RESEARCH ARTICLE

GREAM: A Web Server to Short-List Potentially Important Genomic Repeat Elements Based on Over-/Under-Representation in Specific Chromosomal Locations, Such as the Gene Neighborhoods, within or across 17 Mammalian Species

Darshan Shimoga Chandrashekar[1,2], **Poulami Dey**[1,2], **Kshitish K. Acharya**[1,3]*

1 Institute of Bioinformatics and Applied Biotechnology (IBAB), Biotech Park, Electronic City, Bengaluru (Bangalore), 560100, Karnataka state, India, **2** Manipal University, Manipal, 576104, Karnataka state, India, **3** Shodhaka Life Sciences Pvt. Ltd., IBAB, Biotech Park, Bengaluru (Bangalore), 560100, Karnataka state, India

* kshitish@ibab.ac.in

OPEN ACCESS

Citation: Chandrashekar DS, Dey P, Acharya KK (2015) GREAM: A Web Server to Short-List Potentially Important Genomic Repeat Elements Based on Over-/Under-Representation in Specific Chromosomal Locations, Such as the Gene Neighborhoods, within or across 17 Mammalian Species. PLoS ONE 10(7): e0133647. doi:10.1371/journal.pone.0133647

Editor: Marc Robinson-Rechavi, University of Lausanne, SWITZERLAND

Received: March 8, 2015

Accepted: June 30, 2015

Published: July 24, 2015

Data Availability Statement: All relevant data is available via Figshare (http://dx.doi.org/10.6084/m9.figshare.1476780) as well as via the GREAM home page (http://resource.ibab.ac.in/GREAM/).

Funding: This work was supported by an institutional grant to Institute of Bioinformatics and Biotechnology (IBAB) from Department of Electronics and Information Technology (DeitY), Government of India, under 'Centre of Excellence for Research and Teaching in Bioinformatics'.

Abstract

Background

Genome-wide repeat sequences, such as LINEs, SINEs and LTRs share a considerable part of the mammalian nuclear genomes. These repeat elements seem to be important for multiple functions including the regulation of transcription initiation, alternative splicing and DNA methylation. But it is not possible to study all repeats and, hence, it would help to short-list before exploring their potential functional significance via experimental studies and/or detailed *in silico* analyses.

Result

We developed the 'Genomic Repeat Element Analyzer for Mammals' (GREAM) for analysis, screening and selection of potentially important mammalian genomic repeats. This web-server offers many novel utilities. For example, this is the only tool that can reveal a categorized list of specific types of transposons, retro-transposons and other genome-wide repetitive elements that are statistically over-/under-represented in regions around a set of genes, such as those expressed differentially in a disease condition. The output displays the position and frequency of identified elements within the specified regions. In addition, GREAM offers two other types of analyses of genomic repeat sequences: a) enrichment within chromosomal region(s) of interest, and b) comparative distribution across the neighborhood of orthologous genes. GREAM successfully short-listed a repeat element (MER20) known to contain functional motifs. In other case studies, we could use GREAM to short-list repetitive elements in the azoospermia factor a (AZFa) region of the human Y chromosome

Competing Interests: The corresponding author declares an affiliation to the private company, Shodhaka Life Sciences Pvt. Ltd. This company supports basic research and his affiliation with this company in no way alters the pure academic nature of the work being reported. Further, he also hereby confirms that 'this affiliation does not alter our adherence to PLOS ONE policies on sharing data and materials'.

and those around the genes associated with rat liver injury. GREAM could also identify five over-represented repeats around some of the human and mouse transcription factor coding genes that had conserved expression patterns across the two species.

Conclusion

GREAM has been developed to provide an impetus to research on the role of repetitive sequences in mammalian genomes by offering easy selection of more interesting repeats in various contexts/regions. GREAM is freely available at http://resource.ibab.ac.in/GREAM/.

Introduction

Repetitive sequences are known to occupy significant portions of the genomes of many complex multicellular organisms [1]. Following earlier indications that all such repeats may not be just ‘junk’ DNA [2–6] researchers have been paying increased attention to these genomic repeat sequences in various contexts, which include gene expression regulation, as described below.

Barbara McClintock indicated the role of 'transposable elements' in ‘genic expression' very early [7, 8]. Involvement of genomic repeat elements in the transcription regulation via defined chromatin loops was suggested based on the high density of transposons in scaffold/matrix associated regions [9]. Multiple experimental evidences for the role of many repeat elements in the regulation of gene expression have accumulated in the recent years. In fact, genomic repeat elements are now indicated to have a significant role in the transcriptional control and/or regulatory networks. For example, Lynch et al [10] found the DNA transposon MER20 to be enriched in the neighborhood of differentially regulated genes in the endometrial stroma and provided evidence for the transcription factor binding abilities of DNA-motifs present within this repeat element. The density of repeat elements and/or their methylation levels have been suggested to influence the regulation of expression of neighboring genes [11–13]. Kunarso et al [14] indicated significant contribution of transposable elements to the occurrence of some of the transcription factor (OCT4 and NANOG) binding sites. They suggested that transposable elements have incorporated new genes into the core regulatory network of embryonic stem cells in humans and mice, and that species-specific transposable elements have substantially altered the transcriptional circuitry of pluripotent stem cells. Lowe et al [15] also suggested a role for certain mobile elements in shaping the gene regulatory networks across mammalian genomes. This suggestion was made because such elements contributed to at least 5.5% of the mammal-exclusive nonexonic conserved elements located in the gene deserts with a strong preference for the neighborhood of genes involved in the development and transcriptional regulation. The MER121 repeat class seems to be conserved in the orthologous genomic locations and may play a cis-regulatory or structural role in mammalian genomes [16]. Genomic repeats have been specifically implicated in stress-responses of plants via gene expression regulation [17, 18]. However, more research efforts are needed to investigate the possible role of the repeat elements in transcriptional regulation and regulatory networks in different species, tissues and conditions.

Genomic repeat elements seem to be important from other perspectives such as evolution, genomic stability, alternative splicing and pathogenicity [19–34]. For example, studying repetitive sequences/elements may help to understand the organization and evolution of eukaryotic genomes [19–22]. Repeat elements have been reported to be associated with some types of cancer [23–24]. Interspersed repeat elements at 'breakpoints' may be relevant to specific reciprocal translocations, which in turn may be central to the pathogenesis of chronic myeloid leukemia [25]. Insertion of the repeat elements have caused in new exons from introns (exonization) or introns from exons (intronization) [26–27] and thus contributed to diversity in terms of alternative splicing in mammals [28–29]. Differential methylation patterns associated with the repeat elements within introns may be associated with the alternative splicing, and seed coat colors in soybean lines [30]. Zabala et al [31] suggested radiation induced temporal changes in methylation within some of the repeat elements in the mouse genome. This study also noted the dependence of such changes on the genetic background (type of strain), gender and the type of the repeat elements.

Thus, apart from the need to explore genomic repeat elements in the context of transcriptional regulation, there is a need to study them in the context of other cellular/molecular processes in mammalian species. But, at present it would be extremely difficult, or even impossible to explore every repeat element present in specific regions such as the neighborhood of a set of genes with interesting and shared expression patterns.

In this context, it is necessary to screen all genomic repeat elements and short-list the interesting ones, which may have a higher chance of being functionally relevant. For instance, it would benefit researchers to select a few repeat elements such as those enriched in the regions around the transcription start sites of a set of differentially transcribed genes. It would then be feasible to explore the types of TFBSs and their abundance within each selected repeat element, hypothesize about their possible functions and/or experimentally test such functional roles. Analyzing the distribution of repeat elements conserved around a set of the orthologous genes can be another type of study where a preliminary short-listing can help. Similarly, it would be interesting to study the repetitive element(s) enriched in two or more species in the neighborhood of genes with constitutive expression in all species considered [see the utility section below for more examples].

There has been a need for bioinformatics support to analyze a variety of possible functions of the repeat elements [35]. Different tools have already been developed for detection and masking analyses of the repeat elements [36]. A comparison of these tools [37] reveals that none of them allows identification of over-/under-represented repeat elements in the neighborhood of a given set of genes. TranspoGene [38] allows querying with multiple gene names and reveals the repeat elements from intronic, exonic and proximal promoter (up to 250 base pairs upstream) regions of the genes. But even this tool cannot be used to analyze the repeat element distribution across genes. Thus, there is a requirement to short-list the repeats with a statistically significant difference in their distribution within the regions of interest, compared with the distribution in the whole genome.

Hence, a new online computational tool, 'Genomic Repeat Element Analyzer for Mammals' (GREAM, http://resource.ibab.ac.in/GREAM) has been developed to identify over-/under-represented elements in, a) the promoters or other neighborhoods across a set of query genes from a species, b) specific chromosomal regions of a species, and c) the promoters or other neighborhoods of orthologous genes from different species.

Materials and Methods

Data collection

RepeatMasker [http://www.repeatmasker.org/] based repeat annotation tracks of the mammalian genome sequences were downloaded from NCBI Genome FTP site (ftp://ftp.ncbi.nlm.nih.gov/genomes/MapView/). Similarly, gene-related information for all 17 mammalian species was obtained from NCBI Gene FTP site (ftp://ftp.ncbi.nih.gov/gene/). Gene identifiers (official gene symbol, Entrez Gene id, Ensembl Gene id, RefSeq mRNA id and UniGene id) corresponding to all mammalian species considered were downloaded from Ensembl BioMart (http://asia.ensembl.org/biomart/martview). Orthologous gene information was downloaded from NCBI HomoloGene Build 67 (ftp://ftp.ncbi.nih.gov/pub/HomoloGene/). Additional information about the repetitive elements was downloaded from Repbase 19.07 (http://www.girinst.org/repbase/). The gene ontology (GO) details and GO-gene association files were downloaded from Gene Ontology Consortium (http://geneontology.org/page/download-ontology, http://geneontology.org/page/download-annotations).

Existing software and programming languages used to create GREAM

Apache 2.2 [http://httpd.apache.org/] was used as a web server on a 64-bit machine with 32 GB RAM and 3 TB HDD space, having Red hat Linux as operating system. MySQL RDBMS [MySQL 5.1.69, http://www.mysql.com/] was used for the data storage. GUI was developed using PERL-CGI scripts. Web technologies such as Cascade Style Sheets (CSS) and JavaScript (JSS) were used for implementing user-friendly features. PHP 5.3.3 was used to create a graphical representation of repetitive element distribution in the neighborhood of all query genes. Open flash chart [http://teethgrinder.co.uk/open-flash-chart/] was used to create the histograms representing relative abundance of the repetitive element classes within the gene neighborhoods.

Obtaining genome-wide frequency of occurrence for every repetitive element based on two types of statistical analyses

The genome-wide statistics for the frequency of occurrence of each repeat element was first calculated in two ways: a) by considering the fraction of the genes that have the specific repeat element in their neighborhood (gene counts); b) by considering the fraction of the specific repeat element among all repeats in the genome (repeat counts).

Genome-wide statistics for a repeat element based on the 'gene counts' is the ratio of the 'number of genes, each of which has that specific repeat element at least once in its neighborhood' to 'all genes'. This ratio can be represented by the following formula:

$$\frac{\textit{Count of genes with the specific repeat in their neighborhood}}{\textit{Total number of genes in the genome}}$$

Genome-wide frequency of a specific repeat element based on ‘repeat counts’ is the ratio of 'occurrence of that repeat element' to the 'occurrence of all repeat elements'—in specific gene neighborhoods. This ratio can be represented by the following formula:

$$\frac{\textit{Count of the specific repeat element around the genes}}{\textit{Count of all repeat elements around the genes}}$$

According to the above-mentioned formulae, PERL scripts were written to calculate the genome-wide frequency of occurrences for all repetitive elements—for each type and size of the gene neighborhoods. The types of gene neighborhood included upstream transcription start site (TSS), downstream of TSS, downstream of gene end, on both sides of the TSS [around TSS] or before and after the gene [around the gene]. In genes known to code for multiple, alternatively spliced transcripts, the 5’-most TSS have been picked. Downstream of the TSS may include one or both UTRs and one or more exons as well as introns. The different sizes considered around the genes or the TSS were 10 kb, 20 kb, 50 kb, 100 kb, 200 kb, 500 kb, 1 mb, 1.5 mb, 2 mb, 2.5 mb and 3 mb. These pre-computed genome-wide frequencies of the repeat elements will be compared with the observed frequency from around the genes in a query set and then its difference with the corresponding genome-wide frequency would be computed. Codes have been written to compute the significance of the difference as described below.

Estimating of over-/under-representation of repetitive elements

GREAM estimates the observed frequency of a specific repeat element for a given gene-set, based on the 'gene counts' as,

$$\frac{Count\ of\ query\ genes\ with\ the\ specific\ repeat\ in\ their\ neighborhood}{Total\ number\ of\ query\ genes}$$

OR based on 'repeat counts' as,

$$\frac{Count\ of\ the\ specific\ repeat\ in\ the\ neighborhood\ of\ all\ query\ genes}{Count\ of\ all\ repeats\ in\ the\ neighborhood\ of\ all\ query\ genes}$$

For both types of analyses, the ratio of the observed frequency of occurrence of each repetitive element in the query region to the corresponding genome-wide frequency would be determined. This ratio indicates the over-/under-representation of each repeat. The significance of the ratio (i.e., the difference between the two frequencies) would be assessed by the binomial probability [39] for each repeat element, using the following formula:

$$P(X = k) = \binom{n}{k} p^k\ (1 - p)^{n-k}$$

In case of the analysis using the 'gene counts', 'p' is the genome-wide frequency of occurrence (described in the previous section) of a specific repeat element, 'n' is the total number of query genes submitted and 'k' is the total number of query genes having the repeat element at least once in their neighborhood.

In case of the analysis by using the 'repeat counts', 'p' is the genome-wide frequency of occurrence (described in the previous section) of a specific repeat element, 'n' is the total occurrence of all repetitive elements in the neighborhood of all query genes (for 'Analyze gene-set' and 'Analyze orthologous gene-set' feature) or in the specified chromosomal location (for 'Analyze locus' feature), and 'k' is the total occurrence of a specific repeat element in the neighborhood of all query genes (for 'Analyze gene-set' and 'Analyze orthologous gene-set' feature) or in a specific chromosomal location (for 'Analyze locus' feature).

If the calculated probability value (P) is less than 0.05, over-/under-representation of specific repetitive element was considered to be significant.

Estimation of repeat element coverage on mammalian genomes

Genome size for human and mouse was obtained from the genome reference consortium [http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/], while for other species the information was obtained from Ensembl genome browser [40]. The size of the genome covered by repeat elements was computed by parsing the repeat annotation tracks of specific genome assemblies (downloaded from NCBI Genome ftp).

Transcription factor binding site (TFBS) analysis of repetitive element sequences

Sequences of the repetitive elements obtained from Repbase 19.07 were screened for potential TFBSs using the CLOVER tool (http://zlab.bu.edu/clover/) [41]. Position weight matrices of

Table 1. The contribution of repeat elements to nuclear genomes in 17 mammalian species.

Species	Genome assembly	Total size (in nucleotides)	Repeat coverage (in nucleotides)	Percentage of genome covered by repeat elements
Human	GRCh37.p5-Primary Assembly	3,101,788,170	1,441,122,130	46.46
Mouse	GRCm38-C57BL/6J	2,730,855,475	1,186,206,269	43.44
Rat	Rnor_5.0-Primary Assembly	2,573,362,844	1,103,002,536	42.86
Cow	Btau_4.6.1-Primary Assembly	2,649,685,036	1,347,191,424	50.84
Chimpanzee	Pan_troglodytes-2.1.4-Primary Assembly	2,995,917,117	1,458,594,710	48.69
Rabbit	OryCun2.0-Primary Assembly	2,604,023,284	1,115,709,440	42.85
Dog	CanFam3.1-Primary Assembly	2,392,715,236	1,004,202,711	41.97
Elephant	Loxafr3.0-Primary Assembly	3,118,565,340	1,456,307,490	46.70
Gibbon	Nleu_3.0-Primary Assembly	2,756,591,777	1,392,833,579	50.53
Gorilla	gorGor3.1-Primary Assembly	2,828,888,833	1,344,286,159	47.52
Horse	EquCab2.0-Primary Assembly	2,428,790,173	1,046,969,418	43.11
Monkey	Mmul_051212-Primary Assembly	3,093,871,206	1,466,974,355	47.42
Marmoset	Callithrix jacchus-3.2-Primary Assembly	2,759,289,125	1,301,711,202	47.18
Opposum	MonDom5-Primary Assembly	3,501,660,299	1,951,703,622	55.74
Orangutan	P_pygmaeus_2.0.2-Primary Assembly	3,109,347,532	1,563,466,038	50.28
Pig	Sscrofa10.2-Primary Assembly	3,024,658,544	1,040,672,576	34.41
Platypus	Ornithorhynchus_anatinus-5.0.1-Primary Assembly	1,917,748,604	889,518,843	46.38

doi:10.1371/journal.pone.0133647.t001

TFBSs are downloaded from JASPAR 2014 (http://jaspardev.genereg.net/) [42]. The locations of potential TFBSs were listed using CLOVER (with default threshold score).

Case study for validation

Lynch et al [10] reported association of a DNA transposon, MER20 in rewiring the regulatory network of differentially regulated stromal genes. They reported enrichment of MER20 within 200KB of each end of differentially expressed genes, identified TFBSs within MER20 (using TRANSFAC database), and confirmed their transcription factor binding ability via ChIP-qPCR experiments. Hence, MER20 formed a good case for validating the efficiency of GREAM; it is perhaps the only available repeat with such evidences. We downloaded processed RNA-seq data for differentiated and undifferentiated human endometrial stromal cells from GEO (GSE30708), obtained 1,149 differentially expressed genes having a read count of >20 in either differentiated or undifferentiated endometrial stromal cells as mentioned in the original study, and analyzed the repeat element distribution in the same neighborhood region (around gene, 200KB). Repeat enrichment was analyzed using statistics based on 'gene counts'.

Results and Discussion

The distribution of repeat elements in the mammalian genomes was first reviewed by quantifying them in the latest genome-builds available (Table 1). 'Annotated repeat elements' constitute 34 to 55% of known mammalian genomes. Non-mammalian genomes seem to have significantly lesser proportion of such repeats [43]. But there can be exceptions. For example, repeat elements contribute to 77% of the genome in *Rana sculenta* [1, 44]. In general, the distributions noted by us were similar to the observation made earlier by other researchers for human and mouse species [1, 44]. The estimations may vary to some extent based on the method used. Compared to the conventional method (Repeat Masker, which involves aligning the repeat

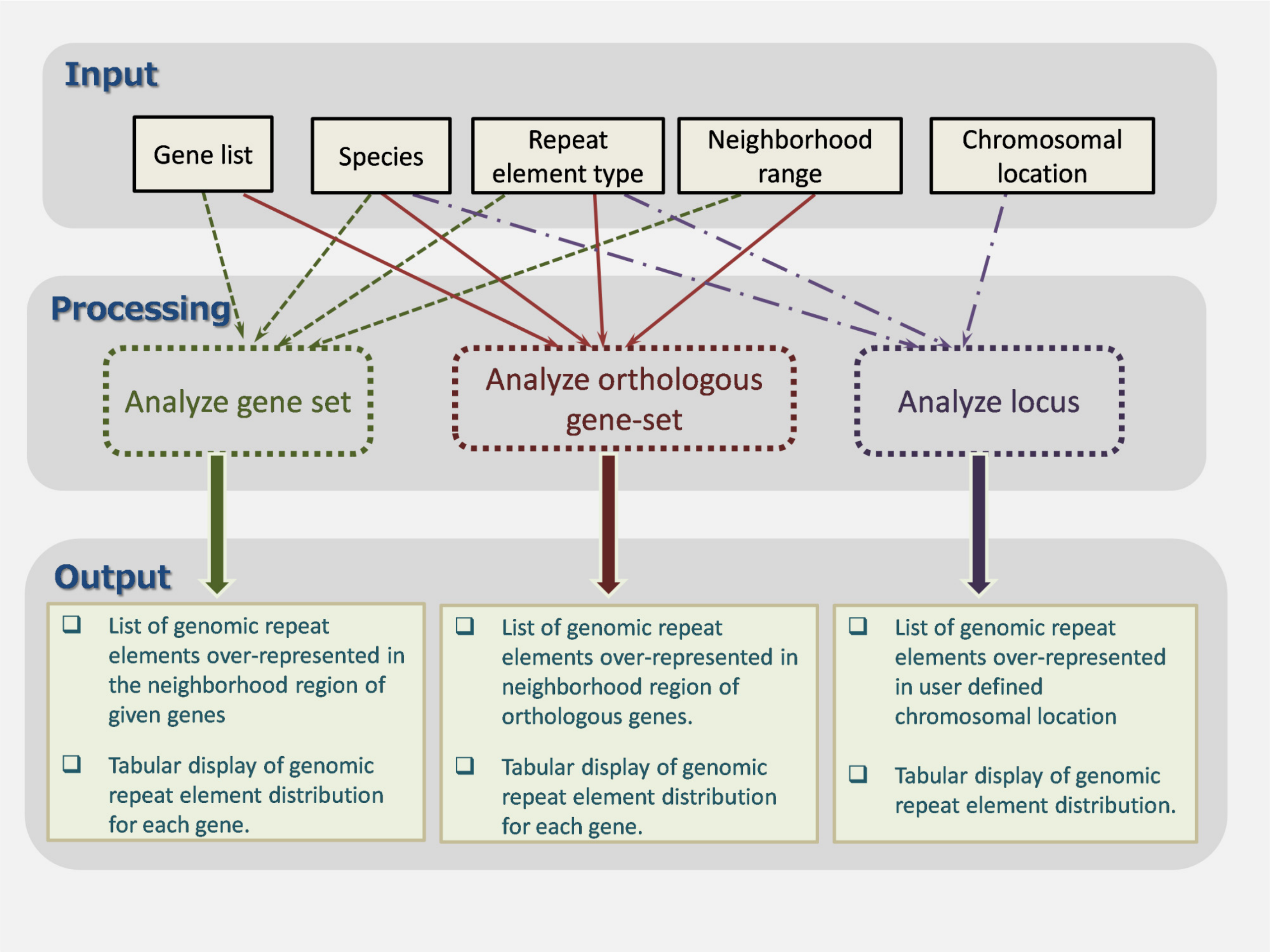

Input
Gene list
Species
Repeat element type
Neighborhood range
Chromosomal location
Processing
Analyze gene set
Analyze orthologous gene-set
Analyze locus
Output
❑ List of genomic repeat elements over-represented in the neighborhood region of given genes
❑ Tabular display of genomic repeat element distribution for each gene.
❑ List of genomic repeat elements over-represented in neighborhood region of orthologous genes.
❑ Tabular display of genomic repeat element distribution for each gene.
❑ List of genomic repeat elements over-represented in user defined chromosomal location
❑ Tabular display of genomic repeat element distribution.

Fig 1. Workflow of the GREAM web server.

doi:10.1371/journal.pone.0133647.g001

consensus sequences with the genome) of repeat element identification, a novel *de novo* strategy (P-cloud) estimated higher (i.e. 69%) repeat element coverage of the human genome [45].

Overview of GREAM

In Fig 1, the overall workflow of GREAM is schematically represented.

GREAM allows three types of repeat element analysis: 1) the feature 'Analyze gene-set' facilitates analysis of the repeat element distribution within the neighborhood of a query gene-set; 2) the feature 'Analyze orthologous gene-set' facilitates analysis of the repeat element distribution in the neighborhood of a query gene-set and their orthologous genes; 3) the feature 'Analyze locus' facilitates analysis of the repeat element distribution within specific chromosomal location(s).

Input data. GREAM requires a set of genes as an input for both 'Analyze gene-set' and 'Analyze orthologous gene-set' features. User can either paste the gene identifiers in the text area provided or upload the gene list as a text file. Official gene symbols, Entrez gene identifiers, Ensembl gene identifiers, RefSeq mRNA identifiers or the Unigene identifiers can be used to upload the gene-list. But the submitted query list should be homogeneous. The species can be chosen from the options provided in a drop down menu. 'Analyze gene-set' can be used for 17 mammalian species. 'Analyze locus' can be used for all species except elephants and gibbons as the contigs were not assembled into chromosomes for these two species. 'Analyze orthologous gene-set' can be used for only seven species (human, mouse, rat, cow, dog, chimpanzee and monkey) as HomoloGene includes orthologous gene information exclusively for these mammalian species.

The tool allows user to specify the gene neighborhood and the population statistics to be considered ('gene counts' or 'repeat counts') for short-listing important repeat elements. In case of 'Analyze orthologous gene-set', additional species need to be selected. For 'Analyze locus' feature, GREAM requires specification of a) a chromosomal location (for example, Chr1:3011498–3120000 or Chr4|NT_113885.1:5000–100000 or ChrUn|NT_167210.1:100–150000), b) mammalian species and c) type of genomic repeat element of interest ('all' repeats scanned with default settings).

There is an option within each of the three types of analyses for users to provide a title for the job submitted to keep track of their own analysis, and also to use e-mail address for receiving a link for the output on completion of submitted job.

Job handling. Each submitted job will be stored and processed based on the order of submission using a PERL script that runs on the server. User can monitor the submitted job's status (which can be either "in queue", "being processed" or "completed") through the 'job status' page. On completion of the job, the user will be redirected to the result page. In addition, a link to the result page will be sent to the e-mail address–if provided. The result of the analysis of a specific job can be accessed till one week of its submission date.

Processing input. In 'Analyze gene-set', the query genes are first subjected to gene identifier conversions, using a local copy of the Ensembl Biomart, to obtain corresponding NCBI Entrez gene identifiers. Their chromosomal locations are then obtained using the gene annotation track from NCBI Gene and based on the input specifications, the neighborhood to be analyzed are obtained for all query genes. Using the repeat annotation tracks downloaded from NCBI, the distribution of all or user predefined repeat elements within the neighborhood regions of all query genes will be obtained. Observed frequency of each repeat element is compared with its pre-computed genome-wide frequency, for the user-specified statistical parameters ('gene counts' or 'repeat counts').

In 'Analyze locus', the user predefined chromosomal location will be scanned for the presence of all or selected repetitive elements using repeat annotation tracks downloaded from NCBI. Then, the observed frequency of occurrence of each repeat element in the given region will be compared with their genome-wide frequency of occurrence within the entire genome. The analysis is based on 'repeat counts' statistics.

In 'Analyze orthologous gene-set', the gene identifier conversion of query genes will be performed as mentioned before. A local copy of the 'HomoloGene' database will then be used for obtaining orthologous genes corresponding to the query gene-set and the selected mammalian species. Gene neighborhood of query gene-set and each of orthologous gene-sets will be scanned separately for all or user predefined repetitive elements. The observed frequency for the query-set will be compared with the genome-wide frequency of occurrence of each repeat element, for each species as per user-specified statistics ('based on gene counts' or 'based on repeat counts'). Finally, the repeat elements that show significant over-/under-representation in the query gene-set and its orthologous gene-sets from at least one or other species are identified.

In all cases, the binomial probability (see methods for more detail) will be used to calculate the significance of difference of the frequency in the query regions versus the comparable genome-wide frequency of occurrence of each repeat element.

Output. 'Analyze gene-set' feature results in a table of the repeat elements and their location within neighborhood and a flash chart showing the percentage of detected repeat elements of different types for each queried gene. Also a summary table will be displayed with details of the repeat elements that are statistically over-/under-represented within the specified size and type of neighborhood. GREAM allows users to visualize the distribution of one or more significant repeat elements within the regions of interest. This feature is available when a user chooses one of the following as the type of neighborhood region: 'around TSS', 'upstream of TSS', 'downstream of TSS' or 'downstream of gene end'.

In 'Analyze locus' feature, the output page includes a table for the repeat elements and their location within the queried chromosomal location and a flash chart showing percentage of the detected repeat elements of different types. A summary table provides details of the repeat elements that are statistically over-/under-represented within the given chromosomal location.

In 'Analyze orthologous gene-set' feature, similar results are provided for all queried genes and their orthologous counterparts of chosen mammalian species: a) a table of the repeat elements and their location within the neighborhood and b) a flash chart showing percentage of the detected repeat elements of different types. Summary tables reveal the details of the repeat elements that are statistically over-/under-represented within specific regions around the query genes and the corresponding orthologous genes.

Each repeat element represented in tables will be hyperlinked to a web page showing further details—including the sequence and potential transcription factor binding sites within the sequence. GREAM facilitates ontology analysis of genes that have enriched repeat elements in their neighborhood.

Validating GREAM using a case study

Lynch et al [10] showed that MER20, a DNA transposon is enriched within a 200kb neighborhood of endometrially expressed genes. They suggested the involvement of MER20 in co-ordinating gene expression regulation via the TFBSs harbored within it. By using 'Analyze gene-set' feature of GREAM, we investigated the neighborhood of 1,149 differentially expressed genes obtained from their study. The goal was to find out if GREAM can identify one of the repeat elements detected independently by another group of researchers, and also known to harbor TFBSs with proven TF-binding capacities. GREAM indeed found MER20 to be significantly (p-value: 0.0234) over-represented in the neighborhood of 943 (82%) of the query genes.

It is important to note two aspects of GREAM's validation: a) the tool short-listed a repeat element that was independently identified earlier using a different statistical approach (Yates corrected chi-square test used by Lynch et al [10], as against the 'binomial probability of repeat counts' employed here); b) the detected repeat element was already shown to have TFBSs with functional evidences via ChIP-qPCR [10]

GREAM could detect other potentially important repeats. In fact, 21 repeat elements were over-represented with p-value < = 0.0234, in the regions studied, including L1MB7, MIR3, L1ME4a, L1MC4 and MLT1D with p-value 0.0001 or less (see S1 Table for more details). Our results indicated the probability of a larger number of repeat elements in addition to MER20, having a functional role in regulating transcription of the pregnancy-related genes investigated originally [10] and selected for the analysis in our study. But it should be noted that there could be false positives among the repeats short-listed by the tool. Further experimental investigations are needed to confirm or reject the relevance of genomic repeats indicated by GREAM.

Utility of GREAM

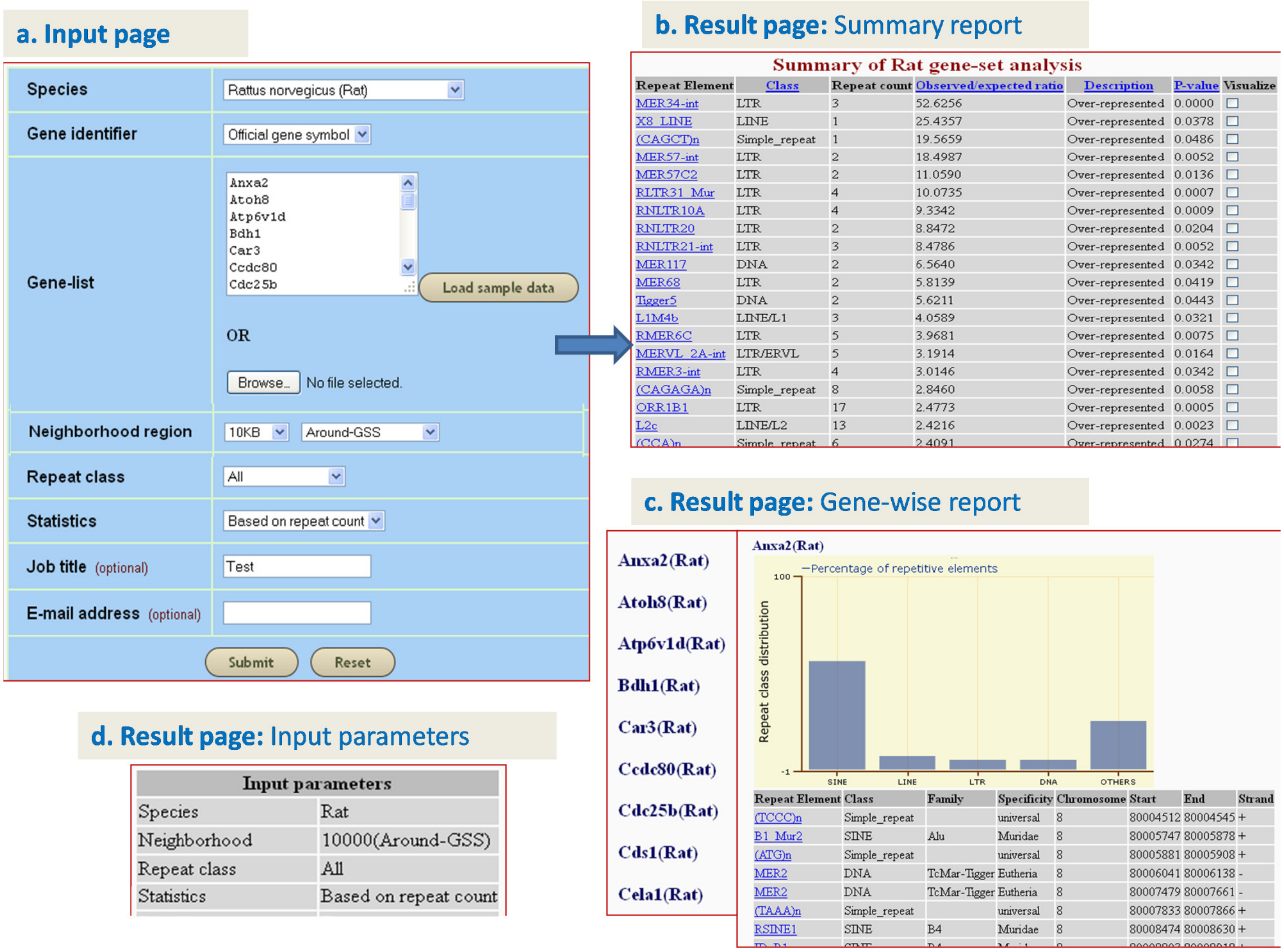
The above-mentioned case study illustrates the use of GREAM for short-listing the genomic repeat elements around genes that may have a functional significance. Thus, the tool can be used to short-list important repeat elements for detailed *in silico* analysis/experimentation. The following questions represent more examples of research problems that can be addressed using GREAM:

1. Are there repetitive elements associated with the epigenetic regulation of promoters of certain genes sharing functionalities or expression pattern?
2. Is there an abundance of any specific type of repeat element in the neighborhood of chromosomal regions that are frequently associated with genetic recombination or a specific type of structural abnormality?
3. Are there X- or Y-chromosome-associated repeat elements preserved across mammalian species and, if yes, in which regions?
4. Are there specific type of repeat elements around genes that are near telomeres and/or centromeres?

We describe three case studies for a better illustration of the utilities (the repeat elements detected for each of these case studies are listed in supporting information:

A. Analyzing the repeat element distribution in the neighborhood of rat genes associated with general liver injury

'Analyze gene-set' option was used to identify over-represented repeat elements in the neighborhood of 64 rat genes reported to be associated with the liver injury based on gene expression profile [46]. Input parameters such as species (rat), gene identifier (official gene symbol), neighborhood region (10 kb nt, around TSS), repeat class (all), statistics (based on 'repeat counts') and job name ('rat-liver-injury-set') were provided in addition to the list of 64 genes [Fig 2A]. On submission, the job was scheduled for the processing, and after processing the result page revealed the following: 1) 'Summary report' table with a list of 81 significant repeat elements (65 over-represented and 16 under-represented) from the neighborhood of the 64 query genes. The significance of many of the over-represented repeats was remarkable with very low p-values. For example, the top five repeats were over-represented with a p-value of less than 10^{-4} and included three LTR retrotransposons (ERVB1_2-LTR_RN, MER34-int, RLTR18-int) two DNA transposons (Zaphod2, Ricksha_c). On the other hand, even the top five under-represented repeats (LINEs: Lx, Lx5, Lx8b, L1_Rn and L1_Rn2) had only a moderate p-value ($p<0.021$) (S4and S5Tables). When the analysis was repeated with the statistics option of 'gene counts', 45 over-represented and 3 under-represented repeat elements were found, with moderate p-values. While the three under-represented repeat elements were LINEs (Lx, Lx8b and Lx5) ($p<0.003$), the top five ($p< = 0.006$) over-represented repeat elements included LTR retrotransposons (RLTR31_Mur, MTEa and MER67B), a SINE element (ID4_) and a LINE element (L2c). Of these 48 elements, 43 were the same as those obtained with 'repeat counts' option (S6and S7Tables). The table showed observed/expected ratio of repeat elements, corresponding p-values and count of each repeat type [Fig 2B] 2) 'Gene-wise report', with an account of the repeat element distribution in the neighborhood of each of 64 genes submitted in the form of histogram and table [Fig 2C]. The results page also display the 'Input parameters' provided initially [Fig 2D].



a. Input page

Species	Rattus norvegicus (Rat)
Gene identifier	Official gene symbol
Gene-list	Anxa2 Atoh8 Atp6v1d Bdh1 Car3 Ccdc80 Cdc25b OR Browse... No file selected.
Neighborhood region	10KB Around-GSS
Repeat class	All
Statistics	Based on repeat count
Job title (optional)	Test
E-mail address (optional)	

b. Result page: Summary report

Summary of Rat gene-set analysis

Repeat Element	Class	Repeat count	Observed/expected ratio	Description	P-value	Visualize
MER34-int	LTR	3	52.6256	Over-represented	0.0000	□
X8_LINE	LINE	1	25.4357	Over-represented	0.0378	□
(CAGCT)n	Simple_repeat	1	19.5659	Over-represented	0.0486	□
MER57-int	LTR	2	18.4987	Over-represented	0.0052	□
MER57C2	LTR	2	11.0590	Over-represented	0.0136	□
RLTR31_Mur	LTR	4	10.0735	Over-represented	0.0007	□
RNLTR10A	LTR	4	9.3342	Over-represented	0.0009	□
RNLTR20	LTR	2	8.8472	Over-represented	0.0204	□
RNLTR21-int	LTR	3	8.4786	Over-represented	0.0052	□
MER117	DNA	2	6.5640	Over-represented	0.0342	□
MER68	LTR	2	5.8139	Over-represented	0.0419	□
Tigger5	DNA	2	5.6211	Over-represented	0.0443	□
L1M4b	LINE/L1	3	4.0589	Over-represented	0.0321	□
RMER6C	LTR	5	3.9681	Over-represented	0.0075	□
MERVL_2A-int	LTR/ERVL	5	3.1914	Over-represented	0.0164	□
RMER3-int	LTR	4	3.0146	Over-represented	0.0342	□
(CAGAGA)n	Simple_repeat	8	2.8460	Over-represented	0.0058	□
ORR1B1	LTR	17	2.4773	Over-represented	0.0005	□
L2c	LINE/L2	13	2.4216	Over-represented	0.0023	□
(CCA)n	Simple_repeat	6	2.4091	Over-represented	0.0274	□

c. Result page: Gene-wise report

Anxa2(Rat)
Atoh8(Rat)
Atp6v1d(Rat)
Bdh1(Rat)
Car3(Rat)
Ccdc80(Rat)
Cdc25b(Rat)
Cds1(Rat)
Cela1(Rat)

Anxa2(Rat)

Repeat Element	Class	Family	Specificity	Chromosome	Start	End	Strand
(TCCC)n	Simple_repeat		universal	8	80004512	80004545	+
B1_Mur2	SINE	Alu	Muridae	8	80005747	80005878	+
(ATG)n	Simple_repeat		universal	8	80005881	80005908	+
MER2	DNA	TcMar-Tigger	Eutheria	8	80006041	80006138	-
MER2	DNA	TcMar-Tigger	Eutheria	8	80007479	80007661	-
(TAAA)n	Simple_repeat		universal	8	80007833	80007866	+
RSINE1	SINE	B4	Muridae	8	80008474	80008630	+

d. Result page: Input parameters

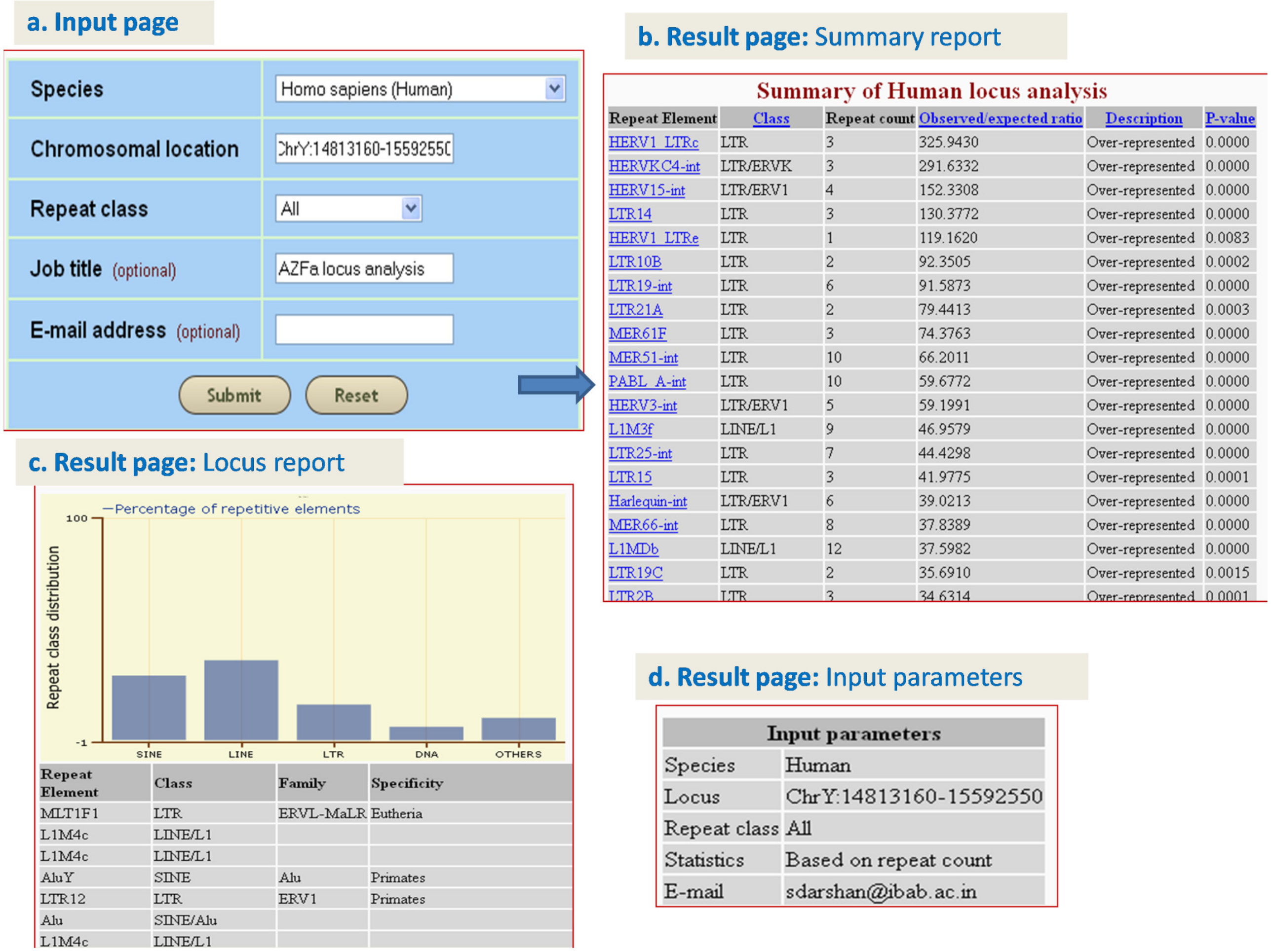
Input parameters	
Species	Rat
Neighborhood	10000(Around-GSS)
Repeat class	All
Statistics	Based on repeat count

Fig 2. Snapshots of GREAM illustrating use of 'analyze gene-set' option on a gene-set from Tawa et al [46].

doi:10.1371/journal.pone.0133647.g002

B. Analyzing the repeat element distribution in AZFa region of human Y chromosome

'Analyze locus' option was used to identify the repeat elements that are enriched in the AZFa locus region of human Y chromosome. This locus, which has 3 genes (USP9Y, DBY and UTY) that directly influence male fertility [47] was submitted as a query (chromosome Y: 14813160 to 15592550). Additional parameters such as species (human), repeat class (All) and job name ('AZFa locus analysis') were specified [Fig 3A]. 'Summary report' table was displayed listing 111 repeat elements (91 over-represented and 20 under-represented) in the region of interest along with the counts of repeat elements, their observed/expected ratios and corresponding P-values [Fig 3B]. 'Locus report' provided distribution of the repeat elements in AZFa region in the form of histogram and table [Fig 3C], and 'Input parameters' were also displayed [Fig 3D]. The results included 35 over-represented repeats with significant p-value ($p<10^{-4}$). Examples include LTR retrotransposons (HERV1_LTRc, HERVKC4-int, HERV15-int, LTR14, LTR19-int), LINEs (L1M3f, L1MDb, L1M3c, L1M4c, L1MB4, L1MCa) and DNA transposons (Tigger4, Tigger7, Tigger2). Similarly, the under-represented elements included two LINEs (L2a and L2c) and three SINEs (MIR, MIRb and MIR3) with $p<10^{-4}$ (S2and S3Tables).



a. Input page

Species	Homo sapiens (Human)
Chromosomal location	ChrY:14813160-1559255
Repeat class	All
Job title (optional)	AZFa locus analysis
E-mail address (optional)	

b. Result page: Summary report

Summary of Human locus analysis

Repeat Element	Class	Repeat count	Observed/expected ratio	Description	P-value
HERV1_LTRc	LTR	3	325.9430	Over-represented	0.0000
HERVKC4-int	LTR/ERVK	3	291.6332	Over-represented	0.0000
HERV15-int	LTR/ERV1	4	152.3308	Over-represented	0.0000
LTR14	LTR	3	130.3772	Over-represented	0.0000
HERV1_LTRe	LTR	1	119.1620	Over-represented	0.0083
LTR10B	LTR	2	92.3505	Over-represented	0.0002
LTR19-int	LTR	6	91.5873	Over-represented	0.0000
LTR21A	LTR	2	79.4413	Over-represented	0.0003
MER61F	LTR	3	74.3763	Over-represented	0.0000
MER51-int	LTR	10	66.2011	Over-represented	0.0000
PABL_A-int	LTR	10	59.6772	Over-represented	0.0000
HERV3-int	LTR/ERV1	5	59.1991	Over-represented	0.0000
L1M3f	LINE/L1	9	46.9579	Over-represented	0.0000
LTR25-int	LTR	7	44.4298	Over-represented	0.0000
LTR15	LTR	3	41.9775	Over-represented	0.0001
Harlequin-int	LTR/ERV1	6	39.0213	Over-represented	0.0000
MER66-int	LTR	8	37.8389	Over-represented	0.0000
L1MDb	LINE/L1	12	37.5982	Over-represented	0.0000
LTR19C	LTR	2	35.6910	Over-represented	0.0015
LTR2B	LTR	3	34.6314	Over-represented	0.0001

c. Result page: Locus report

Repeat Element	Class	Family	Specificity
MLT1F1	LTR	ERVL-MaLR	Eutheria
L1M4c	LINE/L1		
L1M4c	LINE/L1		
AluY	SINE	Alu	Primates
LTR12	LTR	ERV1	Primates
Alu	SINE/Alu		
L1M4c	LINE/L1		

d. Result page: Input parameters

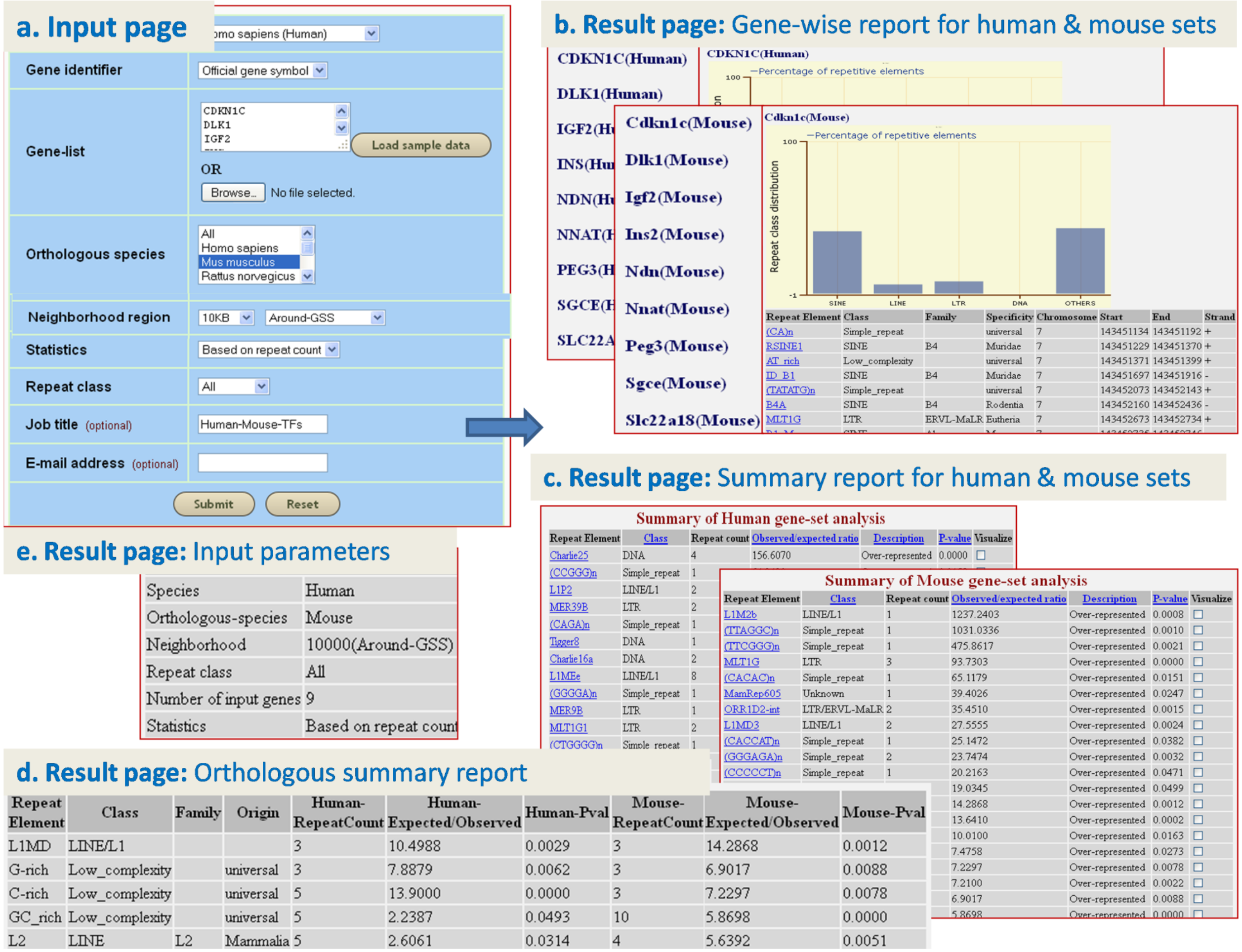
Input parameters	
Species	Human
Locus	ChrY:14813160-15592550
Repeat class	All
Statistics	Based on repeat count
E-mail	sdarshan@ibab.ac.in

Fig 3. Snapshots of GREAM illustrating use of 'analyze locus' option on AZFa region.

doi:10.1371/journal.pone.0133647.g003

C. Analyzing the repeat element distribution in the neighborhood of genes coding for transcription factors and showing conserved expression patterns across human and mouse

Steinhoff et al [48], found that the levels of 9 transcription factor-coding mRNAs (CDKN1C, DLK1, IGF2, INS, NDN, NNAT, PEG3, SGCE and SLC22A18) had a positive correlation across 22 human and mouse tissues. We used 'analyze orthologous gene-set' feature of GREAM to identify the repeat elements enriched in the neighborhood of the orthologs of interest. The selected 9 genes were input along with other parameters such as species (human), gene identifier (gene symbol), orthologous species (mouse), neighborhood region (10 kb nt, around TSS), statistics (based on 'repeat counts'), repeat class (all) and job name ('Human-Mouse-TFs') [Fig 4A]. The 'Summary report' table for the human gene-set listed 24 over-represented repeat elements and 3 under-represented ones, while the table for mouse orthologs listed 25 over-represented repeat elements and 3 under-represented [Fig 4B]. 'Gene-wise report' for human and mouse gene-set showed distribution of repeat elements in neighborhood of individual genes in the form of histogram and table [Fig 4C]. 'Orthologous summary report' listed 5 repeat elements (L1MD, G-rich, C-rich, GC_rich and L2) commonly enriched in the human gene-set and their mouse ortholog gene-set [Fig 4D], and 5) 'Input parameters', were also



	ɔmo sapiens (Human)
Gene identifier	Official gene symbol
Gene-list	CDKN1C DLK1 IGF2 Load sample data OR Browse... No file selected.
Orthologous species	All Homo sapiens Mus musculus Rattus norvegicus
Neighborhood region	10KB Around-GSS
Statistics	Based on repeat count
Repeat class	All
Job title (optional)	Human-Mouse-TFs
E-mail address (optional)	

Repeat Element	Class	Family	Specificity	Chromosome	Start	End	Strand
(CA)n	Simple_repeat		universal	7	143451134	143451192	+
RSINE1	SINE	B4	Muridae	7	143451229	143451370	+
AT_rich	Low_complexity		universal	7	143451371	143451399	+
ID_B1	SINE	B4	Muridae	7	143451697	143451916	-
(TATATG)n	Simple_repeat		universal	7	143452073	143452143	+
B4A	SINE	B4	Rodentia	7	143452160	143452436	-
MLT1G	LTR	ERVL-MaLR	Eutheria	7	143452673	143452734	+

Summary of Human gene-set analysis

Repeat Element	Class	Repeat count	Observed/expected ratio	Description	P-value	Visualize
Charlie25	DNA	4	156.6070	Over-represented	0.0000	□
(CCGGG)n	Simple_repeat	1				
L1P2	LINE/L1	2				
MER39B	LTR	2				
(CAGA)n	Simple_repeat	1				
Tigger8	DNA	1				
Charlie16a	DNA	2				
L1MEe	LINE/L1	8				
(GGGA)n	Simple_repeat	1				
MER9B	LTR	1				
MLT1G1	LTR	2				
(CTGGG)n	Simple_repeat	1				

Summary of Mouse gene-set analysis

Repeat Element	Class	Repeat count	Observed/expected ratio	Description	P-value	Visualize
L1M2b	LINE/L1	1	1237.2403	Over-represented	0.0008	□
(TTAGGC)n	Simple_repeat	1	1031.0336	Over-represented	0.0010	□
(TTCGGG)n	Simple_repeat	1	475.8617	Over-represented	0.0021	□
MLT1G	LTR	3	93.7303	Over-represented	0.0000	□
(CACAC)n	Simple_repeat	1	65.1179	Over-represented	0.0151	□
MamRep605	Unknown	1	39.4026	Over-represented	0.0247	□
ORR1D2-int	LTR/ERVL-MaLR	2	35.4510	Over-represented	0.0015	□
L1MD3	LINE/L1	2	27.5555	Over-represented	0.0024	□
(CACCAT)n	Simple_repeat	1	25.1472	Over-represented	0.0382	□
(GGGAGA)n	Simple_repeat	2	23.7474	Over-represented	0.0032	□
(CCCCCT)n	Simple_repeat	1	20.2163	Over-represented	0.0471	□
			19.0345	Over-represented	0.0499	□
			14.2868	Over-represented	0.0012	□
			13.6410	Over-represented	0.0002	□
			10.0100	Over-represented	0.0163	□
			7.4758	Over-represented	0.0273	□
			7.2297	Over-represented	0.0078	□
			7.2100	Over-represented	0.0022	□
			6.9017	Over-represented	0.0088	□
			5.8698	Over-represented	0.0000	□

Species	Human
Orthologous-species	Mouse
Neighborhood	10000(Around-GSS)
Repeat class	All
Number of input genes	9
Statistics	Based on repeat count

Repeat Element	Class	Family	Origin	Human-RepeatCount	Human-Expected/Observed	Human-Pval	Mouse-RepeatCount	Mouse-Expected/Observed	Mouse-Pval
L1MD	LINE/L1			3	10.4988	0.0029	3	14.2868	0.0012
G-rich	Low_complexity		universal	3	7.8879	0.0062	3	6.9017	0.0088
C-rich	Low_complexity		universal	5	13.9000	0.0000	3	7.2297	0.0078
GC_rich	Low_complexity		universal	5	2.2387	0.0493	10	5.8698	0.0000
L2	LINE	L2	Mammalia	5	2.6061	0.0314	4	5.6392	0.0051

Fig 4. Snapshots of GREAM illustrating use of 'analyze orthologous gene-set' on human transcription factor genes and their mouse orthologs from Steinhoff et al [48].

doi:10.1371/journal.pone.0133647.g004

displayed [Fig 4E]. Result details are available at S8, S9, S12, S13 and S16 Tables. When the analysis was repeated with the statistics option of 'gene counts', for the human gene-set, 13 over-represented and 5 under-represented repeat elements were found. Of these 18 repeats, 16 were the same as those obtained using the 'repeat counts' option. In case of mouse ortholog gene-set, 13 over-represented and 4 under-represented repeat elements were found. Of these 17 repeats, 16 were the same as those obtained using the 'repeat counts' option for mouse. G-rich and C-rich elements were found to be enriched in the human gene-set and their mouse ortholog gene-set. These two were found to be common among human and mouse gene-sets when 'repeat counts' option was used. Result details are available at S10, S11, S14, S15 and S17 Tables.

Limitations

While short-listing the statistically over-/under-represented repeat elements is one way of finding potentially functional ones, many such short-listed elements may not be really associated with functions. But, in absence of alternatives, this approach can assist scientists in exploring repetitive elements. Considering short-listed repeats with stringent p-value cut offs such as $p<0.005$, may reduce the number of false positives obtained from GREAM.

The time taken by GREAM for processing the jobs submitted is directly proportional to the size of the gene-set, neighborhood regions considered and type of statistics used. Analysis using the 'repeat counts' option usually takes longer time than 'gene counts' option. Analyses with larger regions (1 mb and above) can take a few minutes to hours, depending on the number of genes (please see user guide of the tool for more details).

Conclusions

In the context of exploring multiple possible functional roles, GREAM would be useful for selecting important genomic repeat elements present around specific genes or in chromosomal regions of interest. With the help of this web server, many researchers can now raise new questions and/or progress faster in their investigation of the potential functions of repetitive sequences.

Supporting Information

S1 Table. Summary of repeat element analysis on human differentially regulated genes of endometrial stromal cells.
(DOCX)

S2 Table. Summary of over-represented repeat elements found within AZFa locus of the human Y chromosome which influence male fertility.
(DOCX)

S3 Table. Summary of under-represented repeat elements found within AZFa locus of the human Y chromosome which influence male fertility.
(DOCX)

S4 Table. Summary of repeat elements, over-represented (based on 'repeat counts') in the neighborhood of 64 rat genes associated with general rat injury.
(DOCX)

S5 Table. Summary of repeat elements, under-represented (based on 'repeat counts') in the neighborhood of 64 rat genes associated with general rat injury.
(DOCX)

S6 Table. Summary of repeat elements, over-represented (based on 'gene counts') in the neighborhood of 64 rat genes associated with general rat injury.
(DOCX)

S7 Table. Summary of repeat elements, under-represented (based on 'gene counts') in the neighborhood of 64 rat genes associated with general rat injury.
(DOCX)

S8 Table. Summary of repeat elements, over-represented (based on 'repeat counts') in the neighborhood of 9 human transcription factor genes.
(DOCX)

S9 Table. Summary of repeat elements, under-represented (based on 'repeat counts') in the neighborhood of 9 human transcription factor genes.
(DOCX)

S10 Table. Summary of repeat elements, over-represented (based on 'gene counts') in the neighborhood of 9 human transcription factor genes.
(DOCX)

S11 Table. Summary of repeat elements, under-represented (based on 'gene counts') in the neighborhood of 9 human transcription factor genes.
(DOCX)

S12 Table. Summary of repeat elements, over-represented (based on 'repeat counts') in the neighborhood of mouse orthologs of 9 human transcription factor genes.
(DOCX)

S13 Table. Summary of repeat elements, under-represented (based on 'repeat counts') in the neighborhood of mouse orthologs of 9 human transcription factor genes.
(DOCX)

S14 Table. Summary of repeat elements, over-represented (based on 'gene counts') in the neighborhood of mouse orthologs of 9 human transcription factor genes.
(DOCX)

S15 Table. Summary of repeat elements, under-represented (based on 'gene counts') in the neighborhood of mouse orthologs of 9 human transcription factor genes.
(DOCX)

S16 Table. Summary of repeat elements, commonly over-represented (based on 'repeat counts') in the neighborhood of 9 human transcription factor genes and their mouse orthologs.
(DOCX)

S17 Table. Summary of repeat elements, commonly over-represented (based on 'gene counts') in the neighborhood of 9 human transcription factor genes and their mouse orthologs.
(DOCX)

Acknowledgments

The authors also thank Mr. Anish KM, Mrs. Madhura Adavkar, Mr. Venkata Shiva Prasad and Ms. Pallavi Gupta, students from IBAB, for their efforts in data collection from different repositories and validation.

Author Contributions

Conceived and designed the experiments: KKA. Analyzed the data: DSC. Wrote the paper: DSC KKA. Created the web server and performed the case study for validation: DSC. Identified the statistical methods for analysis and contributed to discussions in that regard: PD.

References

1. Biémont C, Vieira C. Genetics: junk DNA as an evolutionary force. Nature. 2006 Oct 5; 443(7111):521–4 PMID: 17024082

2. Emery PT, Robinson TE, Duddington R, Brookfield JF. What is the impact of transposable elements on host genome variability? Proc Biol Sci. 1999 Aug 22; 266(1429):1677–83 PMID: 10501038

3. Nekrutenko A, Li WH. Transposable elements are found in a large number of human protein-coding genes. Trends Genet. 2001 Nov; 17(11):619–21 PMID: 11672845

4. Tang W, Gunn TM, McLaughlin DF, Barsh GS, Schlossman SF, Duke-Cohan JS. Secreted and membrane attractin result from alternative splicing of the human ATRN gene. Proc Natl Acad Sci U S A. 2000 May 23; 97(11):6025–30 PMID: 10811918

5. Esnault C, Maestre J, Heidmann T. Human LINE retrotransposons generate processed pseudogenes. Nat Genet. 2000 Apr; 24(4):363–7 PMID: 10742098

6. Morgan HD, Sutherland HG, Martin DI, Whitelaw E. Epigenetic inheritance at the agouti locus in the mouse. Nat Genet. 1999 Nov; 23(3):314–8 PMID: 10545949

7. McClintock B. The origin and behavior of mutable loci in maize. Proc Natl Acad Sci U S A. 1950 Jun; 36 (6):344–55 PMID: 15430309

8. McClintock B. Chromosome organization and genic expression. Cold Spring Harb Symp Quant Biol. 1951; 16:13–47 PMID: 14942727

9. Jordan IK, Rogozin IB, Glazko GV, Koonin EV. 2003. Origin of a substantial fraction of human regulatory sequences from transposable elements. Trends Genet. 19(2):68–72 PMID: 12547512

10. Lynch VJ, Leclerc RD, May G, Wagner GP. 2011. Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. Nat Genet. 2011, 43(11):1154–9 doi: 10.1038/ng.917 PMID: 21946353

11. Estécio MR, Gallegos J, Vallot C, Castoro RJ, Chung W, Maegawa S, et al. Genome architecture marked by retrotransposons modulates predisposition to DNA methylation in cancer. Genome Res. 2010 Oct; 20(10):1369–82 doi: 10.1101/gr.107318.110 PMID: 20716667

12. Xie M, Hong C, Zhang B, Lowdon RF, Xing X, Li D, et al, DNA hypomethylation within specific transposable element families associates with tissue-specific enhancer landscape. Nat Genet. 2013 Jul; 45 (7):836–41 doi: 10.1038/ng.2649 PMID: 23708189

13. Hollister JD, Gaut BS. Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. Genome Res. 2009 Aug; 19 (8):1419–28 doi: 10.1101/gr.091678.109 PMID: 19478138

14. Kunarso G, Chia NY, Jeyakani J, Hwang C, Lu X, Chan YS, et al. Transposable elements have rewired the core regulatory network of human embryonic stem cells. Nat Genet. 2010 Jul; 42(7):631–4 doi: 10.1038/ng.600 PMID: 20526341

15. Lowe CB, Bejerano G, Haussler D. Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. Proc Natl Acad Sci U S A. 2007 May 8; 104(19):8005–10 PMID: 17463089

16. Kamal M, Xie X, Lander ES. A large family of ancient repeat elements in the human genome is under strong selection. Proc Natl Acad Sci U S A. 2006 Feb 21; 103(8):2740–5 PMID: 16477033

17. Ito H, Gaubert H, Bucher E, Mirouze M, Vaillant I, Paszkowski J. An siRNA pathway prevents transgenerational retrotransposition in plants subjected to stress. Nature. 2011 Apr 7; 472(7341):115–9 doi: 10.1038/nature09861 PMID: 21399627

18. Makarevitch I, Waters AJ, West PT, Stitzer M, Hirsch CN, Ross-Ibarra J, et al. Transposable elements contribute to activation of maize genes in response to abiotic stress. PLoS Genet. 2015 Jan 8; 11(1): e1004915 doi: 10.1371/journal.pgen.1004915 PMID: 25569788

19. Mehrotra S, Goyal V. Repetitive sequences in plant nuclear DNA: types, distribution, evolution and function. Genomics Proteomics Bioinformatics. 2014 Aug; 12(4):164–71 doi: 10.1016/j.gpb.2014.07.003 PMID: 25132181

20. Grandi FC, An W. Non-LTR retrotransposons and microsatellites: Partners in genomic variation. Mob Genet Elements. 2013 Jul 1; 3(4):e25674 PMID: 24195012

21. López-Flores I, Garrido-Ramos MA. The repetitive DNA content of eukaryotic genomes. Genome Dyn. 2012; 7:1–28 doi: 10.1159/000337118 PMID: 22759811

22. Haider S, Cordeddu L, Robinson E, Movassagh M, Siggens L, Vujic A, et al. The landscape of DNA repeat elements in human heart failure. Genome Biol. 2012 Oct 3; 13(10):R90 doi: 10.1186/gb-2012-13-10-r90 PMID: 23034148

23. Ade C, Roy-Engel AM, Deininger PL. Alu elements: an intrinsic source of human genome instability. Curr Opin Virol. 2013 Dec; 3(6):639–45 doi: 10.1016/j.coviro.2013.09.002 PMID: 24080407

24. Belan E. LINEs of evidence: noncanonical DNA replication as an epigenetic determinant. Biol Direct. 2013 Sep 13; 8:22 doi: 10.1186/1745-6150-8-22 PMID: 24034780

25. Mattarucchi E, Guerini V, Rambaldi A, Campiotti L, Venco A, Pasquali F, et al. Microhomologies and interspersed repeat elements at genomic breakpoints in chronic myeloid leukemia. Genes Chromosomes Cancer. 2008 Jul; 47(7):625–32 doi: 10.1002/gcc.20568 PMID: 18398823

26. Sorek R. The birth of new exons: mechanisms and evolutionary consequences. RNA. 2007 Oct; 13 (10):1603–8 PMID: 17709368

27. Sela N, Mersch B, Gal-Mark N, Lev-Maor G, Hotz-Wagenblatt A, Ast G. Comparative analysis of transposed element insertion within human and mouse genomes reveals Alu's unique role in shaping the human transcriptome. Genome Biol. 2007; 8(6):R127 PMID: 17594509

28. Sela N, Mersch B, Hotz-Wagenblatt A, Ast G (2010) Characteristics of Transposable Element Exonization within Human and Mouse. PLoS ONE 5(6):e10907 doi: 10.1371/journal.pone.0010907 PMID: 20532223

29. Krull M, Petrusma M, Makalowski W, Brosius J, Schmitz J. Functional persistence of exonized mammalian-wide interspersed repeat elements (MIRs). Genome Res. 2007 Aug; 17(8):1139–45 PMID: 17623809

30. Newman MR, Sykes PJ, Blyth BJ, Bezak E, Lawrence MD, Morel KL, et al. The methylation of DNA repeat elements is sex-dependent and temporally different in response to X radiation in radiosensitive and radioresistant mouse strains. Radiat Res. 2014 Jan; 181(1):65–75 doi: 10.1667/RR13460.1 PMID: 24397437

31. Zabala G, Vodkin LO. Methylation affects transposition and splicing of a large CACTA transposon from a MYB transcription factor regulating anthocyanin synthase genes in soybean seed coats. PLoS One. 2014 Nov 4; 9(11):e111959 doi: 10.1371/journal.pone.0111959 PMID: 25369033

32. Ocalewicz K. Telomeres in fishes. Cytogenet Genome Res. 2013; 141(2–3):114–25 doi: 10.1159/000354278 PMID: 23988378

33. Li J, Akagi K, Hu Y, Trivett AL, Hlynialuk CJ, Swing DA, et al. Mouse endogenous retroviruses can trigger premature transcriptional termination at a distance. Genome Res. 2012 May; 22(5):870–84 doi: 10.1101/gr.130740.111 PMID: 22367191

34. Wolff EM, Byun HM, Han HF, Sharma S, Nichols PW, Siegmund KD, et al. Hypomethylation of a LINE-1 promoter activates an alternate transcript of the MET oncogene in bladders with cancer. PLoS Genet. 2010 Apr 22; 6(4):e1000917 doi: 10.1371/journal.pgen.1000917 PMID: 20421991

35. Schönbach C. From masking repeats to identifying functional repeats in the mouse transcriptome. Brief Bioinform. 2004 Jun; 5(2):107–17 PMID: 15260892

36. Lerat E. Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. Heredity. 2010 104(6):520–33 doi: 10.1038/hdy.2009.165 PMID: 19935826

37. Chandrashekar DS, Acharya KK. Comparison of non-commercial repeat element analysis tools. Startbioinfo. 2015; Available: http://startbioinfo.com/REAnalysisTools

38. Levy A, Sela N, Ast G. 2008. Transpogene and microTranspoGene: transposed elements influence on the transcriptome of seven vertebrates and invertebrates. Nucl. Acids Res. 36(Database issue):D47–52 PMID: 17986453

39. Ross SM. Introduction to probability and statistics for engineers and scientists. Fourth edition. Elsevier Publication, London; 2010; pp 147

40. Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, et al. Ensembl 2015. Nucleic Acids Res. 2015 Jan; 43(Database issue):D662–9 doi: 10.1093/nar/gku1010 PMID: 25352552

41. Frith MC, Fu Y, Yu L, Chen JF, Hansen U, Weng Z. Detection of functional DNA motifs via statistical over-representation. Nucleic Acids Res. 2004 Feb 26; 32(4):1372–81 PMID: 14988425

42. Mathelier A, Zhao X, Zhang AW, Parcy F, Worsley-Hunt R, Arenillas DJ, et al. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. Nucleic Acids Res. 2014 Jan; 42(Database issue):D142–7 doi: 10.1093/nar/gkt997 PMID: 24194598

43. Sela N, Kim E, Ast G. The role of transposable elements in the evolution of non-mammalian vertebrates and invertebrates. Genome Biol. 2010; 11(6):R59 doi: 10.1186/gb-2010-11-6-r59 PMID: 20525173

44. BIONUMBERS: The database of useful biological numbers. Available: http://kirschner.med.harvard.edu/files/bionumbers/Transposable%20element%20contents%20of%20some%20model%20organisms.pdf

45. de Koning AP, Gu W, Castoe TA, Batzer MA, Pollock DD. Repetitive elements may comprise over two-thirds of the human genome. PLoS Genet. 2011 Dec; 7(12):e1002384 doi: 10.1371/journal.pgen.1002384 PMID: 22144907

46. Tawa GJ, AbdulHameed MD, Yu X, Kumar K, Ippolito DL, Lewis JA, et al. Characterization of chemically induced liver injuries using gene co-expression modules. PLoS One. 2014 Sep 16; 9(9):e107230 doi: 10.1371/journal.pone.0107230 PMID: 25226513

47. Foresta C, Moro E, Ferlin A. Y chromosome microdeletions and alterations of spermatogenesis. Endocr Rev. 2001 Apr; 22(2):226–39 PMID: 11294825

48. Steinhoff C, Paulsen M, Kielbasa S, Walter J, Vingron M. Expression profile and transcription factor binding site exploration of imprinted genes in human and mouse. BMC Genomics. 2009 Mar 31; 10:144 doi: 10.1186/1471-2164-10-144 PMID: 19335913