. discoideum* | 403 | 4 | 18 |
| *E. coli* | 88 | 1 | - |
| *G. stearothermophilus* | 62 | 1 | - |
| *H. sapiens* | 732 | 98 | 22 |
| *S. cerevisiae* | 189 | 1 | 18 |
| *S. oleracea* | 2111 | 450 | 24 |

References

1. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet j. 2011;17:10. doi:10.14806/ej.17.1.200.

2. Andrews S. FastQC: a quality control tool for high throughput sequence data.: https://www.bioinformatics.babraham.ac.uk/ projects/fastqc/. Accessed: 2016-09-10. 2010.

3. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucl Acids Res 1997;25:955–64.

4. Balakrishnan R, Park J, Karra K, Hitz BC, Binkley G, Hong EL, Sullivan J, Micklem G, Cherry JM. YeastMine--an integrated data warehouse for Saccharomyces cerevisiae data as a multipurpose tool-kit. Database (Oxford) 2012;2012:bar062. doi:10.1093/database/bar062.

5. Hoffmann A, Fallmann J, Vilardo E, Mörl M, Stadler PF, Amman F. Accurate mapping of tRNA reads. Bioinformatics 2018;34:1116–24. doi:10.1093/bioinformatics/btx756.

6. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 2010;26:841–2. doi:10.1093/bioinformatics/btq033.

7. Hoffmann S, Otto C, Kurtz S, Sharma CM, Khaitovich P, Vogel J, Stadler PF, Hackermüller J. Fast mapping of short sequences with mismatches, insertions and deletions using index structures. PLoS Comput Biol 2009;5:e1000502. doi:10.1371/journal.pcbi.1000502.

8. Ernst FGM, Erber L, Sammler J, Jühling F, Betat H, Mörl M. Cold adaptation of tRNA nucleotidyltransferases: A tradeoff in activity, stability and fidelity. RNA Biol 2018;15:144–55. doi:10.1080/15476286.2017.1391445.