Genomics 110 (2018) 257–262

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

RTFAdb: A database of computationally predicted associations between retrotransposons and transcription factors in the human and mouse genomes

Gökhan Karakülah*

İzmir International Biomedicine and Genome Institute (iBG-İzmir), Dokuz Eylül University, 35340, İnciraltı, İzmir, Turkey

ARTICLE INFO

Keywords:
Retrotransposons
Transcription factors
Gene regulation
Mammalian genomes
Bioinformatics

ABSTRACT

In recent years, retrotransposons have gained increasing attention as a source of binding motifs for transcription factors (TFs). Despite the substantial roles of these mobile genetic elements in the regulation of gene expression, a comprehensive resource enabling the investigation of retrotransposon species that are bound by TFs is still lacking. Herein, I introduce for the first time a novel database called RTFAdb, which allows exploring computationally predicted associations between retrotransposons and TFs in diverse cell lines and tissues of human and mouse. My database, using over 3.000 TF ChIP-seq binding profiles collected from human and mouse samples, makes possible searching more than 1.500 retrotransposon species in the binding sites of a total of 596 TFs. RTFAdb is freely available at http://tools.ibg.deu.edu.tr/rtfa/ and has the potential to offer novel insights into mammalian transcriptional networks by providing an additional layer of information regarding the regulatory roles of retrotransposons.

1. Introduction

Retrotransposons (also called class I elements) are a class of transposable element (TE) that retain the ability to mobilize and insert themselves into novel regions of the host genome through RNA intermediate [1]. Retrotransposon mobilization begins with transcription of an RNA intermediate from a genomic copy, followed by its reverse transcription into DNA by the retrotransposon-encoded reverse transcriptase [2]. In this way, they accumulate in DNA and contribute to genome size expansion [3] and genome evolution [4]. Previous studies indicate their crucial roles in a wide range of biological events, including disease [5], development [6,7], and differentiation [8].

In recent years, retrotransposons have been gaining increasing attention due to their regulatory roles in gene expression (for review see [9]). They can act as novel promoters that introduce new transcription start sites for nearby genes, and thus contribute to remodeling of gene structures and generation of novel transcript variants [10,11]. For example, Peaston *et al.* reported that certain mouse-specific MT retrotransposons provide novel promoters and first exons for a group of genes and regulate their expression levels in mouse oocytes and zygotes [12]. In a more recent study, it has been shown that the insertion of MT-C retrotransposon into the intronic region of the mouse Dicer gene, essential for microRNA biogenesis, leads to the gain of a retrotransposon-derived novel promoter for the host gene and the generation of an oocyte-specific Dicer isoform [13].

In addition to providing novel promoters to their host genes, retrotransposons can harbor *cis-regulatory* sequences, and therefore act as enhancers. In this way, they influence transcriptional activity of nearby genes by providing binding sites for transcription factors (TFs). One early example of a retrotransposon-derived transcription factor binding site (TFBS) is associated with the well-characterized oncogene human tumor suppressor protein p53 (TP53) [14]. Wang *et al.* reported that five members of LTR10 and MER61 retrotransposon families are over-represented in the binding sites of TP53, and they show TP53-dependent enhancer activity [14]. Another chromatin immunoprecipitation followed by sequencing (ChIP-seq) based study exhibited the significance of retrotransposons in the evolution of gene regulatory networks in human and mouse embryonic stem cells [15]. The study clearly presented that ~25% of OCT4 and NANOG binding sites were contributed by TEs, in particular by the endogenous retrovirus (ERV) 1 retrotransposon family, in mammalian species. In addition, a systematic computational analysis of publicly available DNase I hypersensitivity data sets demonstrated that ERV retrotransposons are the source of a majority of regulatory elements in human [16]. Recent studies have also provided novel evidence for the regulatory roles of ERVs in mammalian immune defense [17] and embryogenesis [18].

To date, a few computational tools and methodologies have been developed to facilitate exploring the regulatory roles of TEs in gene expression. Among these, a simple analysis pipeline proposed by Conley and Jordan is utilized for the identification of potential TE-derived

* Corresponding author.
E-mail address: gokhan.karakulah@deu.edu.tr.

https://doi.org/10.1016/j.ygeno.2017.11.002
Received 20 September 2017; Received in revised form 31 October 2017; Accepted 14 November 2017
Available online 17 November 2017

TFBSs from ChIP-seq data sets [19]. Similarly, Ramsay and Bourque suggested a practical pipeline that allows testing whether a certain repeat element is significantly overrepresented in a given set of ChIP-seq peaks [20]. Another study, using z-score based statistics, introduced a solid computational methodology for the enrichment analysis of TEs observed in the binding sites of TFs [21]. The GSuite Hyperbrowser [22] also provides users with a set of tools to perform statistical analyses of genomic features, including transcription factor binding sites derived from Encyclopedia of DNA Elements (ENCODE) [23] and repeat element annotations from RepeatMasker. However, it currently allows the acquisition and investigation only of human data sets from the ENCODE database. Most recently, the PlanTEnrichment tool has been established to predict TEs that potentially influence the expression of nearby genes in plants [30]. Despite the key roles of retrotransposons that harbor regulatory sequences for TFs, to date, a comprehensive database offering convenient access to the associations between TEs and TFs is still lacking.

In this study, I report for the first time a living database called RTFAdb, which functions as an exploration tool for computationally predicted associations of retrotransposons and TFs in the human and mouse genomes. Complementing the current TE databases, RTFAdb allows researchers, by using publicly available TF ChIP-seq binding profiles, to investigate (i) retrotransposon species which are over-represented in the binding sites of a given TF, (ii) TFs significantly associated with the retrotransposon of interest, and (iii) all putative retrotransposon—TF associations within the available tissue or cell types.

2. Materials and methods

2.1. Data collection and processing

Fig. 1 shows the steps of data processing and database development. I downloaded all available human and mouse TF ChIP-seq data in BED file format and their respective experiment metadata from the ENCODE web site (https://www.encodeproject.org/). To achieve high confidence binding site coordinates of the human and mouse TFs, I considered only the replicated TF ChIP-seq experiments with optimal and conservative irreproducible discovery rate (IDR) thresholded peaks for further downstream analysis. The genomic locations of the human and mouse repeat elements (RepeatMasker annotation) in text file format were downloaded from the UCSC genome annotation database [24]. To determine the retrotransposon species that intersect with TFBSs, I made use of the intersectBed command (intersectBed -f 0.1 -u -a RepeatMasker.bed -b TF·ChIP-seq.bed) of bedtools [25]. Additionally, the shuffleBed command (shuffleBed -i TF·ChIP-seq.bed -g genome.chrom.sizes) was utilized to randomly permute the genomic locations of each TF ChIP-seq BED file. I utilize these randomly permuted regions as null basis while testing statistical significance of each retrotransposon observed in the binding sites of TFs. To increase the confidence of my statistical analysis, I iterated the shuffleBed command 100 times for each BED file. Afterwards, I identified all retrotransposon species, which were overlapping randomly permuted regions in the BED files generated by shuffleBed. The total number of each retrotransposon in the respective genome, and the frequencies of each retrotransposon overlapping TFBSs and randomly permuted regions, were calculated with *cut*, *sort*, and *uniq* command line utilities of the Linux environment.

2.2. Identification of statistically overrepresented retrotransposons in the binding sites of transcription factors

To calculate the significance level of each retrotransposon—TF association observed in the ChIP-seq experiments, I utilized previously introduced methodology based on binomial test statistics [20]: let R denote a specific retrotransposon in the respective genome and let T a certain transcription factor of the species of interest. Then the probability value of the R—T association in a ChIP-seq experiment is described as follows:

$$The\ p-value\ of\ the\ R—T\ association = \frac{n!}{(n-X)!X!}(p)^X(1-p)^{n-X}$$

- where X is the total number of R that intersect with the binding sites of T in the ChIP-seq experiment
- n is the total count of R in the respective genome
- p is the ratio of the total number of R that intersect with the randomly permuted genomics intervals and the frequency of R in the respective genome (n).

I made use of the binom.test (binom.test (X, n, p, alternative = “greater”, conf.level = 0.95)) function of R (https://www.r-project.org/; v3.3.2) [26] statistical computing environment for the calculation of the p-value of each retrotransposon—TF pair and its confidence interval. Thereby, I identified all statistically over-represented retrotransposons in a set of TF ChIP-seq peaks.

2.3. Implementation of database and user interfaces

The database and user interfaces of RTFAdb were written in the LAMP (Linux, Apache, MySQL, PHP) development environment. RTFAdb is a living database that comprises distinct data sets, such as retrotransposon annotations in EMBL format collected from Repbase (http://www.girinst.org/repbase/) [27], metadata of TF ChIP-seq

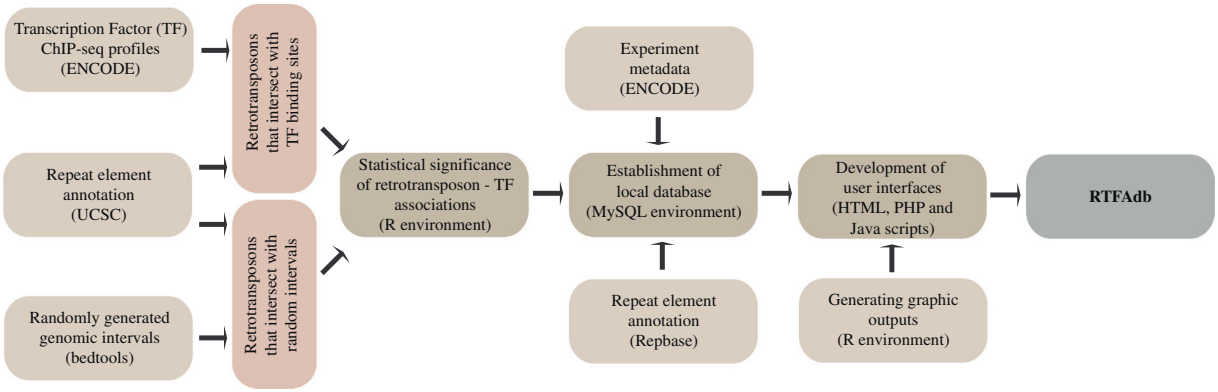


Fig. 1. Summary of the workflow of the study. RTFAdb uses diverse public data sets from ENCODE, UCSC Genome Browser, and Repbase. After the identification of retrotransposons species that intersect with the binding sites of TFs, I calculated statistically overrepresented retrotransposons in the TFBSs. RTFAdb was developed in the LAMP environment, and all putative retrotransposon—TF associations were stored in a local MySQL relational database management system. The user interfaces of RTFAdb were built using HTML5 markup and PHP scripting languages, and were enhanced with JQuery and AJAX technologies.

experiments, and binomial test statistics of each retrotransposon—TF pairs. I stored all data sets in my local MySQL (https://www.mysql.com/; v5.5.52) relational database tables. RTFAdb was located on a Linux (https://www.centos.org/; CentOS Linux release 7.3(1611)) machine with Apache hypertext transfer protocol server (https://www.apache.org/; v2.4.6). The user interfaces of RTFAdb were developed with Hyper Text Markup Language 5 (HTML5) with responsive Pure CSS modules (https://purecss.io/; v0.6.2). Database queries of RTFAdb were developed with structured query language, and the communication between the user interfaces of RTFAdb and the MySQL database was implemented in PHP (http://www.php.net/; v5.4.16). The user interfaces of RTFAdb were also improved with JQuery (https://jquery.com/) and Asynchronous JavaScript and XML methods to provide users with a responsive and practical search experience. In addition, the on-the-fly dot plot graphs produced by RTFAdb in each database query are performed with the ggplot2 package [28] of the R environment.

3. Results

3.1. The content of RTFAdb database

RTFAdb was built based on public ChIP-seq profiles, and retrotransposon annotations from RepeatMasker and Repbase. I downloaded over 3.000 (as of May 2017) TF binding profiles for both human and mouse from the ENCODE project web site. The basic statistics regarding the content of RTFAdb are shown in Table 1. The database currently allows for exploring putative associations of approximately 1500 retrotransposon species in the binding sites of a total of 596 human and mouse TFs. Per the RepeatMasker annotation of hg19 genome assembly, I found that MIRb was the richest retrotransposon species observed in the human genome. However, B3 is the most abundant for the mouse reference genome (mm10). Total number of the 10 most abundant retrotransposons in the human and mouse genomes are shown in Fig. 2. When I performed the binomial test statistics among as many as 1.600.000 retrotransposon—TF associations, I identified approximately 734.000 statistically significant retrotransposon—TF pairs (p-value $<$ 0.05), suggesting the potential role of retrotransposons in the regulation of gene expression.

3.2. Search by transcription factor

RTFAdb has been developed in a user-friendly manner and comprises three distinct search modules, which allow the user to explore computationally predicted retrotransposon—TF associations in the organism of interest (Fig. 3a). Using the “Search by Transcription Factor” module, accessible on the home page of RTFAdb, the user can retrieve all statistically overrepresented retrotransposons in the binding sites of the selected TF. Once the target TF, appropriate ChIP-seq peak type, and p-value parameters are chosen from the pull-down menu, it lists all retrotransposon—TF associations below the given p-value threshold in the Results page (Fig. 3b). The interactive results table appearing in the results page permits searching of any keyword within the table and sorting by column headers. RTFAdb also provides sequence information

Table 1
The basic statistics regarding the content of RTFAdb.

Entity	Total number	
	Human	Mouse
ChIP-seq binding profiles	2.778	457
Transcription factors (TF)	548	48
Retrotransposons	702	880
Cell/tissue type	79	29
Significant* retrotransposon – TF associations	702.597	31.659

* p-value $<$ 0.05

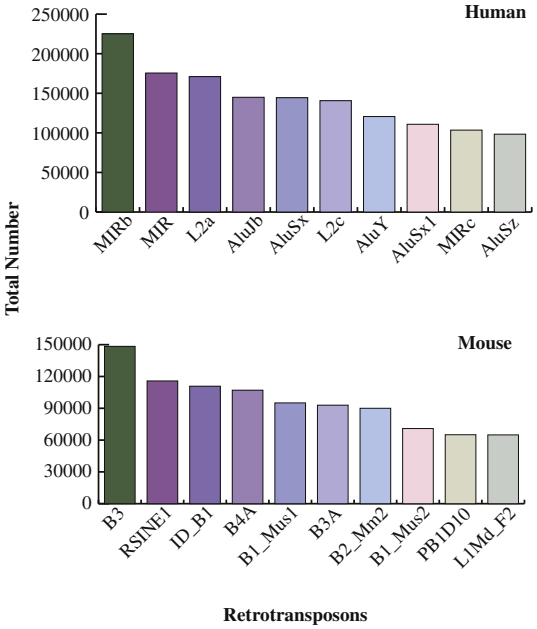


Fig. 2. The 10 most abundant retrotransposons in human and mouse genomes. Per the RepeatMasker annotation of UCSC Genome Browser, the most abundant retrotransposons in the human and mouse genomes are MIRb and B3, respectively. While Alu repeat elements dominate a large proportion of the human genome, short-interspersed nuclear elements are rich in the mouse genome.

(if available; Fig. 3c) of the retrotransposon of interest, and ChIP-seq experiment details when the user clicks on retrotransposon names and experiment IDs. Moreover, the confidence interval of the selected p-value and the parameters used in the binomial test statistics are shown as a bar plot and in table format in a new window upon clicking the corresponding p-value hyperlink (Fig. 3d).

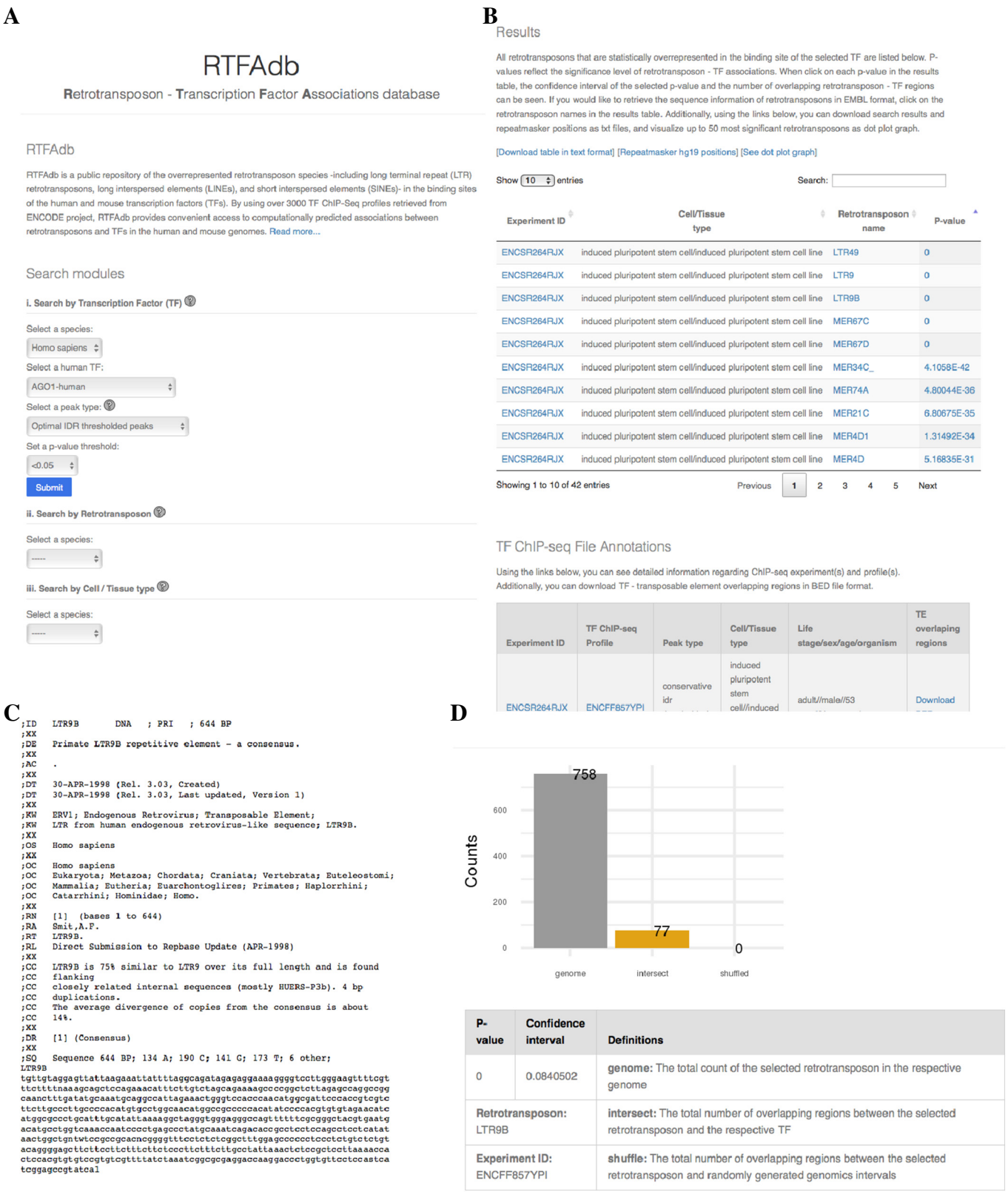
To visualize up to 50 most statistically significant retrotransposons, observed in the TFBSs, in portable network graphics (PNG) format, the “see dot plot graph” link located above the results table can be utilized. Additionally, RTFAdb allows the download of genomic coordinates of retrotransposons in the selected organism and putative associations as separate tab delimited files. The detailed metadata of ChIP-seq experiment(s) associated with the selected TF are exhibited in the “TF ChIP-seq File Annotations” table. By clicking the experiment and ChIP-seq profile hyperlinks, users are directed to the corresponding ENCODE project page so that they can review the summary of the ChIP-seq experiment and download the respective ChIP-seq peaks file in BED file format. The genomic intervals of the retrotransposons species that overlap with the binding sites of the user-defined TF are provided as a separate hyperlink and can be downloaded in BED format.

3.3. Search by retrotransposon name

This module enables researchers to investigate whether the retrotransposon of interest is significantly overrepresented in the binding sites of any TF. After the appropriate parameters are selected, and search operation is performed, all putative retrotransposon—TF associations along with their pre-calculated significance levels and the summary of ChIP-seq experiments can be investigated. Since I use a similar layout for the results pages in the user interface design, all search modules provide data in the same type and format. The details regarding modules and how to interpret the results are provided on the tutorial page of RTFAdb.

3.4. Search by cell or tissue type

One of the unique features of the RTFAdb database is the ability to


A
RTFAdb
Retrotransposon - Transcription Factor Associations database
RTFAdb
RTFAdb is a public repository of the overrepresented retrotransposon species -including long terminal repeat (LTR) retrotransposons, long interspersed elements (LINEs), and short interspersed elements (SINEs)- in the binding sites of the human and mouse transcription factors (TFs). By using over 3000 TF ChIP-Seq profiles retrieved from ENCODE project, RTFAdb provides convenient access to computationally predicted associations between retrotransposons and TFs in the human and mouse genomes. Read more...
Search modules
i. Search by Transcription Factor (TF)
Select a species:
Homo sapiens
Select a human TF:
AGO1-human
Select a peak type:
Optimal IDR thresholded peaks
Set a p-value threshold:
<0.05
Submit
ii. Search by Retrotransposon
Select a species:
iii. Search by Cell / Tissue type
Select a species:
B
Results
All retrotransposons that are statistically overrepresented in the binding site of the selected TF are listed below. P-values reflect the significance level of retrotransposon - TF associations. When click on each p-value in the results table, the confidence interval of the selected p-value and the number of overlapping retrotransposon - TF regions can be seen. If you would like to retrieve the sequence information of retrotransposons in EMBL format, click on the retrotransposon names in the results table. Additionally, using the links below, you can download search results and repeatmasker positions as txt files, and visualize up to 50 most significant retrotransposons as dot plot graph.
[Download table in text format] [Repeatmasker hg19 positions] [See dot plot graph]
Show 10 entries
Search:
Experiment ID | Cell/Tissue type | Retrotransposon name | P-value
ENCSR264RJX | induced pluripotent stem cell/induced pluripotent stem cell line | LTR49 | 0
ENCSR264RJX | induced pluripotent stem cell/induced pluripotent stem cell line | LTR9 | 0
ENCSR264RJX | induced pluripotent stem cell/induced pluripotent stem cell line | LTR9B | 0
ENCSR264RJX | induced pluripotent stem cell/induced pluripotent stem cell line | MER67C | 0
ENCSR264RJX | induced pluripotent stem cell/induced pluripotent stem cell line | MER67D | 0
ENCSR264RJX | induced pluripotent stem cell/induced pluripotent stem cell line | MER34C_ | 4.1058E-42
ENCSR264RJX | induced pluripotent stem cell/induced pluripotent stem cell line | MER74A | 4.80044E-36
ENCSR264RJX | induced pluripotent stem cell/induced pluripotent stem cell line | MER21C | 6.80675E-35
ENCSR264RJX | induced pluripotent stem cell/induced pluripotent stem cell line | MER4D1 | 1.31492E-34
ENCSR264RJX | induced pluripotent stem cell/induced pluripotent stem cell line | MER4D | 5.16835E-31
Showing 1 to 10 of 42 entries
Previous 1 2 3 4 5 Next
TF ChIP-seq File Annotations
Using the links below, you can see detailed information regarding ChIP-seq experiment(s) and profile(s). Additionally, you can download TF - transposable element overlapping regions in BED file format.
Experiment ID | TF ChIP-seq Profile | Peak type | Cell/Tissue type | Life stage/sex/age/organism | TE overlaping regions
ENCSR264RJX | ENCFF857YPI | conservative idr | induced pluripotent stem cell//induced | adult//male//53 | Download
C
;ID LTR9B DNA ; PRI ; 644 BP
;XX
;DE Primate LTR9B repetitive element - a consensus.
;XX
;AC .
;XX
;DT 30-APR-1998 (Rel. 3.03, Created)
;DT 30-APR-1998 (Rel. 3.03, Last updated, Version 1)
;XX
;KW ERV1; Endogenous Retrovirus; Transposable Element;
;KW LTR from human endogenous retrovirus-like sequence; LTR9B.
;XX
;OS Homo sapiens
;XX
;OC Homo sapiens
;OC Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
;OC Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
;OC Catarrhini; Hominidae; Homo.
;XX
;RN [1] (bases 1 to 644)
;RA Smit,A.F.
;RT LTR9B.
;RL Direct Submission to Repbase Update (APR-1998)
;XX
;CC LTR9B is 75% similar to LTR9 over its full length and is found
;CC flanking
;CC closely related internal sequences (mostly HUERS-P3b). 4 bp
;CC duplications.
;CC The average divergence of copies from the consensus is about
;CC 14%.
;XX
;DR [1] (Consensus)
;XX
;SQ Sequence 644 BP; 134 A; 190 C; 141 G; 173 T; 6 other;
LTR9B
tgttgtaggagttattaagaaattattttaggcagatagagaggaaaaggggtccttgggaagttttcgt
ttcttttaaagcagctccagaaacatttcttgtctagcagaaaagccccggctcttagagccaggccgg
caanctttgatatgcaaatgcaggccattagaaactgggtccacccaacatggcgattcccaccgtcgtc
ttcttgcccttgccccacatgtgcctggcaacatggccgccccacatatccccacgtgtgtagaacatc
atggcgccct gcatttgcatattaaaaggctagggtgggagggccagtttttcgcgggctacgtgaatg
acatgcctggtcaaaccaatcccctgagccctatgcaaatcagacaccgcctcctccagcctcctcatat
aactggctgntwtccgccgcacncggggtttcctctctcggctttggagccccctcctctgtctctgt
acaggggagcttctttctttcttctccctttcttttcttgcctattaaactctcgctccttaaaacca
ctccacgtgtgtccgtgtcgttttatctaaatcggcgcgaggaccaaggacccttggtgttcctccastca
tcggagccgtatcal
D

P-value | Confidence interval | Definitions
0 | 0.0840502 | genome: The total count of the selected retrotransposon in the respective genome
Retrotransposon: LTR9B | | intersect: The total number of overlapping regions between the selected retrotransposon and the respective TF
Experiment ID: ENCFF857YPI | | shuffle: The total number of overlapping regions between the selected retrotransposon and randomly generated genomics intervals

Fig. 3. Screenshots of the RTFAdb home and results pages. (a) RTFAdb home page includes three search modules, which enable searching by TF, retrotransposon name, and tissue/cell type. (b) A searchable results table in the results page lists all putative retrotransposon—TF associations, along with their statistical significance and experiment metadata. (c) Retrotransposon sequences can be shown in EMBL format. (d) The parameters utilized in probability calculation of retrotransposon—TF associations can be visualized in a separate window when clicking the respective *p*-value hyperlink in the results table.

provide all significant putative associations between retrotransposons and TF in the selected cell or tissue type. This module is helpful for those wishing to conduct a single query using all available retrotransposon species against existing TF binding profiles in a particular cell or tissue. Currently, the RTFAdb tool offers two options for peak type selection: (i) optimal and (ii) conservative peak sets. I

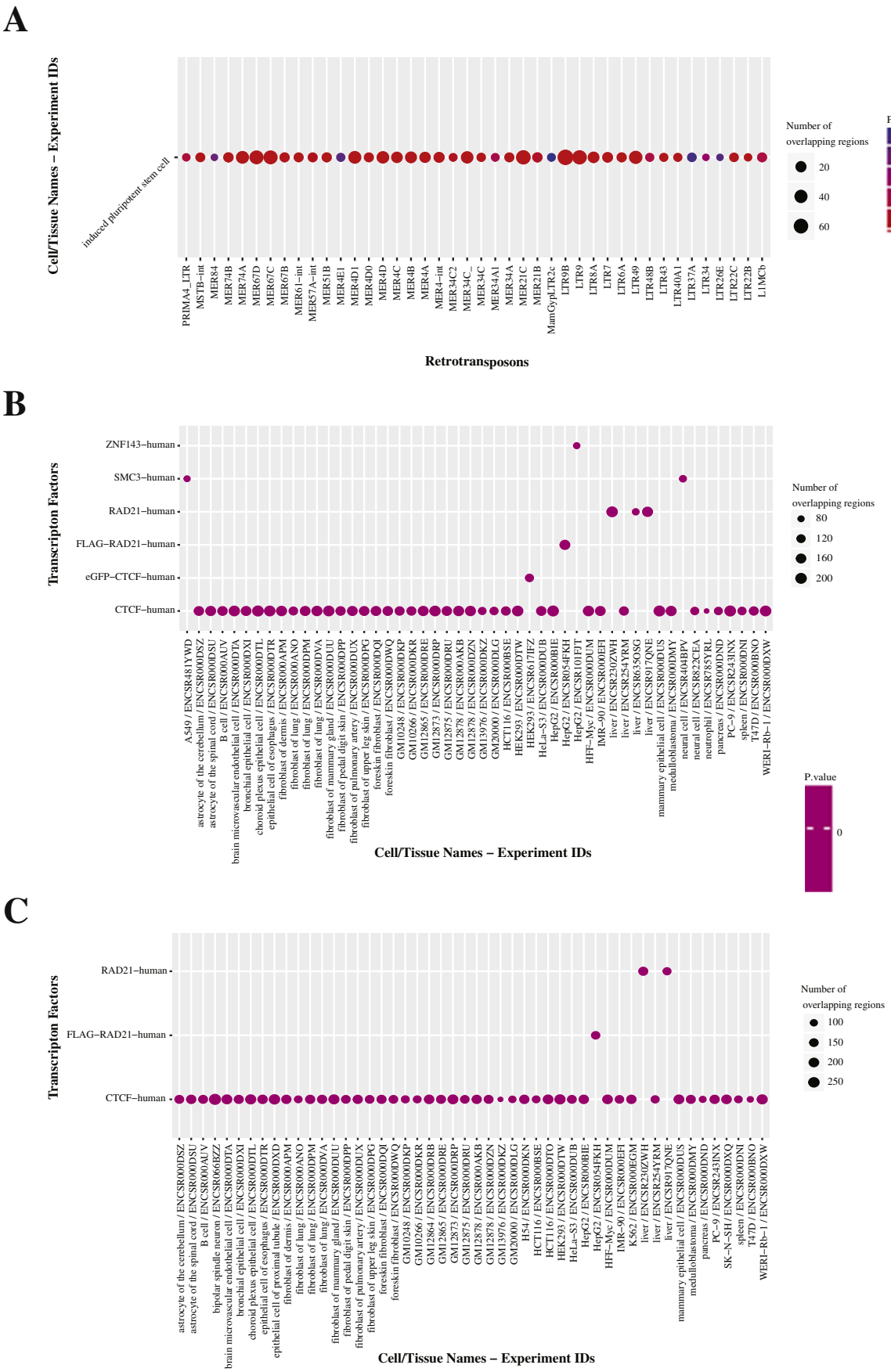


Fig. 4. Dot plot graphs from sample searches of RTFAdb. (a) The graph shows retrotransposons that are statistically overrepresented in the binding sites of human POU5F1. The color scale reflects the level of statistical significance, and the black dots on the upper right side provide information regarding the total number of overlapping regions of retrotransposons in the binding sites of the selected TF. (b–c) The graphs indicate TFs significantly associated with LTR41 and LTR50 retrotransposons, respectively. Both retrotransposons are highly observed in CTCF binding regions.

recommend conservative peak sets with a p-value < 0.0001 threshold for more stringent analysis results.

3.5. Case studies

To examine whether RTFAdb yields reliable results that are in agreement with the current literature, I tested my tool with two well-known retrotransposon—TF associations. For example, when the "Search by Transcription Factor" module of RTFAdb is utilized with the following criteria: (i) organism: Human, (ii) TF: POU5F1 (also known as OCT4), (iii) peak type: Conservative IDR thresholded peaks, and (iv) p-value threshold: 0.0001, the tool retrieves 42 statistically significant associations (p-value < 0.0001) within human-induced pluripotent stem cell, of which the majority are members of the ERV superfamily. My tool detected LTR9B as one of the most significant retrotransposons that is found to be significantly enriched in the TFBSs (Fig. 4a). This result concurs with the previous findings from Kunarso *et al.* [15], which experimentally revealed the association of LTR9B with POU5F1. A 2013 study [29] also indicates that particular specific retrotransposons, such as LTR41 and LTR50, are bound by CTCF. RTFAdb successfully recognized these associations when I performed a search using the "Search by Retrotransposon" module within the conservative IDR thresholded peaks of human ChIP-seq profiles. My search results, consistent with previous findings [29], propose that CTCF is one of the most notable candidates for LTR41 and LTR50 (Fig. 4b and c).

4. Conclusions and future works

In this study, I introduced a unique and pivotal resource called RTFAdb for not only the mobile DNA research community but also for those exploring retrotransposon species putatively providing binding sites for TFs in a specific cell or tissue type. My regularly updated tool, integrating publicly available TF ChIP-seq profiles, is designed to have a user-friendly interface, and I believe that it will support future studies on the regulatory roles of retrotransposon species in human and mouse species. RTFAdb currently includes over 3.000 TF ChIP-seq binding profiles. However, its expandable architecture and my automated analysis pipeline allow the addition of new data sets as they become available. My future work will involve TF ChIP-seq profiles collected from diverse model organisms, and I plan to integrate a new module into RTFAdb that allows the calculation of statistical significance of retrotransposons within a given TF ChIP-seq peak set profile. In conclusion, RTFAdb has the potential to offer novel insights into mammalian transcriptional networks by providing an additional layer of information regarding the regulatory roles of retrotransposons. RTFAdb can be searched online at http://tools.ibg.deu.edu.tr/rtfa/.

Acknowledgements

The author also would like to extend his thanks to Izmir Biomedicine and Genome Institute (iBG-izmir), Dokuz Eylül University, Izmir, Turkey for its support in implementing the project.

References

[1] T. Wicker, F. Sabot, A. Hua-Van, J.L. Bennetzen, P. Capy, B. Chalhoub, A. Flavell, P. Leroy, M. Morgante, O. Panaud, et al., A unified classification system for eukaryotic transposable elements, Nat. Rev. Genet. 8 (2007) 973–982.
[2] H.L. Levin, J.V. Moran, Dynamic interactions between transposable elements and their hosts, Nat. Rev. Genet. 12 (2011) 615–627.
[3] C. Sun, D.B. Shepard, R.A. Chong, J. Lopez Arriaza, K. Hall, T.A. Castoe, C. Feschotte, D.D. Pollock, R.L. Mueller, LTR retrotransposons contribute to genomic gigantism in plethodontid salamanders, Genome Biol. Evol. 4 (2012) 168–183.
[4] R. Cordaux, M.A. Batzer, The impact of retrotransposons on human genome evolution, Nat. Rev. Genet. 10 (2009) 691–703.
[5] A. Peixoto, M. Pinheiro, L. Massena, C. Santos, P. Pinto, P. Rocha, C. Pinto, M.R. Teixeira, Genomic characterization of two large Alu-mediated rearrangements of the BRCA1 gene, J. Hum. Genet. 58 (2013) 78–83.
[6] J.L. Garcia-Perez, T.J. Widmann, I.R. Adams, The impact of transposable elements on mammalian development, Development 143 (2016) 4101–4114.
[7] H. Kano, I. Godoy, C. Courtney, M.R. Vetter, G.L. Gerton, E.M. Ostertag, H.H. Kazazian, L1 retrotransposition occurs mainly in embryogenesis and creates somatic mosaicism, Genes Dev. 23 (2009) 1303–1312.
[8] A. Morales-Hernandez, F.J. Gonzalez-Rico, A.C. Roman, E. Rico-Leo, A. Alvarez-Barrientos, L. Sanchez, A. Macia, S.R. Heras, J.L. Garcia-Perez, J.M. Merino, et al., Alu retrotransposons promote differentiation of human carcinoma cells through the aryl hydrocarbon receptor, Nucleic Acids Res. 44 (2016) 4665–4683.
[9] R.A. Elbarbary, B.A. Lucas, L.E. Maquat, Retrotransposons as regulators of gene expression, Science 351 (2016) aac7247.
[10] G.J. Faulkner, Y. Kimura, C.O. Daub, S. Wani, C. Plessy, K.M. Irvine, K. Schroder, N. Cloonan, A.L. Steptoe, T. Lassmann, et al., The regulated retrotransposon transcriptome of mammalian cells, Nat. Genet. 41 (2009) 563–571.
[11] P. Mita, J.D. Boeke, How retrotransposons shape genome regulation, Curr. Opin. Genet. Dev. 37 (2016) 90–100.
[12] A.E. Peaston, A.V. Evsikov, J.H. Graber, W.N. de Vries, A.E. Holbrook, D. Solter, B.B. Knowles, Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos, Dev. Cell 7 (2004) 597–606.
[13] M. Flemr, R. Malik, V. Franke, J. Nejepinska, R. Sedlacek, K. Vlahovicek, P. Svoboda, A retrotransposon-driven dicer isoform directs endogenous small interfering RNA production in mouse oocytes, Cell 155 (2013) 807–816.
[14] T. Wang, J. Zeng, C.B. Lowe, R.G. Sellers, S.R. Salama, M. Yang, S.M. Burgess, R.K. Brachmann, D. Haussler, Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53, Proc. Natl. Acad. Sci. U. S. A. 104 (2007) 18613–18618.
[15] G. Kunarso, N.Y. Chia, J. Jeyakani, C. Hwang, X.Y. Lu, Y.S. Chan, H.H. Ng, G. Bourque, Transposable elements have rewired the core regulatory network of human embryonic stem cells, Nat. Genet. 42 (2010) (631-U111).
[16] P.E. Jacques, J. Jeyakani, G. Bourque, The majority of primate-specific regulatory sequences are derived from transposable elements, PLoS Genet. 9 (2013).
[17] E.B. Chuong, N.C. Elde, C. Feschotte, Regulatory evolution of innate immunity through co-option of endogenous retroviruses, Science 351 (2016) 1083–1087.
[18] Y. Yasuhiko, Y. Hirabayashi, R. Ono, LTRs of endogenous retroviruses as a source of Tbx6 binding sites, Front. Chem. 5 (2017).
[19] A.B. Conley, I.K. Jordan, Identification of transcription factor binding sites derived from transposable element sequences using ChIP-seq, Computational Biology of Transcription Factor Binding, 674 2010, pp. 225–240.
[20] L. Ramsay, G. Bourque, in silico methods to identify exapted transposable element families, Methods Mol. Biol. 1400 (2016) 33–45.
[21] A. Testori, L. Caizzi, S. Cutrupi, O. Friard, M. De Bortoli, Cora', D., Caselle, M., The role of transposable elements in shaping the combinatorial interaction of transcription factors, BMC Genomics 13 (2012).
[22] B. Simovski, D. Vodak, S. Gundersen, D. Domanska, A. Azab, L. Holden, M. Holden, I. Grytten, K. Rand, F. Drablos, et al., GSuite HyperBrowser: integrative analysis of dataset collections across the genome and epigenome, Gigascience 6 (2017).
[23] I. Dunham, A. Kundaje, S.F. Aldred, P.J. Collins, C. Davis, F. Doyle, C.B. Epstein, S. Frietze, J. Harrow, R. Kaul, et al., An integrated encyclopedia of DNA elements in the human genome, Nature 489 (2012) 57–74.
[24] K.R. Rosenbloom, J. Armstrong, G.P. Barber, J. Casper, H. Clawson, M. Diekhans, T.R. Dreszer, P.A. Fujita, L. Guruvadoo, M. Haeussler, et al., The UCSC genome browser database: 2015 update, Nucleic Acids Res. 43 (2015) D670–681.
[25] A.R. Quinlan, BEDTools: the Swiss-Army tool for genome feature analysis, Current Protocols in Bioinformatics/Editoral Board, Andreas D Baxevanis [et al], 47, 11 12 2014, pp. 11–34.
[26] R.C. Team, R: a Language and Environment For Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2016 (2015. URL: http.www.R-project.org).
[27] W.D. Bao, K.K. Kojima, O. Kohany, Repbase update, a database of repetitive elements in eukaryotic genomes, Mobile DNA-Uk 6 (2015).
[28] H. Wickham, ggplot2: elegant graphics for data analysis, Use R, 2009, pp. 1–212.
[29] P.C. Schwalie, M.C. Ward, C.E. Cain, A.J. Faure, Y. Gilad, D.T. Odom, P. Flicek, Co-binding by YY1 identifies the transcriptionally active, highly conserved set of CTCF-bound regions in primate genomes, Genome Biol. 14 (2013) R148.
[30] G. Karakulah, A. Suner, PlanTEnrichment: a tool for enrichment analysis of transposable elements in plants, Genomics 109 (5–6) (2017) 336–340.