Bioinformatics, 33(10), 2017, 1563–1564
doi: 10.1093/bioinformatics/btx002
Advance Access Publication Date: 2 February 2017
Applications Note

OXFORD

Genome analysis

MetaPlotR: a Perl/R pipeline for plotting metagenes of nucleotide modifications and other transcriptomic sites

Anthony O. Olarerin-George and Samie R. Jaffrey*

Department of Pharmacology, Weill Medical College, Cornell University, New York, NY, USA

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on September 16, 2016; revised on December 12, 2016; editorial decision on December 30, 2016; accepted on January 5, 2017

Abstract

Summary: An increasing number of studies are mapping protein binding and nucleotide modifications sites throughout the transcriptome. Often, these sites cluster in certain regions of the transcript, giving clues to their function. Hence, it is informative to summarize where in the transcript these sites occur. A metagene is a simple and effective tool for visualizing the distribution of sites along a simplified transcript model. In this work, we introduce MetaPlotR, a Perl/R pipeline for creating metagene plots.

Availability and Implementation: The code and associated tutorial are available at https://github.com/olarerin/metaPlotR.

Contact: srj2003@med.cornell.edu

1 Introduction

With the advent of high-throughput sequencing there has been an explosion of studies mapping protein binding sites and RNA modification sites transcriptome-wide. For example, the ENCODE project has mapped RNA binding sites for over 130 proteins in different cell lines (Van Nostrand *et al.*, 2016). Also, several groups have mapped various modified nucleotides in mRNA (e.g. N6-methyladenosine, N6, 2′-O-dimethyladenosine, N1-methyladenosine, 5-methylcytidine and pseudouridine) at single nucleotide-resolution (Carlile *et al.*, 2014; Dominissini *et al.*, 2016; Ke *et al.*, 2015; Li *et al.*, 2016; Linder *et al.*, 2015; Lovejoy *et al.*, 2014; Squires *et al.*, 2012).

A fundamental question raised by these mapping projects is where in the transcript do sites preferentially occur. This is important because the location of these sites, for example RNA modifications, can provide clues as to their biogenesis and function. A useful tool for visualizing this is the metagene. A metagene is a frequency plot of sites along a transcript model comprised of a 5′UTR, coding sequence and 3′UTR. Figure 1 shows an example of a metagene for N6-methyladenosine (m6A) generated by MetaPlotR showing the characteristic stop codon-proximal peak (Dominissini *et al.*, 2012; Meyer *et al.*, 2012).

While the concept of a metagene is straightforward, its implementation is not trivial. A typical workflow entails: (1) converting site coordinates from genomic to transcriptomic space; (2) identifying the region of the transcript in which sites occur; (3) projecting these sites onto a virtual transcript space (metagene coordinates) and (4) plotting the metagene coordinates as a histogram or density plot. In this work, we developed MetaPlotR, a Perl and R pipeline, to easily generate metagenes for any organism for which a genome and transcript annotation is available through the UCSC genome browser database.

2 Implementation

The MetaPlotR pipeline consists of four Perl scripts for data preprocessing and one R script for plotting the metagene.

The first Perl script in the pipeline, **make_annot_bed.pl**, creates a *master annotation file* (BED format; see https://genome.ucsc.edu/FAQ/FAQformat.html#format1 for description of format) of every nucleotide in a given transcriptome. The script is supplied with the locations of a directory containing the genome of interest and a gene prediction file containing the genomic coordinates for all genes and

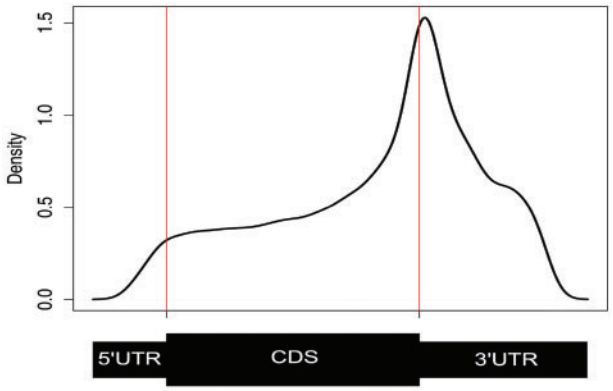

Fig. 1. Metagene of m6A sites. This metagene of N-6 methyladenosine sites (m6A) was generated with MetaPlotR. The metagene shows the characteristic peak in RNA methylation around the CDS end/3′UTR start. The metagene also highlights an optional feature of MetaPlotR—re-scaling of the 5′UTR, CDS and 3′UTR plotted lengths to reflect the average size of these regions in the queried dataset

their various features/regions. Both the genome and gene prediction file can be downloaded from the UCSC genome browser (Kent *et al.*, 2002). We provide examples of how to obtain these files in the manual.

The second Perl script, **size_of_cds_utrs.pl**, creates a file cataloging the transcriptomic coordinates of the start and end sites of the transcript regions (i.e. 5′UTR, CDS and 3′UTR). It takes the master annotation file as input and outputs a region annotation file. This file is necessary for determining the distance of queried sites from the transcriptomic features (i.e. transcriptional start site, start codon, stop codon and transcript end).

The next Perl script, **annotate_bed_file.pl**, annotates the user supplied BED file (containing single nucleotide genomic coordinates of sites of interest). It serves as a wrapper for (i.e. convenient way to run) Bedtools Intersect (Quinlan and Hall, 2010) and essentially labels every line in the user supplied BED file with the matching line (i.e. same coordinates) in the master annotation file. The resulting file is called the annotated query file.

The last Perl script, **rel_and_abs_dist_calc.pl** identifies the region of the transcript in which the user supplied sites fall and converts the transcriptomic coordinates to metagene coordinates. In other words, it projects the transcriptomic coordinates onto a representative and simplified transcript model. Sites that occur in the 5′UTR have a value from 0 to 1, where 0 and 1 represent the 5′ and 3′ ends of the 5′UTR, respectively. Similarly, sites in the CDS have a value from 1 to 2 and the 3′UTR 2 to 3. This script takes as input the annotated query file and the region annotation file. The outputted file generated from this step contains the metagene coordinates necessary to generate the metagene plot.

Finally, to visualize the metagene, we provide a tutorial for rendering the images in R. The R script takes the metagene coordinates produced by rel_and_abs_dist_calc.pl, extracts the columns of interest and plots absolute or relative distance metagenes with various optional features as demonstrated in the tutorial/manual.

3 Concluding remarks

Here we provide a simple pipeline to create metagene plots. The only requirement is that the appropriate genome and gene prediction file exists in the UCSC genome browser database. The goal of this pipeline was to make it as simple as possible for biologist with little to no bioinformatics experience to readily generate metagene plots of nucleotide modifications, protein binding sites or any other transcriptomic sites.

Funding

This research was supported by NIH grant R01DA037755 (SRJ) and in part by National Cancer Institute (NCI) Grant NIH T32 CA062948 (AO). Anthony Olarerin-George, Ph.D., holds a Postdoctoral Enrichment Program Award from the Burroughs Wellcome Fund.

Conflict of Interest: none declared.

References

Carlile,T.M. *et al.* (2014) Pseudouridine profiling reveals regulated mRNA pseudouridylation in yeast and human cells. *Nature*, **515**, 143–146.

Dominissini,D. *et al.* (2016) The dynamic N1-methyladenosine methylome in eukaryotic messenger RNA. *Nature*, **530**, 441–446.

Dominissini,D. *et al.* (2012) Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature*, **485**, 201–206.

Kent,W.J. *et al.* (2002) The human genome browser at UCSC. *Genome Res*, **12**, 996–1006.

Ke,S. *et al.* (2015) A majority of m 6 A residues are in the last exons, allowing the potential for 3′ UTR regulation. *Genes Dev.*, **29**, 2037–2053.

Linder,B. *et al.* (2015) Single-nucleotide-resolution mapping of m6A and m6Am throughout the transcriptome. *Nat. Methods*, **12**, 767–772.

Li,X. *et al.* (2016) Transcriptome-wide mapping reveals reversible and dynamic N1-methyladenosine methylome. *Nat. Chem. Biol.*, **12**, 311–316.

Lovejoy,A.F. *et al.* (2014) Transcriptome-wide mapping of pseudouridines: pseudouridine synthases modify specific mRNAs in *S. cerevisiae*. *PLoS One*, **9**, e110799.

Meyer,K.D. *et al.* (2012) Comprehensive analysis of mRNA methylation reveals enrichment in 3' UTRs and near stop codons. *Cell*, **149**, 1635–1646.

Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.

Squires,J.E. *et al.* (2012) Widespread occurrence of 5-methylcytosine in human coding and non-coding RNA. *Nucleic Acids Res.*, **40**, 5023–5033.

Van Nostrand,E.L. *et al.* (2016) Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat. Methods*, **13**, 508–514.