Proc. Annual Meeting of the Austrian Society for Biomedical Engineering 2021
DOI: 10.3217/978-3-85125-826-4-25

T3E: TRANSPOSABLE ELEMENT ENRICHMENT ESTIMATOR

Michelle Almeida da Paz and Leila Taher

Institute of Biomedical Informatics, TU Graz, Austria

michelle.almeidadapaz@tugraz.at

Because of their repetitive nature, short sequencing reads derived from transposable elements (TEs) cannot be unambiguously mapped to the reference genome. As a result, most genomic analyses neglect over 50% of the human genome. Here, we present T3E, an algorithm to characterize the histone modifications associated with TEs from Chromatin Immunoprecipitation Sequencing (ChIP-seq) data. T3E relies on the structure of the ChIP-seq control experiment to assess enrichment. When applying T3E to five ChIP-seq libraries we found consistently fewer enrichments compared to a strategy that assumes a random distribution of the reads across the genome, suggesting that the latter has a high false positive rate. This provides a framework for the functional analysis of TEs.

Keywords— *Transposable element, enrichment, ChIP-seq, histone modifications*

Introduction

Over half of the human genome consists of repetitive sequences, including transposable elements (TEs) [1]. Based on their sequence and transposition mechanisms, TEs have been hierarchically classified into several groups/subgroups [2]. TEs are contributors to regulatory network evolution, playing role as host promoters, enhancers, and forming silencer/insulator regions [3]. To study the exaptation of TEs as cis-regulatory elements, we aim to quantitatively investigate the relationship between epigenetic histone modifications and TE groups/subgroups.

Repetitive sequences pose several analytical challenges to current short-read sequencing technologies. Specifically, reads originating from repetitive sequences will often map to multiple loci (multi-mappers) and cannot be unambiguously assigned to any region of the genome. The problem has been tackled in different ways. For example, some strategies simply use one random mapping, increasing the number of mapped reads, but reducing the precision of the mapping [4]. Others discard multi-mappers from the analysis and use only uniquely mapped reads [5].

Here, we present T3E, an algorithm that identifies TE groups featuring enrichment for specific histone modifications using chromatin immunoprecipitation followed by sequencing (ChIP-seq) data.

Methods

Selection of ChIP-seq datasets

We selected five ChIP-seq samples for the H9 cell line from the ENCODE Project data repository [6]: H3K4me1 and H3K4me3 (active euchromatin), H3K9me3 and H3K27me3 (repressed heterochromatin), and H4K8ac (both euchromatin and heterochromatin). All single-ended sequencing libraries were generated by the laboratory of Zhiping Weng, UMass Medical School.

ChIP-seq reads quality control and mapping

The raw data quality of all samples and their respective "input" controls (FASTQ files) were assessed using FASTQC [7]. Sequencing adapters were removed and low-quality reads (minimum Phred score of 10) were filtered out/or trimmed. Mapping was performed using BWA mem [8] against the GRCh37/hg19 assembly of the human genome with the parameter "-a", which outputs all found alignments for the single-end reads. The resulting mappings (BAM files) were processed with SAMtools [9] and BEDtools [10] to filter out unmapped reads, non-chromosomal scaffolds, and reads mapping to the mitochondrial chromosome (Tab. 1). The "input" controls were processed in the same manner.

Table 1. ENCODE Project ChIP-seq libraries considered in this study. * Number of processed reads. ENCFF969KKW has 9,862,491 reads and a read length of 30 base pairs (bp). ENCFF416GCS has 16,845,808 reads and a read length of 36 bp. The read length considered for the sample is the same for the corresponding "input" control.

Histone modifications	File accession	* Read count	Input file accession
H4K8ac	ENCFF974MOD	5508640	ENCFF969KKW
H3K9me3	ENCFF776JLA	7054172	ENCFF969KKW
H3K4me3	ENCFF909NXO	9771318	ENCFF416GCS
H3K27me3	ENCFF212TLT	12002119	ENCFF416GCS
H3K4me1	ENCFF210BMG	16779354	ENCFF416GCS

TE groups/subgroups

Repeat annotation for the GRCh37/hg19 assembly of the human genome was obtained from the RepeatMasker track of the UCSC Genome Browser [11]. Repeat annotation was processed to filter out simple repeats (micro-satellites), low complex repeats, satellite DNA, RNA repeats (including RNA, tRNA, rRNA, snRNA, scRNA, srpRNA, non-TE elements and uncommon repeats (less than 100 instances). Adjacent and overlapping TE instances of the same group/subgroup were merged.

Published by Verlag der TU Graz
Graz University of Technology

DOI: 10.3217/978-3-85125-826-4-25

Reads associated with a TE group/subgroup
The contribution of a read to a TE group/subgroup considers the fraction of a read mapping to a given TE instance and the total number of mappings for the read in the genome:

$$C_K = \sum_{k \in K} \sum_{r \in S} \sum_{i=1}^{N_r} \frac{l_{k\,r_i}}{N_r\, L_r} \quad (1)$$

where K is the set of all instances of a TE group/subgroup in the genome, S is the set of all reads in the sample, N_r is the number of mappings of read r, L_r is the length of read r, and $l_{k\,r_i}$ is the number of nucleotides of the ith mapping of read r overlapping with TE instance k, where $\{l_{k\,r_i} \in \mathbb{Z}_0^+ : 0 \leq l_{k\,r_i} \leq L_r\}$.

Input-based background probability distribution
The estimated probability of a mapping starting at position n on chromosome $c \in \{1, 2, \ldots, 22, X, Y\}$ of the genome is calculated based on the reads in the "input" library and defined by:

$$p_n = \frac{\sum_{r \in K} \frac{1}{N_r}}{\sum_{r \in M} \frac{L_r}{N_r}} \quad (2)$$

where L_r is the length of read r, N_r is the number of mappings of read r, K is the set of all read mappings on chromosome c overlapping n, and M is the set of all read mappings on chromosome c. Note that nucleotides with zero coverage have no probability assigned and are consequently excluded from the analysis.
We sample genomic positions from the corresponding empirical cumulative distribution for a given chromosome using discrete sample. Then, among the reads mapping to that position, we randomly select one. Finally, we identify all other mappings of the selected read (if any). The process is repeated as many times as there are reads in the ChIP-seq library of interest, resulting in a simulated "input" library of the same size of the ChIP-seq library.

TE group enrichment analysis
For each ChIP-seq library, we simulated *N*=100 "input" libraries. For each of them, we computed the number of reads associated with a TE group/subgroup as described above.
For each TE group/subgroup, the number of reads in the ChIP-seq library was compared to the number in the simulated "input" libraries using a permutation test. A P-value was calculated as the number of simulated "input" libraries with a number of reads higher than or equal to the number of reads associated with the TE group/subgroup in the ChIP-seq library divided by *N*. A fold-change (FC) was computed as the ratio between the number of reads associated with the TE group/subgroup in the ChIP-seq library divided by the average of the number of reads associated with the TE group/subgroup across all *N* simulated "input" libraries.
Note that enrichment was calculated for TE groups/subgroups, not for individual TE instances.

Uniform background distribution
A more traditional method to define a background distribution assumes a uniform distribution of the sequencing reads across the genome. Thereby, the reads of a ChIP-seq library are randomly shuffled across the entire genome.

Computational specifications and execution time
The algorithm is written in Python 3 and was executed in two machines using Python 3.8.5. Three samples were processed on a computer with AMD Ryzen 9 3900X, 12 cores, with in total 128 GB of RAM and running Linux version 5.8.0-41-generic (machine 1). Two samples were processed using a computer with AMD Ryzen Threadripper 3970X, 32 cores, with in total 128 GB of RAM and running Linux version 5.8.0-44-generic (machine 2). The execution time increased approximately linearly with the library size (Fig. 1).

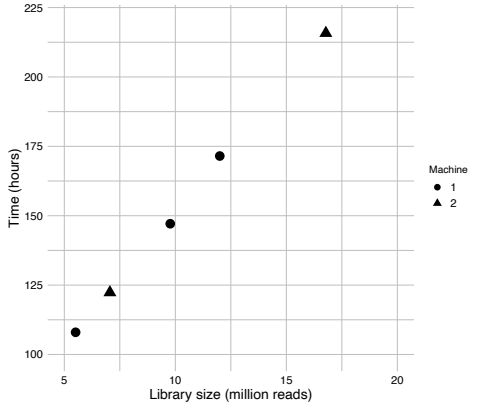

Figure 1. Execution time as a function of the ChIP-seq library size. Note that the times correspond to two different machines.

Results

In T3E the number of mappings observed for a given read is taken into account to quantify the contribution of a read to a TE group/subgroup. By doing this, every single nucleotide mapping onto a TE instance is counted and weighted by the uncertainty of where multi-mapper reads come from. As background for the enrichment analysis, the algorithm constructs a probability distribution of read mappings based on the read mappings in the ChIP-seq control experiment (Fig. 2).
In total, the repeat annotation comprises 860 different TE groups/subgroups covering 44.83% of the human genome. Read mapping statistics evaluation shows a substantial percentage of alignments uniquely mapping on TE regions (Fig. 3), indicating that although different instances of the same TE group/subgroup have repetitive sequences, they are not identical. Multi-mapper reads also mapped to non-TE regions, indicating the presence of other genomic repetitive sequences or non-annotated TEs.
The number of reads that are expected to be mapped by chance to TE groups/subgroups computed based

Published by Verlag der TU Graz
Graz University of Technology

DOI: 10.3217/978-3-85125-826-4-25

on the reads in the "input" library is strongly correlated with that computed assuming a uniform distribution (Fig. 4). However, LINE and SINE major groups exhibit clear deviations, in particular for LINEs using the ENCFF416GCS "input" library (Fig. 4).

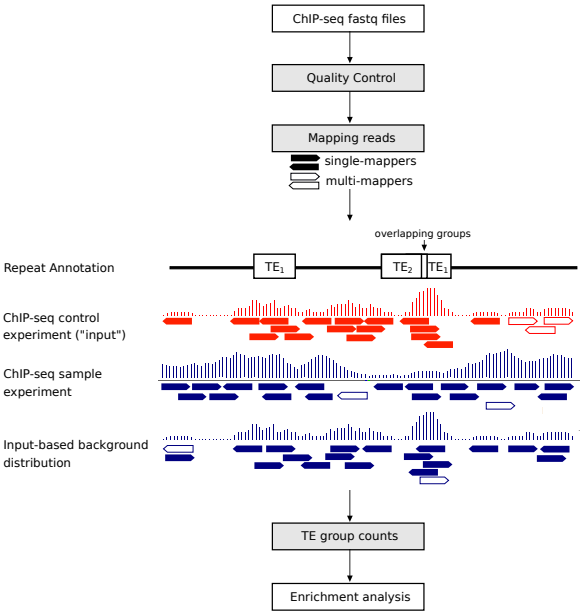

Figure 2. T3E algorithm strategy. The structure of the "input" control is used to construct the background. Overlapping of different TE groups are shown and reads mapping on this region contribute partially to both groups.

Consistently, T3E identified fewer TE groups/subgroups featuring histone modification enrichments. On average, T3E found only 11.11% of the TE groups/subgroups identified based on the uniform background distribution (Fig. 5). In total, 11 TE groups/subgroups showed enrichment: 4 to H3K4me3 (including LTRs, MER57E3 and HERVH-int, all classified into ERV1 - endogenous retroviruses group) and 7 to H3K9me3 (LTR12E and 6 SVA retrotransposons).

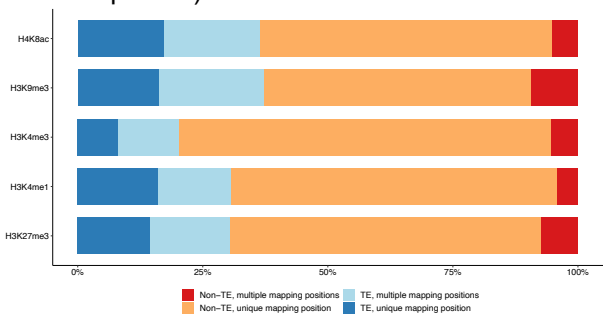

Figure 3. Read mapping statistics. On average, 10,223,121 reads comprised 64.04% uniquely mapped on non-TE, 15.77% multiple mapped on TE regions, 14.32% uniquely mapped on TEs, and 5.88% multiple mapped on non-TEs.

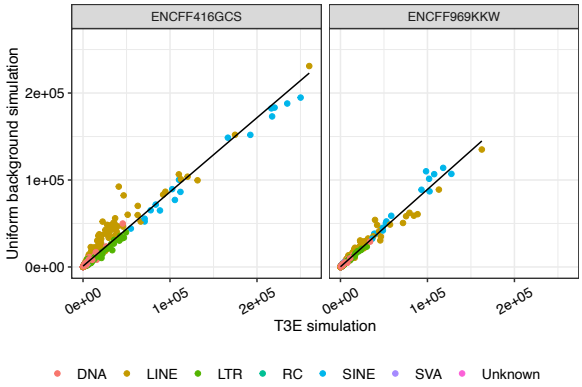

Figure 4. "Input" control simulations. Comparison of T3E input-based background and uniform background distribution counts as the average of 10 iterations. Each dot represents one TE group/subgroup. Unknown represents ancient TE not yet classified [12].

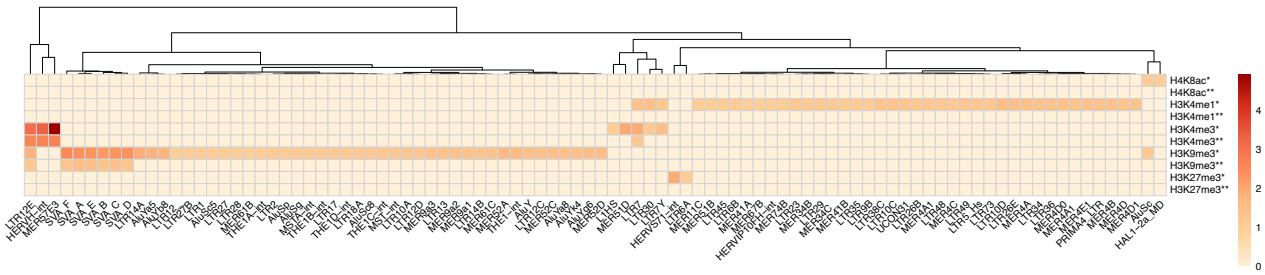

Figure 5. TE group/subgroups enrichment for different histone modifications. Only TE group/subgroups exhibiting a P-value ≤ 0.05 and a $\log_2$ FC ≥ 1 were considered enriched. Light-coloured cells indicate no enrichment. Red intensity varies according to $\log_2$ FC values, from 1.0 to the maximum found ($\log_2$ FC = 4.99). (*) Uniform background distribution. (**) T3E. Columns were clustered using Euclidean distance and complete linkage.

Discussion

Although a considerable effort has been done to study TEs in an integrative manner, several challenges are faced when mapping TE-derived sequences and predicting the enrichment of TEs in high-throughput sequencing, such as ChIP-seq data. Long reads increase the uniqueness of sequenced fragments and

Published by Verlag der TU Graz
Graz University of Technology

DOI: 10.3217/978-3-85125-826-4-25

also help in the mapping step, but still, it is a limiting factor in ChIP-seq and many other applications.

Like the repEnrich [13] method, T3E accounts for the uncertainty in the mapping of multi-mapper reads by dividing by the total number of mappings. In addition, T3E was developed to use the structure of the "input" library to estimate TE enrichments. Thus, the probability of observing a mapping at a given genomic position reflects the read distribution of the "input" control. Our approach avoids the bias of a uniform background, which does not reflect the read mappings, since TEs are not uniformly distributed across the genome [14] and the read mappings in the "input" control have a specific distribution. This is reflected in the decrease in the number of TE groups/subgroups showing enrichment. Furthermore, T3E's strategy of randomly sampling read mappings based on the "input" library takes into account potential library preparation biases. It also eliminates the need for normalization in enrichment computations, preventing the removal of true biological variations.

In summary, T3E is more conservative compared to other current approaches and has the benefit of estimating TE enrichment of groups/subgroups at a nucleotide resolution, without the need of further normalizations. Although this study is a proof of principle, it provides a framework for the analysis of the regulatory functions of TEs.

Acknowledgements

This research was funded by the Austrian Science Fund (FWF) [P33437-B].

References

[1] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature. 2001 Feb 15;409(6822):860–921.

[2] Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 2007 Dec;8(12):973–82.

[3] Etchegaray E, Naville M, Volff J-N, Haftek-Terreau Z. Transposable element-derived sequences in vertebrate development. Mob DNA. 2021 Dec;12(1):1.

[4] Teissandier A, Servant N, Barillot E, Bourc'his D. Tools and best practices for retrotransposon analysis using high-throughput sequencing data. Mob DNA. 2019 Dec;10(1):52.

[5] Jin Y, Tam OH, Paniagua E, Hammell M. TEtranscripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets. Bioinformatics. 2015 Nov 15;31(22):3593–9.

[6] Davis CA, Hitz BC, Sloan CA, Chan ET, Davidson JM, Gabdank I, et al. The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res. 2018 Jan 4;46(D1):D794–801.

[7] Andrews S. FastQC: A Quality Control Tool for High Throughput Sequence Data [Internet]. 2010. Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

[8] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009 Jul 15;25(14):1754–60.

[9] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078–9.

[10] Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010 Mar 15;26(6):841–2.

[11] Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The Human Genome Browser at UCSC. Genome Res. 2002 May 16;12(6):996–1006.

[12] Kojima KK. Human transposable elements in Repbase: genomic footprints from fish to humans. Mob DNA. 2018 Dec;9(1):2.

[13] Criscione SW, Zhang Y, Thompson W, Sedivy JM, Neretti N. Transcriptional landscape of repetitive elements in normal and cancer human cells. BMC Genomics. 2014 Dec;15(1):583.

[14] Ahmed M, Liang P. Transposable Elements Are a Significant Contributor to Tandem Repeats in the Human Genome. Comp Funct Genomics. 2012;2012:1–7.