READemption – A tool for the computational analysis of deep-sequencing-based transcriptome data

Konrad U. Förstner¹,², Jörg Vogel¹ and Cynthia M. Sharma²*

¹Institute for Molecular Infection Biology, University of Würzburg, Würzburg
²Research Centre for Infectious Diseases (ZINF), University of Würzburg, Würzburg

ABSTRACT
Summary: RNA-Seq has become a potent and widely used method to qualitatively and quantitatively study transcriptomes. In order to draw biological conclusions based on RNA-Seq data, several steps some of which are computationally intensive, have to be taken. Our READemption pipeline takes care of these individual tasks and integrates them into an easy-to-use tool with a command line interface. To leverage the full power of modern computers, most subcommands of READemption offer parallel data processing. While READemption was mainly developed for the analysis of bacterial primary transcriptomes, we have successfully applied it to analyze RNA-Seq reads from other sample types, including whole transcriptomes, RNA immunoprecipitated with proteins, not only from bacteria, but also from eukaryotes and archaea.

Availability and Implementation: READemption is implemented in Python and is published under the ISC open source license. The tool and documentation is hosted at http://pythonhosted.org/READemption (DOI: 10.6084/m9.figshare.977849).

Contact: cynthia.sharma@uni-wuerzburg.de; konrad.foerstner@uni-wuerzburg.de

1 INTRODUCTION

RNA-Seq, the examination of cDNA by massively parallel sequencing technologies, is a potent way to perform transcriptome analyses at single-nucleotide-resolution and with a high dynamic range (Wang et al., 2009). It has been successfully used to annotate transcript boundaries and to identify novel transcripts such as small regulatory RNAs in both pro- and eukaryotes (Filiatrault, 2011; transcript boundaries and to identify novel transcripts such as small regulatory RNAs in both pro- and eukaryotes (Filiatrault, 2011; Ozsolak and Milos, 2011). Most prominently, it can be applied to draw biological conclusions based on RNA-Seq data, several steps some of which are computationally intensive, have to be taken. Our READemption pipeline takes care of these individual tasks and integrates them into an easy-to-use tool with a command line interface. To leverage the full power of modern computers, most subcommands of READemption offer parallel data processing. While READemption was mainly developed for the analysis of bacterial primary transcriptomes, we have successfully applied it to analyze RNA-Seq reads from other sample types, including whole transcriptomes, RNA immunoprecipitated with proteins, not only from bacteria, but also from eukaryotes and archaea.

Availability and Implementation: READemption is implemented in Python and is published under the ISC open source license. The tool and documentation is hosted at http://pythonhosted.org/READemption (DOI: 10.6084/m9.figshare.977849).

Contact: cynthia.sharma@uni-wuerzburg.de; konrad.foerstner@uni-wuerzburg.de

1 INTRODUCTION

RNA-Seq, the examination of cDNA by massively parallel sequencing technologies, is a potent way to perform transcriptome analyses at single-nucleotide-resolution and with a high dynamic range (Wang et al., 2009). It has been successfully used to annotate transcript boundaries and to identify novel transcripts such as small regulatory RNAs in both pro- and eukaryotes (Filiatrault, 2011; Ozsolak and Milos, 2011). Most prominently, it can be applied to draw biological conclusions based on RNA-Seq data, several steps some of which are computationally intensive, have to be taken. Our READemption pipeline takes care of these individual tasks and integrates them into an easy-to-use tool with a command line interface. To leverage the full power of modern computers, most subcommands of READemption offer parallel data processing. While READemption was mainly developed for the analysis of bacterial primary transcriptomes, we have successfully applied it to analyze RNA-Seq reads from other sample types, including whole transcriptomes, RNA immunoprecipitated with proteins, not only from bacteria, but also from eukaryotes and archaea.

Availability and Implementation: READemption is implemented in Python and is published under the ISC open source license. The tool and documentation is hosted at http://pythonhosted.org/READemption (DOI: 10.6084/m9.figshare.977849).

Contact: cynthia.sharma@uni-wuerzburg.de; konrad.foerstner@uni-wuerzburg.de

1 INTRODUCTION

RNA-Seq, the examination of cDNA by massively parallel sequencing technologies, is a potent way to perform transcriptome analyses at single-nucleotide-resolution and with a high dynamic range (Wang et al., 2009). It has been successfully used to annotate transcript boundaries and to identify novel transcripts such as small regulatory RNAs in both pro- and eukaryotes (Filiatrault, 2011; Ozsolak and Milos, 2011). Most prominently, it can be applied to draw biological conclusions based on RNA-Seq data, several steps some of which are computationally intensive, have to be taken. Our READemption pipeline takes care of these individual tasks and integrates them into an easy-to-use tool with a command line interface. To leverage the full power of modern computers, most subcommands of READemption offer parallel data processing. While READemption was mainly developed for the analysis of bacterial primary transcriptomes, we have successfully applied it to analyze RNA-Seq reads from other sample types, including whole transcriptomes, RNA immunoprecipitated with proteins, not only from bacteria, but also from eukaryotes and archaea.

Availability and Implementation: READemption is implemented in Python and is published under the ISC open source license. The tool and documentation is hosted at http://pythonhosted.org/READemption (DOI: 10.6084/m9.figshare.977849).

Contact: cynthia.sharma@uni-wuerzburg.de; konrad.foerstner@uni-wuerzburg.de

1 INTRODUCTION

RNA-Seq, the examination of cDNA by massively parallel sequencing technologies, is a potent way to perform transcriptome analyses at single-nucleotide-resolution and with a high dynamic range (Wang et al., 2009). It has been successfully used to annotate transcript boundaries and to identify novel transcripts such as small regulatory RNAs in both pro- and eukaryotes (Filiatrault, 2011; Ozsolak and Milos, 2011). Most prominently, it can be applied to draw biological conclusions based on RNA-Seq data, several steps some of which are computationally intensive, have to be taken. Our READemption pipeline takes care of these individual tasks and integrates them into an easy-to-use tool with a command line interface. To leverage the full power of modern computers, most subcommands of READemption offer parallel data processing. While READemption was mainly developed for the analysis of bacterial primary transcriptomes, we have successfully applied it to analyze RNA-Seq reads from other sample types, including whole transcriptomes, RNA immunoprecipitated with proteins, not only from bacteria, but also from eukaryotes and archaea.

Availability and Implementation: READemption is implemented in Python and is published under the ISC open source license. The tool and documentation is hosted at http://pythonhosted.org/READemption (DOI: 10.6084/m9.figshare.977849).

Contact: cynthia.sharma@uni-wuerzburg.de; konrad.foerstner@uni-wuerzburg.de

1 INTRODUCTION

RNA-Seq, the examination of cDNA by massively parallel sequencing technologies, is a potent way to perform transcriptome analyses at single-nucleotide-resolution and with a high dynamic range (Wang et al., 2009). It has been successfully used to annotate transcript boundaries and to identify novel transcripts such as small regulatory RNAs in both pro- and eukaryotes (Filiatrault, 2011; Ozsolak and Milos, 2011). Most prominently, it can be applied to draw biological conclusions based on RNA-Seq data, several steps some of which are computationally intensive, have to be taken. Our READemption pipeline takes care of these individual tasks and integrates them into an easy-to-use tool with a command line interface. To leverage the full power of modern computers, most subcommands of READemption offer parallel data processing. While READemption was mainly developed for the analysis of bacterial primary transcriptomes, we have successfully applied it to analyze RNA-Seq reads from other sample types, including whole transcriptomes, RNA immunoprecipitated with proteins, not only from bacteria, but also from eukaryotes and archaea.

Availability and Implementation: READemption is implemented in Python and is published under the ISC open source license. The tool and documentation is hosted at http://pythonhosted.org/READemption (DOI: 10.6084/m9.figshare.977849).

Contact: cynthia.sharma@uni-wuerzburg.de; konrad.foerstner@uni-wuerzburg.de
followed by the conversion of the resulting SAM alignment files into BAM files and the generation of mapping statistics.

Coverage calculation: Based on the alignments provided in the BAM files, cDNA coverage files can be generated using the subcommand coverage. It creates several wiggle files that are based on different normalization methods like total number of aligned reads and represent the nucleotide-wise cDNA coverage in a strand specific manner. In order to visually inspect the libraries, these wiggle files can be loaded into common genome browsers.

Gene expression quantification: The read alignments can also be further used by the subcommand gene_quant to calculate gene-wise read counts. For this purpose, annotation files including gene positions in GFF3 format have to be provided. Besides raw gene-wise read countings normalized values – by total number of aligned reads as well as RPKM (Mortazavi et al., 2008) – are returned.

Differential gene expression analysis: For pairwise expression comparison, the subcommand deseq offers statistical analysis based on the approach implemented in DESeq2 (Anders and Huber, 2010) which builds upon the raw read counting. The results of DESeq2 are reformated and supplemented with gene annotations.

Plotting: The final three subcommands called viz_align, viz_genequant and viz_deseq generate several visualizations that help to interpret the result of the subcommand align, gene_quant and deseq, respectively.

READemption requires Python 3.2 or higher (http://python.org) and the libraries matplotlib (Hunter, 2007) and numpy (Oliphant, 2007) as well as the samtools (Li et al., 2009) wrapper pysam (http://pypi.python.org/pypi/pysam/) are needed. The subcommand deseq relies on an R (http://cran.r-project.org) installation and the bioconductor package DESeq2 (Anders and Huber, 2010). Instructions how to install READemption as well as how to execute its subcommands including examples can be found in the documentation.

3 CONCLUSIONS

We present an open source pipeline for the analysis of RNA-Seq data from all domains of life. READemption generates several output files that can be examined with common office suites, graphic programs and genome browsers. Its features make it a useful tool for anybody interested in the computational analysis of RNA-Seq data with the required basic command line skills.

ACKNOWLEDGEMENT

We thank members of the Sharma and the Vogel groups especially Thorsten Bischler for testing and constructive feedback.

Funding: Work in the Sharma laboratory is supported by the ZINF Young Investigator program of the Research Center for Infectious Diseases (ZINF) at the University of Würzburg and the Bavarian BioSysNet program. CMS is supported by the Young Academy program of the Bavarian Academy of Sciences.

REFERENCES


Fig. 1. A) Data and work flow of READemption including the input, output and the performed steps. B) Examples of plots generated by READemption.