4
Computational Identification of Transposable Elements in the Mouse Genome

Daudi Jjingo, Wojciech Makalowski

Repeat sequences cover about 39 percent of the mouse genome and completion of sequencing of the mouse genome [1] has enabled extensive research on the role of repeat sequences in mammalian genomics. This research covers the identification of Transposable elements (TEs) within the mouse transcriptome, based on available sequence information on mouse cDNAs (complementary DNAs) from GenBank [28]. The transcripts are screened for repeats using RepeatMasker [23], whose results are sieved to retain only Interspersed repeats (IRS). Using various bioinformatics software tools as well as tailor made programming, the research establishes: (i) the absolute location coordinates of the TEs on the transcript. (ii) The location of the IRs with respect to the 5'UTR, CDS and 3'UTR sequence features. (iii) The quality of alignment of the TE's consensus sequence on the transcripts where they exist, (iv) the frequencies and distributions of the TEs on the cDNAs, (v) descriptions of the types and roles of transcripts containing TEs. This information has been collated and stored in a relational database (MTEDB) at http://warta.bio.psu.edu/htt_doc/MTEDB/homepage.htm).

1.0 Introduction

1.1 Review

Transposable elements (TEs) are types of repetitive sequences that can move from one place to another and are interspersed throughout most eukaryotic genomes. Their sequence based classification (e.g. SINE [Short Interspersed Element], LINE [Long Interspersed Element], LTR [Long Terminal Repeat]) is frequently based on what is known as a repeat consensus sequence. A repeat consensus sequence [4] is an approximation of an ancestral active TE that is reconstructed from the multiple sequence alignments of individual repetitive sequences. Libraries of such consensus sequences have been compiled and stored in databases like RepBase [5]. Repetitive sequence identifying software like RepeatMasker [3], REPuter [6] and the like rely on such libraries to act as sources of reference sequences against which repetitive sequences from a query sequence(s) are identified. Of these repetitive sequences, transposon-derived IRs or TEs form the largest percentage. In comparison to the human genome, the mouse genome has a higher number of recent TEs that diversify more rapidly than in the human [1]. TEs don't merely

have a pronounced presence in eukaryotic genomes, they also influence these genomes' evolution, structure and functioning in many varied ways; they act as recombination hotspots, facilitate mechanisms for genomic shuffling, and provide ready-to-use motifs for new transcriptional regulatory elements, polyadenylation signals and protein-coding sequences [8]. Though the effects of the mobility of TEs are mostly neutral, in some cases they lead to undesirable mutations, resulting in diseases like Haemophilia B, B-cell lymphoma and other cancers, Neurofibromatosis and many others. These far reaching effects of TEs are an important part of the motivation behind this research. Aswell, some work has been done on inferred molecular functional associations of repeats in mouse cDNAs [7]. This paper concentrates on only TEs in the mouse transcriptome and its coverage is not limited to TEs with some evidence of functionality, but to all TEs in the transcriptome.

1.2 Research objective

With respect to the scope specified above, the research sought, at a minimum to fulfil seven aims and objectives which include establishing; (i) the absolute location coordinates of the TEs on each transcript, (ii) the location of the TEs with respect to the 5'UTR, CDS and 3'UTR sequence features, (iii) the quality of alignment of the TEs' consensus sequence on the transcripts where they exist, (iv) the frequencies and distributions of the TEs on the cDNAs, (v) descriptions of the types and roles of transcripts containing TEs, (vi) collation of the data thus accumulated into a relational database and finally, (vii) the construction of a web-based interface to facilitate access to the information.

2.0 Literature Review.

2.1 Definition and description.

They are discreet units of DNA that move between and within DNA molecules, inserting themselves at random. They are excised or copied from one site and inserted on another site either on the same or on a different DNA molecule. Both their ends are normally inverted repetitive sequences, with the sequence of base pairs on one end being reversed on the other. Their hallmark is non-reliance on another independent form of vector (such as a phage or plasmid DNA), and hence direct movement from one site in the genome to another. [10]

Table 2. Major types of TEs

Type	Replication	Main Families
Interspersed Repeats		
SINE short interspersed	Rely on LINEs	Alu, B1, B2, MIR
LINE long interspersed	Reverse transcription	L!, L2
LTR retrotransposon	Reverse transcription	ERVs, MaLRs
DNA transposon	DNA transposase	Mariner, MERs
Simple Repeats	DNA replication error	

2.6 Effect and Roles of repetitive sequences on the genome

TEs serve as recombination hotspots, providing a mechanism for genomic shuffling and a source of "readyto-use motifs" for new transcriptional regulatory elements, polyadenylation signals, and protein-coding sequences [17]. Transposons can also be disadvantageous. This is either by being inserted into the coding region of a gene, altering some of the gene in the process, or by being inserted upstream of the coding region of a gene in an area important in determining the expression of the gene, for example in an area where a transcription factor would bind to the DNA. The effects of these activities vary, some resulting in genetic disorders like Ornithine aminotransferase deficiency, Haemophilia B, Neurofibromatosis, B-cell lymphoma among others [18]. Some effects of TEs are however neutral or advantageous, some eventually leading to evolutionary novelties like the human glycophorin gene family which evolved through several duplication steps that involved recombination between Alu elements [19] [20] [21] [22]. Genomes evolve by acquiring new sequences and by rearranging existing sequences, and TEs, given their mobility contribute to this process significantly [10]. Transposons also increase the size of the genome, because they leave multiple copies of themselves in the genome and thus occupy upto 38% of the mouse genome [1]. Also, transposons are used in genetic studies, as in the case of being allowed to insert themselves in specific areas so as to "knock out" genes, a technique that turns genes off so that their function can be determined [16].

2.7 Mouse Genome

Genomic resources for the mouse are increasing at an astounding pace and the ability to manipulate the mouse genome and its sequences make the mouse a unique and effective research tool. [4]

2.7.1 Size

All Mus musculus subspecies have the same "standard karyotype" of 20 chromosomes, 19 of them being autosomes and the X and Y sex chromosomes. The 21 chromosomes together are made of a total of 2.751 billion base pairs of nucleotides.

2.7.2 Functional part of the genome.

Genomes show evolutionary conservation over stretches of sequence that have coding potential or any obvious function [23]. However, sequences can only be conserved when selective forces act to maintain their integrity for the benefit of the organism. Thus, conservation implies functionality, even though we may be too ignorant at the present time to understand exactly what that functionality might be in this case [5]. However, looking at functionality in terms of coding potential, the fraction of the mouse genome that is functional is likely to lie somewhere between 5% and 10% of the total DNA present.[5]

2.7.3 Number of genes

The number of genes in the mouse genome has been estimated using using three tiers of input [1]. First, known protein coding cDNAs were mapped onto the genome. Secondly, additional protein-coding genes are predicted using the GenWise [16]. Thirdly, de novo gene predictions from GENSCAN program [17] that are supported by experimental evidence (such as ESTs) are considered. These three strands of evidence are reconciled into a single gene catalogue by using heuristics to merge overlapping predictions, detect pseudogenes and discard misassemblies. These results are then augmented using conservative predictions from the Genie system, which predicts gene structures in the genomic regions delimited by paired 5' and 3'ESTs on the basis of cDNA and EST information from the region. The predicted transcripts are then aggregated into predicted genes based on sequence overlaps. This procedure, has estimated the number of genes in the mouse genome at about 30,000 [1].

2.7.4 Number of transcripts

The number of known mouse transcripts has been determined at about 60,770 represented as full-length mouse complementary DNA sequences [27]. These are clustered into 33,409 'transcriptional units'. Transcriptional unit (TU) refers to a segment of the genome from which transcripts are generated.

3.0 Data Processing

3.1 Background

The number of known full-length mRNA transcripts in the mouse has been greatly expanded by the RIKEN Mouse Gene Encyclopedia project and is currently estimated at about 60,770 [26] [28] clustered into 33,409 'transcriptional units' [27]. Of these transcriptional units, 4,258 are new protein-coding and 11,665 are new non-coding messages, indicating that non-coding RNA is a major component of the transcriptome. 41% of all transcriptional units showed evidence of alternative splicing [26]. In protein-coding transcripts, 79% of splice variations altered the protein product [27]. The Riken Mouse ESTs and cDNAs are deposited in the public databases DDBJ, GenBank, EMBL.

3.2 GenBank

The GenBank database at NCBI [28] provided source of the dataset of the cDNAs of the known mouse transcriptome.

3.2.1 Sequence and data retrieval from GenBank

Table 3. Queries used in data and sequence retrieval

Search	Query	Results
#1	Search Mus musculus[Organism] AND "biomol mrna"* [Properties] AND complete cds [Title]	55664
#2	Search Mus musculus[Organism] AND "biomol mrna"[Properties] AND "srcdb refseq"[Properties]	26562
#3	Search #1 OR #2	82226

The queries above yielded 82226 sequences, downloaded in GenBank and fasta formats (4). These constituted the starting dataset of cDNAs corresponding to the currently known1 mouse transcriptome.

3.3 Eliminating redundancy

3.3.1 Background

GenBank is a highly redundant database. It is thus pertinent that redundancies are expunged from the cDNA dataset.

3.3.2 Patdb

Patdb is software that removes redundancies by merging all identical strings and sub- strings and removing all sequences that are perfect substrings of other sequences. It then concatenates the identifiers of the affected sequences [29]. For example the sequences MEPVQ and MEPVQWT are merged, and if there is another MEPVQWT sequence elsewhere, it is discarded [29]. For our purposes, this not only deals away with redundancies, but also helps assemble the various incomplete cDNA fragments into full-length transcripts. Of the 82226 sequences subjected to patdb, 72697 sequences satisfied the minimum length requirement of patdb (100 nucleotides), meaning that the difference of 9530 sequences were too short to be relevant for the purposes of this research as they are far shorter than the average transcript (usually $>$ 1200 bp). Patdb found 3585 sequences to be either substrings or perfect replicas of other sequences, resulting in a total of 69112 sequences (102037991 bp) as the unique cDNA dataset.

3.4 Mapping Transposable Elements

3.4.1 TE features (identifying characteristics)

As mentioned in chapter 1, TEs have some characteristic distinguishing features, particularly the universal existence of inverted repetitive sequences on either ends of all TEs. This feature and other characteristics like possession of a transposase ORF (Open Reading Frame) and their existence as multiple copies (middle repetitive DNA) within the genome, were the attributes used in computational identification.

3.4.2 TE libraries

Identification of TEs based on their features has enabled the construction of libraries of consensus sequences of various types of TEs. RepBase Update (RU) [32] [33] which is a service of the Genetic Information Research Institute (GIRI) [31] is a comprehensive database of repetitive element consensus sequences. Most prototypic sequences from RU are consensus sequences of large families and subfamilies of repetitive sequences. Smaller families are represented by sequence examples [32].

3.4.3 RepeatMasker

RepeatMasker [3], developed by Arian Smit and Phil Green, is software that screens DNA sequences for low complexity sequences, repetitive/TEs including small RNA pseudogenes, Alus, LINEs, SINEs, LTR * Though the property "biomol mrna" is what is used for the searches, GenBank actually returns cDNAs elements, and others, producing a detailed annotation that identifies all of the repetitive elements in a query sequence [29] [30]. RepeatMasker makes use of RepBase libraries [33], which act as reference points for the identification of repetitive elements in a query sequence. RepeatMasker employs a scoring system to ensure that only statistically significant alignments are shown. It uses statistically optimal scoring matrices derived from the alignments of DNA transposon fossils to their consensus sequences [7]. However, it does not locate all possibly polymorphic simple repetitive sequences. Only di–pentameric and some hexameric repetitive sequences are scanned for, and simple repetitive sequences shorter than 20 bp are ignored.

4.0 Results And Computational Analysis.

4.1 Specialised Object oriented Tools:

Bioperl and Perl: BioPerl [9], is an object oriented form of the Perl programming language [35] which relies mainly on open source Perl modules for bioinformatics, genomics and life science research. It provides reusable Perl modules that facilitate parsing of large quantities of sequence data from various molecular biology programs.

4.2 General analysis:

Within the non-redundant 69112 sequence dataset, command line computational analysis revealed that RepeatMasker identified 47204 repetitive sequences. 20023 of these were simple-repeats, 9583 lowcomplexity repeats and 17598 complex-repeats/TEs. (Table 4)

Table 4. Relative abundances of types of Repetitive sequences

	number of elements*	length occupied	percentage of sequence
SINEs:	9277	1118994 bp	1.10 %
B1s	4260	472916 bp	0.46 %
B2-B4	3852	549367 bp	0.54 %
IDs	654	45780 bp	0.04 %
MIRs	511	50931 bp	0.05 %
LINEs:	1943	572023	bp 0.56 %
LINE1	1567	532555 bp	0.52 %
LINE2	320	33623 bp	0.03 %
L3/CR1	56	5845 bp	0.01 %
LTR elements:	4192	1341750 bp	1.31 %
MaLRs	1220	257647 bp	0.25 %
ERVL	530	160390 bp	0.16 %
ERV_classI	327	95357 bp	0.09 %
ERV_classII	1914	774665 bp	0.76 %
DNA elements:	603	89481 bp	0.09 %
MER1_type	468	68204 bp	0.07 %
MER2_type	74	13225 bp	0.01 %
Unclassified:	47	10171 bp	0.01 %
Total interspersed repeats:		3132419 bp	3.07 %
Small RNA:	133	9704 bp	0.01 %
Satellites:	15	1750 bp	0.00 %
Simple repeats:	19897	891616 bp	0.87 %
Low complexity:	9530	415965 bp	0.41 %

4.3 Filtering out simple and low complexity regions

The emphasis of this research being on TEs, the simple-repeats and low-complexity regions were filtered out by UNIX command line computation to retain a dataset consisting of only TEs/complex repeats. This dataset contains 17598 records representing an equal number of TEs.

4.4 Getting CDS (Coding Sequences) coordinates.

While RepeatMasker output avails the coordinates of the repetitive sequences on a transcript, it does not show the coordinates of the CDS on the transcript. This necessitated the obtaining of CDS coordinates for each transcript identified as possessing a TE by RepeatMasker. This was effected in two different stages.

Stage 1: Involved computing the GI identifiers of all transcripts with TEs. Because some transcripts contain more than one TE, some GI identifiers feature more than once. Thus this stage also involved removing the resultant redundancy. The result was a list of 10213 GI identifiers each with a tab delimited number on its left showing the number of TEs on that particular transcript.

Stage 2: Involved a BioPerl/Perl script which uses the GI list from above to extract the corresponding CDS coordinates for each transcript from the GenBank dataset that was first downloaded (section 3.2.1). The script then stores each GI with its start and end coordinates separated by tabs in a file.

4.5 Computing location and length of each transposable element

Using a Perl language script, the CDS coordinates were used to; implicitly determine the 5'UTR and 3'UTR of each transcript where these exist, establish the total number of TEs in each of the so determined regions with respect to the entire dataset, determine the length of each single TE.

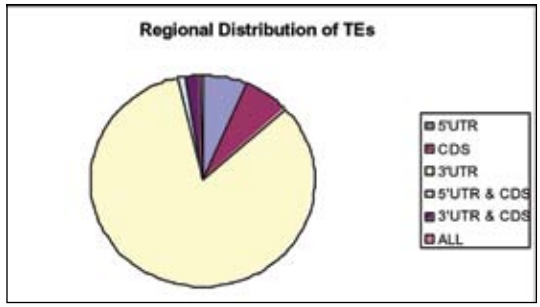
4.5.1 PDB screening of transcripts with TEs in CDS region

The current PDB dataset was downloaded, subjected to patdb to remove redundancies, and then screened against transcripts with TEs in the coding region using blastx. All hits with $>80\%$ identity, alignment length >50, and e -value <0.001 were analysed. None was found to code for a protein of known 3D.

Table 5. Computed TE distribution in sequences and regions.

Initial number of Sequences	82226
Number of nonredundant sequences	69112
Total length (bp)	102037991
GC level (%)	50.01
Sequences with a TE-cassette at all	10213
Sequences with a TE-cassette in CDS	700
TEs lying exclusively in 5'UTR	1179
TEs overlapping 5'UTR and CDS	211
TEs lying exclusively in CDS	1147
TEs lying exclusively in 3'UTR	14618
TEs overlaping3'UTR and CDS	387
TEs overlapping 5'UTR, CDS,& 3'UTR.	59

Fig 2. Pie chart showing relative distribution of TEs in regions



4.6 Average transcript and TE lengths

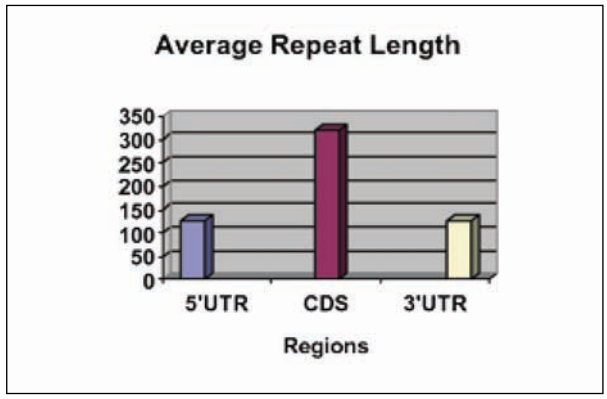
These were calculated from the 69112 non-redundant sequence dataset (see 3.3.2) and the TE lengths files from section 4.5, using a BioPerl/Perl script.

Table 6. Shows average transcript length and average lengths of repetitive sequences in the different regions.

Average transcript length (includes repeatless mRNAs)	1476
Average TE length in 5'UTR	127
Average TE length in CDS	319
Average TE length in 3'UTR	128

averages are calculated to the nearest whole number.

Fig 3. Average regional repetitive sequence lengths.



4.7 Average sequence regional lengths

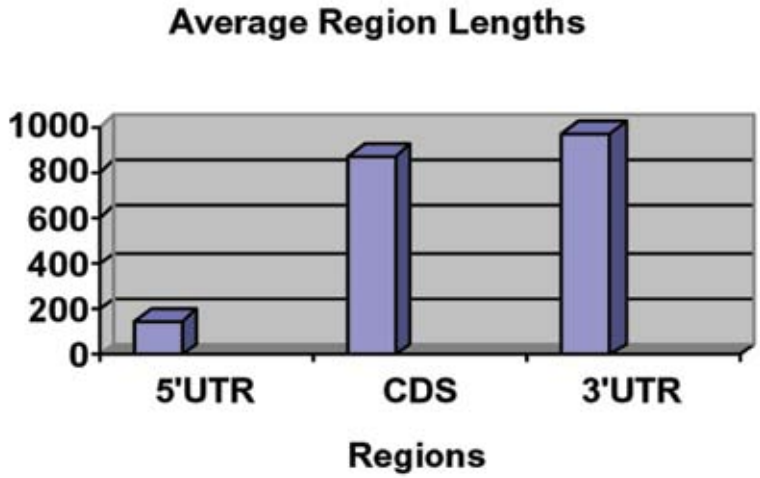
In the context of this research, the primary sequence features are the 5'UTR, CDS and 3'UTR. While these were indirectly alluded to in section s (4.4) and (4.5) for purposes of determining the repetitive sequences in them, this step goes into an

outright and direct calculation of their respective lengths using a Perl/BioPerl script.

Table 7. Average sequence and feature length.

Feature	Average length
5'UTR	150
CDS	877
3'UTR	972
Transcripts (with repetitive sequences)	1998

Fig 4. Average region lengths.



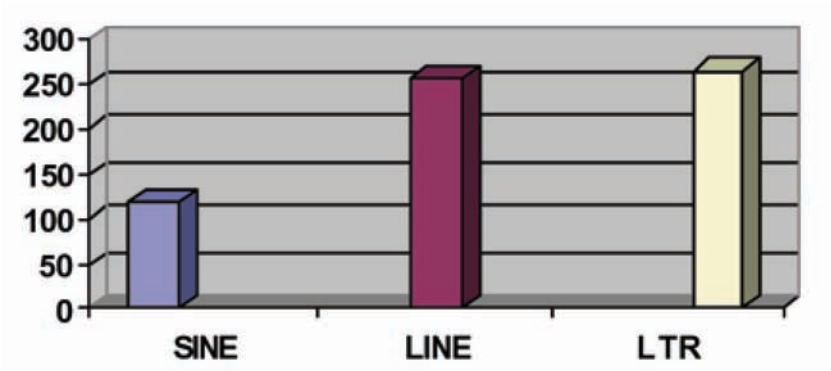
4.8 Average length of major IR families

These were computed using the dataset from the filtering procedure in section 4.3

Table 8. Average lengths of Interspersed repetitive sequence families

IR class	Average length
SINE	119
LINE	257
LTR	265

Fig 5. Average lengths of interspersed repetitive sequences.



4.9 Computing frequencies and occurrence of transposable elements.

TEs tend to move randomly and different transcripts will have varied instances of them. This computational analysis determined the number of transcripts with TEs at all. The analysis was executed in two stages, one involving command line and the other scripted computation. In the first instance, a list of non redundant GI identifiers each tab delimited from the number of TEs it contains (already generated in section 4.4) was analysed to determine occurrence of TEs . It was found that 10213 transcripts or 15% of the 69112 sequences posses TEs. In the second instance the GI list from above was subjected to a Perl script to compute the frequencies of TEs on the transcripts, results of which are shown in the table below.

Table 9. Number of transposable elements per transcript

TEs per mRNA	1	2	3	4	5	6	7	8	9	>=10
Counts	6225	2409	789	366	178	104	58	37	17	30

Fig 6. Showing Distribution of TEs in mouse transcripts

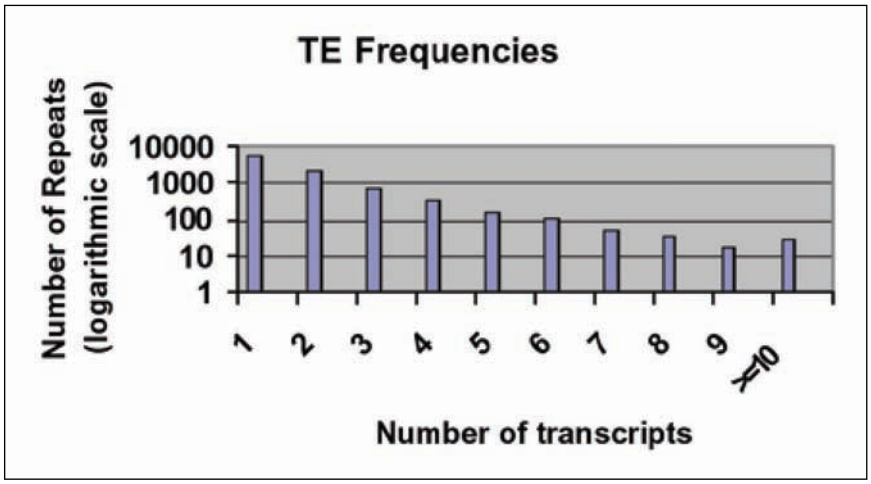
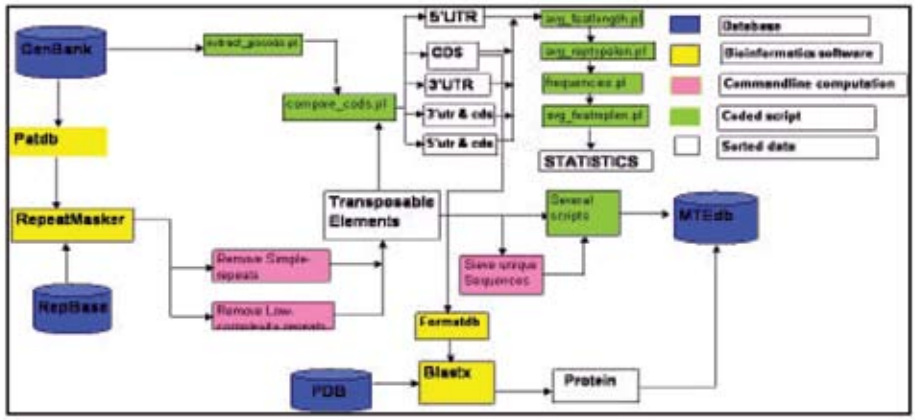


Fig 7. Processing framework used in data processing and computational calculations.



5.0 Discussion

5.1 Quantity versus coverage

Of the 69112 sequences queried, only 10213 or 15% were found to contain repetitive sequences. The actual base pair coverage of these repetitive sequences was 4451205 bp, representing 4.36% of the entire nucleotide set of 102037991 bp. Therefore, within transcripts with CDS information, repetitive sequences cover a much smaller percentage (4.36%) compared to 38% bp coverage in the complete mouse genome. Results in section 4.2 reveal that of the 47204 RepeatMasker identified repetitiv e sequences, low complexity repeats account for 9583 which represents 20% of the total. 20023 simple repeats account for 42%, while complex repeats/TEs represent 37% with a quantity of 17598. However, the coverage in

terms of base pairs is a different picture, with complex repeats (37% by quantity), covering 3.49 times more nucleotides than simple repeats (42% by quantity) and 7.49 times more nucleotides than low complexity repeats (20% by quantity). This is primarily because IRs are much longer in length than their simple and low complexity counterparts, in a measure that more than compensates for their relative inferior quantity.

5.2 Interspersed repetitive sequence occurrence

This project's computational analysis (section 4.3), it must be emphasised, concentrated on IRs/TEs - (excluded simple, satellite and low complexity repeats). The analysis indicated that the mouse genome is dominated by three major classes of IRs (Table 4) namely SINES, LINES and LTR elements. Between themselves, these three classes accounted for 95% of IR (interspersed repeat) sequence coverage. Further still, even within these three, our results revealed a disproportionate presence of some subclasses. The L1 family dominated the LINE class, accounting for 93% of all LINE sequence coverage, leaving L2 and L3 families to share the remaining 6% in a ratio of 3:1 respectively. The SINE class was predominantly shown to be composed of B elements; B1 elements dominated with 44%, followed by B2 elements with 18%, while B3 and B4 elements together contribute 26%. This leaves a mere 12% coverage for the other SINE classes; IDs and MIRs. Within the LTR class, ERV_class II and MaLRs represented 58% and 19% sequence coverage respectively, leaving a dismal 12.07% to be shared among the ERVL and ERV_ class I families.

5.3 Distribution

The position information from the computational analysis in section 4.4 and the statistical results in section 4.5, indicate that IRs can occur in any given region (5'UTR, CDS or 3'UTR) of a transcript. However it appears there is a fairly significant bias towards insertion in the 3'UTR region. The higher numbers of IRs in 3'UTR (see Table 10 below) could be attributed to the fact that this region is longer than both the other two. While that might be part of the reason, a closer examination of the relative lengths of the three regions vis a vis the relative numbers of IRs in each region seems to lend support to the idea of a preference of insertions into the 3'UTR. For instance while the 3'UTR is on average only 1.11 times longer than the CDS region, computational analysis reveals that it contains 15 times as many IRs (Tables 5 and 10). The disparity is not as high in comparison with the 5'UTR but holds nevertheless. 3'UTR is 6 times as long as 5'UTR, but contains 14 time s as many IRs. Perhaps the most conspicuous observation is the statistical revelation of the fact that IR insertions are disproportionately much lower in the CDS region when contrasted with insertion in the UTRs. We have already illustrated this fact with respect to the 3'UTR region. A similar trend can be observed when the CDS region is compared to the 5'UTR region. While the absolute number of IRs within the two regions is approximately equal, with the

ratios being roughly 1:1, it becomes clear that their density within the CDS is much less given the fact that the average CDS region length is 5 times as long as the average 5'UTR length. This observation leads to interesting questions about how IRs or TEs relate to the translated part of the mouse genome. This would imply that while insertions have an equal chance of falling into any region, most IR insertions within the CDS lead to negative consequences for the organism. As a result, these are selected against during genomic evolution. Even at the higher level of transcripts, further evidence of the general undesirability of IR insertion can be adduced, supported by the fact that of the 69112 transcripts that were screened only 10213 or 15% contained any repetitive sequence of any type at all. (section 4.4 and Table 5). Moreover, even within sequences that possessed TEs, the occurrence of sequences containing a given number of TEs dropped logarithmically as the number of TEs per mRNA increased (diagram 6); yet another pointer to the general undesirability of TE insertion. However the mere presence of IRs in the CDS, even though they are relatively fewer in comparison to other regions, bears evidence to the fact that some few insertions into the CDS result in positive or neutral consequences for the host organism and are able to be retained during the course of genomic evolution. One notable and quite interesting observation though, is the fact that the mean length of TEs within the CDS region is significantly higher than those of the other two regions (Table 6). Whether this is of any functional or evolutionary significance is an issue that cannot be appreciably resolved by the scope of this research. In all cases, IRs that overlap or exist in more than one region do not seem to be favoured by genomic evolution. In our case, only 59 or ~0.0025% of the 17598 TEs overlap with all three regions of a transcript, and only 211 or ~0.01% overlap both the 5'UTR and the CDS regions. A mere 387 or 0.02% overlap the CDS and 3'UTR regions.

Table 10. Distribution of repetitive sequences in different mRNA regions

Region	Average Regional Length	Number of IRs
5'UTR	150	1179
CDS	877	1147
3'UTR	972	14618
5'UTR/CDS/3'UT R	1998	59
5'UTR/CDS	N/A	211
3'UTR/CDS	N/A	387

5.4 Possible sources of error

It's pertinent to point out at this point of the discussion that all the results used in the preceding discussion constitute a marginal error arising out of RepeatMasker's identification method. It detects only major simple (di-hexameric repeats) having more than 20 nucleotides, and therefore misses a small number of simple repeats.

Another possible source of error is that a small fraction of potential new mouse repetitive sequences that lack a consensus sequence in RepeatMasker's library against which the query sequences are searched, may either be missed, or incorrectly identified and positioned.

8.0 Conclusions And Further Directions

8.1 Conclusions

While the spread of TEs within a genome is very much a random process, some types of repetitive sequences have been more successful than others. The mouse genome is dominated by three major classes of IRs, namely SINES, LINES and LTRs which between themselves, represent upto 95% of Interspersed repeat coverage in the genome (Table 4). However, even within these three types, some subfamilies are more dominant than others such as the L1 family for the LINEs, the B elements in the SINEs and the ERV_class II in LTRs. The presence of TEs is however much more pronounced in the UTRs than in CDS. This pattern of TE distribution within the transcripts offers two interesting phenomena. On the one hand, the replication and movement of repetitive sequences within the genome as a whole has been a big success, to the extent that they occupy upto 39% of the mouse genome [1]. On the other hand however, their insertion within the CDS, the coding part of the genome, has been relatively minimal. This pattern seems to lend credence to the idea that though insertion of TEs in the genome may lead to desirable evolutionary novelties, for the most part its effects are negative, sometimes fatal and it's therefore selected against by genomic evolutionary pressures and thus their relatively diminished presence in the CDSs. This observation not withstanding, the mere presence of some TEs within the CDSs, attests to the fact that a few of the insertions lead to positive effects and are thus conserved within the region. Their very successful presence and highly conserved status within the non-coding parts of the genome, is further proof to previous observations that even out of the coding region, they serve useful purposes like acting as recombination hotspots [8]. The mouse transposable element database (MTEDB) represents a major resource for the study of functional genomics in the mouse, particularly the complex and intricate phenomena of repetitive sequences in the organism. The database is certainly not perfect, partly because of the continued discovery of novel types of repetitive sequences, shortfalls in the software and methods used among other reasons. It's our belief that though the database may not be perfect, or even complete, it still provides a strong foundation for studying and building of new mouse repetitive sequence data sources, mainly because of the multi dimensional data analysis tools that it offers. A web-based interface for the database, MTEDB can be found at http://warta.bio.psu.edu/htt_doc/MTEDB/homepage.htm.

8.2 Further directions

Taking into account the high abundance and redundancy of repetitive sequences in the mouse genome, it is evidently clear that though the contribution of this and other research efforts is important, there is still an awful lot that science is yet to discover. This is especially so with respect to our knowledge of the biological roles of repetitive sequences and their function in the generation and evolution of the various intricate genomic networks.

Though this research included screening of the data set against the currently known protein structures from PDB, no repetitive sequences were found to have any direct structural information. This was mainly because the PDB is itself still a very limited database, representing only a very small percentage of proteins because the number of known protein structures is still very limited. Further research therefore, could include structural modeling of transcripts with TEs to boost our understanding of the effects of TEs on protein structures.

Another important area of further research could include the study of repetitive sequences in the untranscribed part of the mouse genome and their possible influence on the resultant proteome, say as regulatory regions for gene expression.

Comparison of this project's database with other emerging or existing databases on the mouse genome in general or mouse repetitive sequences in particular is another area of research that would serve to enrich the data set and act as a point of cross reference.

9.0 References

Waterston, R.H, et al. (2002) Initial sequencing and comp arative analysis of the mouse genome. *Nature*, 420. 520-562

Schonbach, C. (2004) From masking repeats to identifying functional repeats in the mouse transcriptome. *Briefings in Bioinformatics*. Vol 5, No.2 107-117.

Smit, A.F.A. and Green, P. RepeatMasker at http://repeatmasker.org

http://www.ncbi.nlm.nih.gov/genome/guide/mouse/

Silver, L.M. (1995) Mouse Genetics Concepts and Applications.Oxford University Press, Oxford.

Kurtz, S. and Schleiermacher, C. (1999) Fast Computation of Maximal Repeats in Complete Genomes, Bioinformatics, 15(5), pp. 426-427.

Nagashima, T. et al. (2004) FREP: A Database of Functional Repeats in Mouse cDNAs. *Nucleic Acid Research, Vol.32*, D471-D475.

Lorenc, A. and Makalowski, W. (2003) Transposable elements and vertebrate protein diversity. Genetica 118: 183-191.

http://www.bioperl.org/

Lewin, B. (1997) Genes. Oxford University Press, Oxford, Newyork, Tokyo, Ibadan.

Makalowski, W. (2001) The human genome structure and organization. Acta Biochimica Polonica 48: 587-598.

Pietrokovski, S. and Henikoff, S. (1997) A helix-turn-helix DNA-binding motif predicted for DNA - mediated transposases. Molecular & General Genetics 254, 689-695.

Dawkins, R. (1982) The Extended Phenotype. Oxford University Press, Newyork.

Russel, P. (2002) iGenetics. Benjamin Cummings publishers, Newyork.

Makalowski, W. (2003) Not Junk After All. Science 300: 1246-1247.

Birney, E. and Durbin, R. (2000) Using GeneWise in Drosophila annotation experiment. Genome Res.10, 547-548.

Burge, C. and Karlin, S. (1997) Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. 268, 78-94.

Maria, R. (1995) The Impact of Short Interspersed Elements (SINEs) on the Host Genome. R.G. Landes company, Austin, Texas USA.

Fukuda, M. (1993) Molecular genetics of the glycophorin A gene cluster. Semin Hematol 30: 138- 151.

Onda, M., Kudo, A., Rearden, M., and Fukuda, M. (1993) Identification of a precursor genomic segment that provided a sequence unique to glycophorin B and E genes. Proc Natl Acad Sci U S A 90: 7220-7224.

Rearden, A., Magnet, A., Kudo, S. and Fukuda, M. (1993) Glycophorin B and glycophorin E genes arose from the glycophorin A ancestral gene via two duplications during primate evolution. J Biol Chem 268: 2260-2267.

Rearden, A., Phan, H., Kudo, S. and Fukuda. M. (1990) Evolution of the glycophorin gene family in the hominoid primates. Biochem Genet 28: 209-222.

Kevles, D. and Hood, L. (1992) The Code of Codes: Scientific and Social Issues in the Human Genome Project (Harvard University Press, Boston)

Milner, C. M. and Campbell, R. D. (1992) Genes, genes and more genes in the human major histocompatibility complex. BioEssays 14: 565-571.

Hasties, N.D. and Bishop, J.O. (1976) The expression of three abundance classes of messenger RNA in mouse tissues. Cell 9: 761-774.

Hidemasa, B., et al. (2003) Systematic Expression Profiling of the Mouse Transcriptome Using RIKEN cDNA Microarrays: Genome Res.2003 June;13 (6b): 13181323.

Okazaki, Y., et al. (2003) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature 420, 563 – 573.

http://www.ncbi.nlm.nih.gov/

Korf, I., Yandell, M. and Bedell, J. (2003) Blast. O'Reilly & Associates, Inc., California, Cambridge, Tokyo.

Lander, E.S., Linton, L.M., et al. (2001) "Initial sequencing and analysis of the human genome." Nature 409:860-921.

http://www.girinst.org/