Bioinformatics, 32(5), 2016, 782–785
doi: 10.1093/bioinformatics/btv647
Advance Access Publication Date: 4 November 2015
Applications Note

Sequence analysis

meRanTK: methylated RNA analysis ToolKit

Dietmar Rieder[1,*], Thomas Amort[2], Elisabeth Kugler[1], Alexandra Lusser[2] and Zlatko Trajanoski[1]

[1]Division of Bioinformatics and [2]Division of Molecular Biology, Biocenter, Medical University of Innsbruck, Innsbruck 6020, Austria

*To whom correspondence should be addressed.

Associate Editor: Ivo Hofacker

Received on February 18, 2015; revised on August 19, 2015; accepted on October 29, 2015

Abstract

Summary: The significance and function of posttranscriptional cytosine methylation in poly(A)RNA attracts great interest but is still poorly understood. High-throughput sequencing of RNA treated with bisulfite (RNA-BSseq) or subjected to enrichment techniques like Aza-IP or miCLIP enables transcriptome wide studies of this particular modification at single base pair resolution. However, to date, there are no specialized software tools available for the analysis of RNA-BSseq or Aza-IP data. Therefore, we developed meRanTK, the first publicly available tool kit which addresses the special demands of high-throughput RNA cytosine methylation data analysis. It provides fast and easy to use splice-aware bisulfite sequencing read mapping, comprehensive methylation calling and identification of differentially methylated cytosines by statistical analysis of single- and multi-replicate experiments. Application of meRanTK to RNA-BSseq or Aza-IP data produces accurate results in standard compliant formats.

Availability and Implementation: meRanTK, source code and test data are released under the GNU GPLv3+ license and are available at http://icbi.at/software/meRanTK/.

Contact: dietmar.rieder@i-med.ac.at

1 Introduction

Modifications of DNA by methylation and hydroxymethylation of the carbon 5 atom of cytosine (m^5C, hm^5C) are well-established epigenetic mechanisms that affect the activity state of genomic regions. In contrast to the relatively limited array of DNA modifications, RNA can be target to >100 chemically distinct modifications, roughly two-thirds of which are methylations (Cantara *et al.*, 2011). Most of these RNA modifications were studied on rRNAs and tRNAs, where m^5C plays a role in structural and metabolic stabilization (Motorin *et al.*, 2010). Surprisingly, little is known about the significance and function of m^5C in poly(A)RNAs. Recently, one of the first studies to globally evaluate RNA cytosine methylation showed that m^5C occurred throughout the human transcriptome in coding and non-coding RNAs (Squires *et al.*, 2012). Moreover, cytosine methylation in the long non-coding RNAs HOTAIR and XIST was found to have the ability to interfere with RNA–protein interaction (Amort *et al.*, 2013).

RNA m^5C modifications can be detected by bisulfite treatment which converts all unmethylated cytosines into uracils leaving methylated cytosines unaffected. An alternative technique is the enrichment of RNAs covalently bound to RNA-methyltransferases (RMT) that were trapped on their substrate either by the precence of 5-azacytidine (Aza-IP) (Khoddami and Cairns, 2013) or by mutating the catalytic center of the respective RMT (miCLIP) (Hussain *et al.*, 2013). When these methods are combined with high-throughput sequencing it is possible to obtain either global (RNA-BSseq) or enzyme-specific (Aza-IP, miCLIP) m^5C-modified transcriptomes. To date, numerous software tools are available for the analysis of DNA-BSseq data (Kunde-Ramamoorthy *et al.*, 2014), but to the best of our knowledge none of these tools supports the analysis of data from RNA-BSseq or Aza-IP, which causes a C to G conversion at the RMT-bound cytosine. Aligners and methylation callers specific for DNA methylation data are of limited use when dealing with RNA. The major shortcomings are due to the fact that transcripts may exist in multiple different splice forms and that they may be transcribed from either of the two DNA strands. At present, a few methylation studies using RNA-BSseq and Aza-IP have been published. However, the analysis strategies used in these works are time-consuming and cumbersome involving many steps utilizing many different—in part commercial—tools and on top they require the development of custom programs (Khoddami and Cairns, 2014; Schaefer, 2015). To address these issues and improve the RNA m^5C data analysis, we developed meRanTK, a tool kit that provides programs for fast mapping of high-throughput RNA-BSseq data and enables accurate identification of methylated and differentially methylated cytosines from BS-seq and Aza-IP experiments at single base pair resolution.

2 Software description and discussion

meRanTK includes five multithreaded programs which enable complete analysis and comparison of m5C transcriptome datasets.

The tools, meRanT and meRanG, use well-established RNAseq-specific short read mappers as core aligning engines and extend them to facilitate mapping of either single- or paired end sequence reads from strand-specific RNA-BSseq libraries to a given reference sequence. Input data may originate from any sequencing platform that produces FASTQ formatted files. The meRanCall methylation caller uses aligned reads to precisely identify and statistically evaluate the positions of methylated cytosines. The experiment comparison tool meRanCompare is designed to detect differentially methylated m^5Cs of two experimental conditions with single or multi-replicate RNA methylation datasets. At last, the annotation tool meRanAnnotate helps to annotate candidate m^5Cs with genomic features such as gene- or transcript names and positional metrics. All included tools were written in Perl and run therefore on a wide variety of computing platforms.

2.1 meRanT

meRanT is an RNA-BSseq alignment tool for mapping sequencing reads to a pre-assembled set of transcripts (e.g. tRNAs, ncRNAs), it first performs a full C→T conversion of RNA-BSseq reads and uses then Bowtie2 (Langmead and Salzberg, 2012) to align them to a C→T-converted reference transcriptome (e.g. RefSeq) that was generated with the indexing mode of meRanT. Since this reference can have multiple transcripts per gene, multi-mapping of individual reads is expected. We distinguish between two classes of multi-mapping reads (MMRs): "class I" MMRs align to transcripts of different genes or to multiple locations on a single transcript; "class II" MMRs align unambiguously to multiple transcripts of the same gene. meRanT is designed to utilize the read mapping coordinates together with a transcript-to-gene assignment to correctly classify MMRs. Only "class II" MMRs are considered for further analysis in which all its valid alignments are inspected. meRanT filters these alignments for correct orientation and then selects from the highest scoring candidates the one mapping to the longest transcript. Finally, after restoring the original read sequence the alignment is stored in the resulting SAM/BAM file. "Class I" MMRs can be stored in a separate SAM/BAM file. Our method to deal with MMRs can "save" many of otherwise ambiguous alignments by assigning them to a "canonical" longest mapping transcript. Thus, it typically increases the mapping efficiency by >50% for single end- and by >20% for the paired end reads.

2.2 meRanG

In contrast to meRanT which aligns bisulfite-reads to a pre-assembled set of transcripts, meRanG aligns RNA-BSseq reads to a bisulfite-converted genome using a splice aware short read mapper. We implemented two variants of meRanG using different aligners: the fast meRanGs uses STAR (Dobin *et al.*, 2013) and memory saving meRanGt uses TopHat2 (Kim *et al.*, 2013). Transcripts can originate from either strand (+/−) of the genome, therefore meRanG runs two parallel aligner instances to map reads to a C→T and to a G→A-converted genome, generated with the indexing mode of meRanG, and saves the highest scoring unique alignment with correct orientation. Bisulfite treatment of RNA causes strong fragmentation (Schaefer *et al.*, 2009). Thus, when using paired end sequencing, the two read pair mates—originating from fragments shorter than twice the read length—can overlap which leads to incorrect methylation rate estimations. To overcome this problem, meRanG can detect these overlaps and either soft- or hard-clip the 3′ ends, each by half of the overlap, so that the genomic region covered by the read pair is counted only once. Finally, meRanG stores uniquely mapped reads in a SAM or BAM file, whereas, MMRs are optionally written to a separate SAM/BAM file. In addition to the alignment files, meRanG can report the read coverage across the entire genome and store it as bedGraph file which may be used for further analysis or visual inspection with a genome browser tool.

meRanT and meRanG can produce M-bias plots which may help detecting potential sequencing or library problems. The fraction $N_{[C]}/N_{[ATG]}$ is plotted at each read position. In an unbiased dataset this plot should present a flat horizontal line since cytosine methylation is expected to occur independently of the read position.

2.3 meRanCall

meRanCall extracts the methylation state of individual cytosines from bisulfite-read alignments obtained with meRanT/meRanG or in its Aza-IP mode from BAM files obtained with conventional short read aligners such as STAR, Bowtie2 or novoalign (Novocraft, 2015). Methylated cytosines are called based on user-supplied minimum thresholds, such as read coverage, non-conversion rate and base quality. Potential PCR duplicates may be filtered by defining a maximum allowed number of identical reads, and potential C-biased read ends may be excluded from the analysis. If control sequences are included in the dataset, meRanCall can determine the overall C→T conversion rate for calculating the *P*-value of the methylation state (Lister *et al.*, 2009) and the *P*-value of the methylation rate (Barturen *et al.*, 2013). If no conversion rate is available, meRanCall uses a Fisher's exact test to calculate the *P*-value for a given candidate methylated cytosine based on the baseline sequencing error. Using these *P*-values a Benjamini-Hochberg correction for multiple hypothesis testing and filtering candidates at a user-defined FDR is performed. Besides the *P*-values meRanCall calculates coverage, C count, methylation rate, 95% confidence intervals, mutation rate and reports the position, strand, reference base and sequence context around a methylated cytosine. All data are stored in a simple tab delimited file and optionally a BED6+3 or narrow-peak BED file may be generated for projecting the methylation data onto a genome browser display.

The result files may then be fed into meRanAnnotate to assign genomic annotations and distance measurements to the individual candidate m^5Cs.

2.4 meRanCompare

meRanTK includes meRanCompare a tool that facilitates the identification of differentially methylated cytosines in two experimental conditions. Result files produced by running meRanCall on different experiments or, in the case of Aza-IP experiments, from IP and Control datasets, will be statistically analyzed and significant differences will be reported. meRanCompare can handle single or multiple replicate experiments using either a Fisher's exact test or a Cochran–Mantel–Haenszel test, respectively, to determine whether the methylation rates of candidate m^{5}Cs are significantly different between the two tested conditions. Fold-change cutoffs for IP enrichment (Aza-IP) or methylation rate (RNA-BSseq) and significance level as well as false discovery rate levels can be set at run time. Moreover, we provide a helper program that estimates differences in the library sizes and reports scaling factors (similar to the method described in DESeq2—Love *et al.*, 2014) which will be considered in the meRanCompare calculation.

2.5 Performance and application

We first applied meRanTK on simulated mouse (mm10) datasets consisting of 70 million 100 bp single- and 60 million paired end reads covering 48 301 and 51 875 m^{5}Cs, respectively (available at the meRanTK website). After stringent quality filtering, read mapping was performed on 12 2.4 GHz cores of a server type computer and m^{5}C candidates were called using meRanCall. Results in Table 1 demonstrate excellent mapping efficiency, high m^{5}C recall rates and a negligible number of false positive m^{5}C calls. We next compared our results with the results obtained by applying a previously proposed analysis strategy (Khoddami and Cairns, 2013). This involves the Useq-package for generating a reference transcriptome with all known and synthetic splice-sites derived from a refFlat annotation file. This set of known and synthetic transcripts is then further combined with a separately generated exon-masked genome, and the resulting sequences are used for generating a bisulfite index for the Novocraft novoalign short read aligner. This multi-step preprocessing and index generation takes about 12 h compared with 4 h required for the single-step index generation of meRanTK. After aligning the simulated RNA-BSseq reads to this reference using novoalign in the bisulfite mode, the alignments in the resulting SAM file have to be transformed to genomic coordinates using the SamTranscriptomeParser—another tool from the Useq package. This way it takes five instead of two steps (meRanTK) to obtain aligned RNA-BSseq reads. The metrics (Useq/novo) in Table 1 indicate that compared with meRanGs this strategy results in comparable alignment—and recall rates. It should be noted that methylation calling was performed with meRanCall in all tests shown (no methylation calling procedure or tool was published with the original RNA-BSseq analysis strategy). The differences are likely due to more stringent settings used in meRanG. However, running on the same hardware, the computation time required for the Useq/novo strategy was more than 100 times longer as compared with meRanGs and still more than 12 times longer as compared with meRanGt. This major improvement in runtime enables researchers to complete an analysis within hours instead of weeks.

Table 1. Benchmark results of meRanG and meRanCall on simulated data

Program	# Filtered	% u/m	Map time	% Recall	% FP	Call time
meRanGs	58 M se	90.2	2 h 09 min	87.80	0.05	7 h 00 min
meRanGs	87 M pe	88.2	3 h 37 min	87.20	0.20	10 h 04 min
meRanGt	58 M se	89.7	18 h 24 min	82.20	0.12	6 h 54 min
meRanGt	87 M pe	84.7	26 h 21 min	82.10	1.73	10 h 54 min
meRanT*	58 M se	93.2	8 h 42 min	87.88	1.84	4 h 42 min
meRanT*	87 M pe	92.3	16 h 06 min	86.49	1.42	4 h 36 min
Useq/novo	58 M se	91.2	233 h 51 min	87.50	0.33	7 h 12 min
Useq/novo	87 M pe	90.0	413 h 06 min	90.10	0.28	10 h 23 min

M = million filtered reads, se = single end reads, pe = paired end reads, u/m=unique mapped, recall = m^{5}Cs recovered from simulated data, FP false positive m^{5}Cs calls, * mm10 RefSeq w/o predictions

We also tested meRanTK on published RNA-BSseq data (Khoddami and Cairns, 2013). 78.23% of 32 M filtered reads could be aligned to the mm10 genome. A large fraction of reads mapped to rDNA repeats or tRNAs so that finally 48.9% mapped to unique genomic locations. Using meRanCall and meRanCompare, we correctly identified 38 out of 39 reported differentially methylated Cs (Supplementary Dataset 3 of Khoddami and Cairns, 2013). Moreover, we correctly missed those m^{5}Cs that were marked as potential false positives in the original paper and identified an additional 1037 previously not reported m^{5}C candidates. We further tested meRanG on additional real data (unpublished) and typically achieved mapping rates of around 92% (83% unique) for single-end and 82% (75% unique) for paired-end reads. The mapping rate for meRanT depends on the transcript database used and was typically around 74% when we used the mouse (mm10) RefSeq database.

To test meRanTK on Aza-IP data we used published dataset for the NSUN2 methyltransferase (Khoddami and Cairns, 2014). We mapped the reads from two replicate and one control experiment to the human genome (hg19) using the STAR aligner allowing for 10% mismatches over the read length and ran Bowtie2 in the "very-sensitive-local" mode to align reads that were not aligned in the first step. We then use meRanCall in its Aza-IP mode to directly call candidate methylated cytosines at the proposed cutoff of 4% C→G conversion and a false discovery rate of 0.01. Using meRanCompare we identified 1496 candidates at an FDR of 0.01, that were present in both replicates with a 3-fold enrichment and a significantly ($P < 0.01$) higher C→G conversion rate compared with the input control experiment. Eight hundred and sixty-six of them were located on tRNAs, 70 on rRNAs, four on ncRNAs and 264 on protein coding genes (292 w/o annotation). Moreover, 456 candidates from our analysis were overlapping with the set of 573 reported candidates from the original study.

3 Conclusion

Previously proposed analysis approaches for RNA cytosine methylation data consist of many time-consuming non-automated steps involving the usage of different tools and still demand writing of custom programs. meRanTK solves these issues and enables easy, time saving and accurate analysis of RNA-BSseq and Aza-IP experiments.

Funding

Tiroler Standortagentur (Project Bioinformatics Tyrol), Austrian Science Fund (FWF) P27024-BBL.

Conflict of Interest: none declared.

References

Amort,T. *et al.* (2013) Long non-coding RNAs as targets for cytosine methylation. *RNA Biol.*, **10**, 1003–1008.

Barturen,G. *et al*. (2013) MethylExtract: high-quality methylation maps and SNV calling from whole genome bisulfite sequencing data. *F1000Research*, **2**, 217.

Cantara,W.A. *et al*. (2011) The RNA modification database, RNAMDB: 2011 update. *Nucleic Acids Res*., **39**, D195–D201.

Dobin,A. *et al*. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.

Hussain,S. *et al*. (2013) NSun2-mediated cytosine-5 methylation of vault non-coding RNA determines its processing into regulatory small RNAs. *Cell Rep*., **4**, 255–261.

Khoddami,V. and Cairns,B.R. (2013) Identification of direct targets and modified bases of RNA cytosine methyltransferases. *Nat. Biotechnol*. **31**, 458–464.

Khoddami,V. and Cairns,B.R. (2014) Transcriptome-wide target profiling of RNA cytosine methyltransferases using the mechanism-based enrichment procedure Aza-IP. *Nat. Protoc*. **9**, 337–361.

Kim,D. *et al*. (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol*., **14**, R36.

Kunde-Ramamoorthy,G. *et al*. (2014) Comparison and quantitative verification of mapping algorithms for whole-genome bisulfite sequencing. *Nucleic Acids Res*. **42**, e43.

Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359.

Lister,R. *et al*. (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, 315–322.

Love,M.I. *et al*. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. **15**, 550.

Motorin,Y. *et al*. (2010) 5-methylcytosine in RNA: detection, enzymatic formation and biological functions. *Nucleic Acids Res*. **38**, 1415–1430.

Schaefer,M. (2015) Chapter Fourteen—RNA 5-methylcytosine analysis by bisulfite sequencing. In: He,C. (ed.) *Methods in Enzymology, RNA Modification*. Academic Press, Salt Lake City, NJ, pp. 297–329.

Schaefer,M. *et al*. (2009) RNA cytosine methylation analysis by bisulfite sequencing. *Nucleic Acids Res*., **37**, e12.

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