CLIFinder: Identification of LINE-1 Chimeric Transcripts in RNA-seq data

Marie-Elisa Pinson, Romain Pogorelcnik, Franck Court, Philippe Arnaud, Catherine Vaurs-Barrière

CLIFinder: Identification of LINE-1 Chimeric Transcripts in RNA-seq data

Marie-Elisa PINSON[1,†], Romain POGORELCNIK[1,†], Franck COURT[1], Philippe ARNAUD[1], Catherine VAURS-BARRIERE[1,*]

[1] GReD, Université Clermont Auvergne, CNRS, INSERM, BP 38, 63001 Clermont-Ferrand, France

† Both authors contribute equally to this work

*To whom correspondence should be addressed.

Abstract

L1 Chimeric Transcripts (LCTs) are initiated by repeated LINE-1 element antisense promoters and include the L1 5'UTR sequence in antisense orientation followed by the adjacent genomic region. LCTs have been characterized mainly using bioinformatics approaches to query dbEST. To take advantage of NGS data to unravel the transcriptome composition, we developed Chimeric LIne Finder (CLIFinder), a new bioinformatics tool. Using stranded paired-end RNA-seq data, we demonstrated that CLIFinder can identify genome-wide transcribed chimera sequences corresponding to potential LCTs. Moreover, CLIFinder can be adapted to study transcription from other repeat types.

Availability: The code is available at: https://github.com/GReD-Clermont/CLIFinder; and for Galaxy users, it is directly accessible in the tool shed at: https://toolshed.g2.bx.psu.edu/view/clifinder/clifinder/

Contact: Catherine.barriere@uca.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Transposable elements represent 45% of the human genome. Due to the challenge of sequence alignment, their role in the transcriptional landscape is poorly studied. However, it is known that LINE-1 retro-elements (L1s) can drive transcription of adjacent genomic regions. For instance, in the 5' untranslated region (5'UTR) of the most recent L1 subfamilies (from the oldest L1PA10 to the most recent L1PA1), an antisense promoter (ASP) has been identified (Speek, 2001; Macia et al., 2011). This ASP can produce L1 Chimeric Transcripts (LCTs) that include the L1 5'UTR sequence in antisense orientation followed by the adjacent genomic region.

LCTs have been characterized mainly using bioinformatics approaches to query dbEST and correspond to spliced ESTs where the L1 5'UTR sequence in antisense orientation is associated with a gene exon (Speek, 2001; Nigumann et al., 2002; Mätlik et al., 2006; Wolf et al., 2010; Criscione et al., 2016). In addition, Cruickshanks and Tufarelli (2009) developed a dedicated biomolecular approach (called LCD) that allowed the identification of 18 LCTs. Nevertheless, these studies remain limited by the a priori definition of such sequences (position of the L1 primer in LCD and selection of only spliced exon-containing ESTs). To take advantage of the strong potential of next generation sequencing to unravel the transcriptome composition, we developed Chimeric LINE Finder (CLIFinder), a new bioinformatics tool that identifies chimeras corresponding to potential LCTs in RNA-seq data.

2 Methods

CLIFinder is a Galaxy tool designed to identify chimeras from one or several samples. This tool was developed mainly to analyze stranded paired-end RNA-seq data. It can also analyze non-stranded paired-end RNA-seq data, but with a higher rate of false-positive chimeras, particularly from events that encompass L1 promoter sequences. Quality and adapter trimming must be done on Fastq files, before their use in CLIFinder.

Finding chimeras

The iterative procedure starts by finding chimeras for each sample (Supplementary Fig. S1). We consider here that Read 1 (R1, starting at the 5'-end) of the stranded pair must contain an L1 sequence. This parameter is customable according to the stranded or non-stranded sequencing and the repeat element considered. For stranded libraries:

- if the first read of the pair (R1) is in the sense of the transcript, R1 is considered to correspond to the transposable element part of the chimera (Fig S1)

- if the first read of the pair is in the opposite sense of the transcript, then the second read of the pair (R2) is considered to be the transposable element part of the chimera.

For non-stranded sequencing, both hypotheses are retained.

Step 1 consists in selecting pairs (R1 and R2) that include R1 with a minimum of X consecutive bp that correspond to repeat element sequences given by the user (for instance, 5'-end LINE-1 sub-families sequences) and an unmapped mate (R2). Only paired reads that respect this condition are retained. Mismatches (Y) can be tolerated or not for L1 mapping. In step 2, these paired reads are filtered using the RepeatMasker tool to retain only reads for which at least Z consecutive bp are not detected as a repeated sequence. To identify chimeras (step 3), the retained filtered paired sequences are aligned to the human reference genome with a flexi-ble maximum insert size between paired reads (W).

Therefore, CLIFinder is a fully customizable tool to detect transcripts initiated by different types of repeat elements with X, Y, Z and W values that can be changed by the user to adapt the filter stringency. The parameter settings depend, notably, on the read length and the L1 reference sequence list given by the user in Step 1. In our hands, using one consensus sequence for each L1 element from the 27 sub-families (Khan et al., 2006) as reference for analyzing 100 to 150 bp paired-end reads, the optimal conditions to identify the largest number of chimeras was X = 30 bp, Y = 6 mismatches for Step 1. For Step 2 only the less stringent condition of Z = 30 bp was used to allow at least the design of a primer in the unique sequence for later validation experiments. Finally, W = 50,000 bp was the maximum insert size between paired reads that allowed the identification of the largest number of spliced chimeras.

Merging chimeras from all samples and concatenation

For each sample, the chimera datasets are stored in two bed files that correspond to each part of the paired-end read alignments. For an overview of all data and comparison of different datasets, CLIFinder concatenates the different files in two unique bed files, one that corresponds to the L1 part (R1) and the other to the unique sequence part (R2) (step 4). At this stage, several potential chimeras can overlap in the first or/and second read dataset. Bedtools merge function is used to group the different reads corresponding to the same locus (±100 bp) (step 5) to eliminate repetitive information that might correspond to the same chimera. This task is done for both files. The Getfasta function from Bedtools is then employed to create a unique fasta file (step 6) that contains all the potential chimeras concatenated from all samples to generate the html report.

Visualizing and downloading the resulting chimeras

CLIFinder results can be visualized on a Galaxy interface. An html table is generated where each line corresponds to a chimera defined by its genomic coordinates (hg19), transcription strand, and number of reads detected in each sample. BLASTn searches against EST and RNA databases are performed for chimera annotation, if data are already available. Users can also download: 1) a tabular file with the same content as the html page; 2) a final annotated file with additional information for each chimera: ID, gene name and strand (if it is localized in an intragenic region), and characteristics of the involved L1 element (family name, position, size, transcription strand) (step 7) (Supplementary Table S1) and 3) a CLIFinder execution report with the number of pairs retained after sample filtering.

CLIFinder was first tested on publicly available metastatic high grade serous ovarian cancer mRNA-seq data (Illumina HiSeq2500, 100bp stranded paired-end, 47 million reads) (Böhm et al., 2016). The original dataset GSM2122741 (Ov1.1) was used to create: 1) Ov1.2, an implemented dataset similar to Ov1.1, but including also the L1-MET LCT sequence (positive control) and a L1PA2 5'UTR antisense sequence (negative control) and 2) Ov1.3, a depleted dataset in which 30% of reads included in Ov1.1 were randomly eliminated and the L1-MET and L1PA2 5'UTR antisense control sequences were added. Two additional datasets that corresponded to mRNA-seq data for the MCF-7 (untreated breast adenocarcinoma) cell line (E-MTAB-3788, Illumina HiSeq2500, 150 bp stranded paired-end, mean 135 million reads) were also analyzed (Philippe et al., 2016). The parameters used for this analysis are described in Supplementary Figure S1. On average, each analysis required few hours depending on sequencing depth (i.e., less than 3 hours for the three Ov datasets and up to 8 hours for the two MCF-7 replicates) using a Quad-Core workstation with 16Go RAM (Supplementary Table S1).

3 Results

All results are available in Supplementary Table S2. From Ov1.1, CLIFinder identified 37 chimeras that corresponded to potential LCTs. For Ov1.2, 38 chimeras were obtained: the previous 37 chimeras and L1-MET LCT (ID_27). Ov1.3 analysis high-lighted only three chimeras: L1-MET and two other chimeras with a reduced number of reads. Some chimeras were associated with multiple L1s (ID_19). In this case, after visualization of the chimera position loaded from the html file on the genome browser UCSC, manual monitoring must be performed to retain the element presumed to initiate the chimera transcription. Attention must also be paid to remove the few chimeras transcribed in the same orientation as the L1 (ID_12_13). Consistent with the RNA-seq specificities, chimeras defined by only one read, displayed R1 and R2 sequences of 100bp. When a chimera is defined by several over-lapping reads, R1 and

R2 size can be increased. Negative (in the case of R1-R2 overlapping) or null distance between R1 and R2 was often observed, suggesting a linear transcription from L1 that continued in the adjacent unique sequence. In some cases, the distance between R1 and R2 was larger than expected (ID_7, ID_30), demonstrating the occurrence of splicing events (Supplementary Fig S2).

To assess whether chimeras detected by CLIFinder in Ov1.1 could correspond to LCTs initiated at L1 ASPs, the L1 elements involved in the 36 chimeras were compared to the list of 14,495 L1s containing a 5'UTR sequence (extracted from RepeatMasker for the 27 sub-families used by CLIFinder). This showed that 33 chimeras involved L1 elements with a 5'UTR. Then, analysis of the L1 sub-families involved in these 33 chimeras revealed that, although CLIFinder used the consensus sequences from the 27 L1 sub-families to identify chimeras, 31 chimeras (94%) involved L1s only from the more recent sub-families (i.e., L1PA1 to L1PA7) that possess ASP. Finally, comparison of the 33 chimeras and of known LCTs demonstrated that five chimeras identified in this metastatic ovarian cancer sample corresponded to EST-LCTs previously found in other tissues. Altogether, these observations suggest that most chimeras identified by CLIFinder from RNA-seq data are true LCTs.

To strengthen these results, two replicates from stranded paired-end untreated breast adenocarcinoma MCF-7 cells were also analyzed. After curation of the chimeras that did not meet the criteria for potential LCTs (i.e., chimeras in the same transcription sense of L1: n= 39; and chimeras associated with L1 but without 5'UTR ASP region: n=8), in total, 125 chimeras were identified in the two replicates (Supplementary Table S2C). Among these chimeras, 118 (94.5%) involved recent L1s and 20 (16%) corresponded to already described LCTs (Fig 1A). Comparison of the genomic position of the chimeras identified in the two MCF-7 replicates indicated that 28 chimeras (22%) were common to both samples (22%) (Fig 1A). This moderate overlap can be explained by the fact that chimeras can be detected only if at least one cDNA fragment encompassing the L1/unique sequence junction is sequenced. When only chimeras defined by at least 2 or 3 reads in one MCF-7 replicate were considered, the overlap between replicates increased to 52% (two reads) and 82% (three reads) (Fig. 1B). This indicates that results are reproducible for highly represented chimeras and that the analysis of different samples together allows assessing precisely LCT expression genome-wide in a specific condition.

Two to three times more chimeras were identified in the MCF-7 RNA-seq datasets than in the Ov1.1 dataset, in agreement with the higher sequencing depth (Supplementary Table S1). Interestingly, among the identified chimeras, nine were common between the MCF-7 RNA-seq

datasets and the Ov1.1 dataset (Fig. 1A). Among these chimeras, some corresponded to already known LCTs (n=4), such as Ov1.1 ID_7 and MCF-7 ID_19 (Supplementary Fig S2) that were previously found in thymus, retinoblastoma and tongue tumor samples. Others corresponded to new loci, such as Ov1.1 ID_30 that seems to be a good candidate because it was recurrently detected in MCF-7 Rep1 and Rep2 (chimeras defined by 2 and 4 reads, respectively) (Supplementary Fig S2, Supplementary Table S2C).

In summary, CLIFinder is a new bioinformatics tool to identify genome-wide transcribed chimera sequences corresponding to LCTs from stranded paired-end RNA-seq datasets. This tool will be useful for genome-wide analyses of LCT expression in different tissues, in normal or pathological conditions. Moreover, CLIFinder can be adapted to study transcription from other repeat types.

Acknowledgements

The authors acknowledge the support of Y Renaud from Byonet (byonet.fr).

Funding

This work was supported by the Plan Cancer-INSERM [CS14085CS, to P.A.], ARC [SFI20121205549, to C.V.B.], the Auvergne Region and the Fonds Européen de Développement Régional (FEDER).

Conflict of Interest: none declared.

References

Böhm, S. et al. (2016) Neoadjuvant Chemotherapy Modulates the Immune Micro-environment in Metastases of Tubo-Ovarian High-Grade Serous Carcinoma. Clin Cancer Res. 22, 3025-3036.

Criscione, S.W. (2016) Genome-wide characterization of human L1 antisense promoter-driven transcripts. BMC Genomics. 17, 463.

Cruickshanks, H.A. and Tufarelli, C. (2009) Isolation of cancer-specific chimeric transcripts induced by hypomethylation of the LINE-1 antisense promoter. Genomics. 94, 397-406.

Khan, H. (2010) Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. Genome Res. 16, 78-87.

Macia, A. (2011) Epigenetic control of retrotransposon expression in human embryonic stem cells. Mol Cell Biol. 31, 300-316.

Mätlik, K. (2006) L1 antisense promoter drives tissue-specific transcription of human genes. J Biomed Biotechnol. 2006, 71753.

Nigumann, P. (2002) Many human genes are transcribed from the antisense promoter of L1 retrotransposon. Genomics. 79, 628-634.

Philippe, C. (2016) Activation of individual L1 retrotransposon instances is restricted to cell-type dependent permissive loci. Elife. 5, e13926.

Speek, M. (2001) Antisense promoter of human L1 retrotransposon drives tran-scription of adjacent cellular genes. Mol Cell Biol. 21, 1973-1985.

Wolff, E.M. (2010) Hypomethylation of a LINE-1 promoter activates an alternate transcript of the MET oncogene in bladders with cancer. PLoS Genet. 6, e1000917.

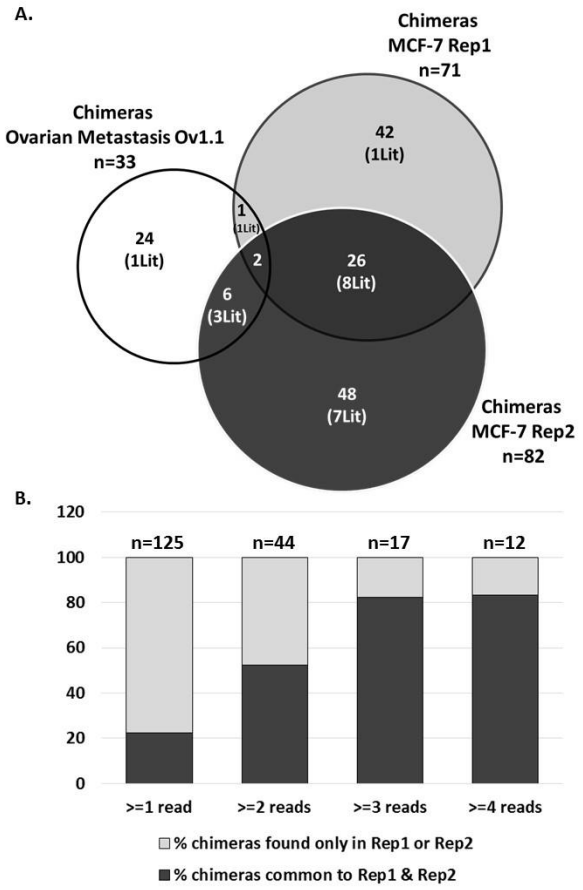

Figure 1: Analysis of the chimeras identified in stranded paired-end RNA-seq datasets from different tissues by CLIFinder

A. Comparison of chimeras identified in the Ov1.1 dataset and the two replicates (Rep1 and Rep2) from the untreated breast cancer cell line MCF-7. (Lit) indicates the number of chimeras corresponding to LCTs already described in the literature. Two to three times more chimeras were identified in the MCF-7 RNA-seq datasets than in the Ov1.1 dataset, in agreement with the higher sequencing depth. Nine chimeras were found both in the Ov1.1 and MCF-7 datasets. Already known LCTs were identified in each dataset.

B. Analysis of the concordance concerning the number of chimeras identified in the two MCF-7 replicates. Among all identified chimeras (detected by at least one read in one replicate), 22% were common to both replicates. When only chimeras depicted by at least 2, 3 or 4 reads in one replicate were considered, the overlap increased respectively to 52%, 82% and 83%, suggesting that CLIFinder analysis results are reproducible.