Christian Schönbach
is a Team Leader at the RIKEN Genomic Sciences Center in Japan and an Associate Faculty member at the Institute for Infocomm Research in Singapore. His main research interests are computational and functional genomics approaches in immunology.

From masking repeats to identifying functional repeats in the mouse transcriptome

Christian Schönbach

Date received (in revised form): 16th March 2004

Abstract

The back-to-back release of the mouse genome and the functionally annotated RIKEN mouse full-length cDNA collection was an important milestone in mammalian genomics. Yet much of the data remain to be explored in terms of biological effects and mechanisms. For example, interspersed repeats account for 39 per cent of the mouse genome sequence and 11 per cent of representative transcripts. A considerable number of transposable repeat elements are still active and propagating in mouse compared with human. While existing repeat databases and tools assist the classification of repeats or identification of new repeats, there is little bioinformatic support towards exploring the extent and role of repeats in transcriptional variation, modulation of protein function, or gene regulatory events. Since the mouse is used as a model organism to study human genes and their disease associations, this review focuses on information extraction and collation that captures the functional context of repeats in mouse transcripts to facilitate the biological interpretation and extrapolation of findings to the human.

Keywords: *interspersed repeats, simple repeats, functional repeats*

INTRODUCTION

Eukaryotic genomes frequently include repeat sequences that are scattered throughout the genome. Apart from segmental duplications of chromosomal regions and small RNAs, the majority of repeats fall into two broad categories: transposon-derived interspersed repeats and simple sequence repeats.[1–3] Both categories can be further classified (Table 1) by the mode of repeat expansion, number of repeating nucleotide units and sequence similarity to a repeat consensus sequence.[4] The consensus sequence[5] represents the approximation of an ancestral active transposon element that is reconstructed from the multiple sequence alignments of individual repeat sequences (inactive copies of the transposable element). The consensus repeat sequences are the core of the repeat reference database RepBase,[6] which is used by the program RepeatMasker (see also Tables 2 and 3). RepeatMasker uses the Smith–Waterman algorithm together with repeat-optimised scoring matrices to compare query sequences against RepBase. A typical RepeatMasker report provides a Smith–Waterman score-based report on repeat classification, matching positions in the query and repeat consensus, orientation, repeat GC content and diversity information among others. Both resources are essential for masking or automatically annotating repeats to obtain a genome-wide view of repeat distribution, frequency and diversification.

The human and mouse genome analyses revealed for example, that transposon-derived interspersed repeats make up the largest fraction of repeats in mammalian genomes. Interspersed repeats constitute 46 per cent of the human and 38 per cent of the mouse genomes as opposed to 3 per cent in the *Fugu* fish genome.[4,7] The distribution of interspersed repeats can be biased. SINEs and LINEs of both mouse and human tend to occupy regions with high G+C and A+T content, respectively. However, the mouse contains a higher number of recent transposon-derived repeats that diversify more rapidly than in the human.

C. Schönbach,
Immunoinformatics Team (formerly Biomed. Knowledge Discovery Team),
Bioinformatics Group,
RIKEN Genomic Sciences Center (GSC),
RIKEN Yokohama Institute,
Suehiro-cho, Tsurumi,
Yokohama, Kanagawa 230-0045,
Japan

Tel: +81 (0) 45 503 9551
Fax: +81 (0) 45 503 9552
E-mail: schoen@postman.riken.jp

Table 1: Classification of mammalian repeat sequences

Category	Expansion	Transmission	Major families
Interspersed repeats			
SINE short interspersed	Depending on LINEs	Vertical[1]	Alu, B1, B2, MIR
LINE long interspersed	Reverse transcription	Vertical	L1, L2
LTR retrotransposon	Reverse transcription similar to retroviruses	Lateral[2] and vertical	MaLRs, ERVs
DNA transposon	DNA transposase	Lateral and vertical	MERs, Mariner
Simple repeats	DNA replication error		
Dinucleotide repeats, triplet repeats		Vertical	Microsatellites or VNTR

Table 2: Databases covering repeats, dedicated repeat databases and collections

Category	Name	URL
Databases		
Reference database	RepBase	http://www.girinst.org/
Part of genome databases	TIGR Rice	http://www.tigr.org/tdb/e2k1/osa1/
	ENSEMBL	http://www.ensembl.org/
	BDGP Fruit fly	http://www.fruitfly.org/
Specialist databases		
Mouse functional repeats	FREP	http://facts.gsc.riken.go.jp/FREP/
*Alu*s	*Alu*Gene	http://alugene.tau.ac.il/
HERVs	HERVd	http://herv.img.cas.cz/
Simple repeats		
Human microsatellites	STRBase	http://www.cstl.nist.gov/div831/strbase/
Various model organisms	Tandem Repeats Database	http://tandem.bu.edu/
	Inverted Repeats Database	http://tandem.bu.edu/
Various plants	PlantSSR	http://www.genome.clemson.edu/projects/ssr/
Human, mouse, rat	GRID	http://grid.abcc.ncifcrf.gov/
Human	RSDB	http://rsdb.csie.ncu.edu.tw/
Bacterial palindromes	BIME	http://www.pasteur.fr/recherche/unites/pmtg/repet/
Collections		
Triplet repeats		http://www.neuro.wustl.edu/neuromuscular/mother/dnarep.htm
Vertebrate retrotransposons		http://zmbe2.uni-muenster.de/expath/alltables.htm

Table 3: Tools for repeat identification

Name	Description	URL
General tools		
RepeatMasker	General repeat finding and masking	http://repeatmasker.genome.washington.edu/
MaskerAid	RepeatMasker with higher performance	http://blast.wustl.edu/maskeraid/
RepeatFinder	Finds repetitive structures in genome sequences	http://www.tigr.org/software/#gfa/
CENSOR	Repeat search and classification based on RepBase	http://www.girinst.org/Censor_Server.html
Interspersed repeats		
Alu Blast		http://www.genome.ou.edu/alu_blast.html
Simple repeats		
REPuter	Detects direct and degenerate simple repeats	http://www.genomes.de/
Tandem Repeat Finder	Finds exact and degenerate tandem repeats	http://tandem.bu.edu/trf/trf.html
PTRfinder	Identifies perfect tandem repeats	http://ncisgi.ncifcrf.gov/~collinsj/Tandem_Repeats/
Binary Repeat Align	Aligns tandem repeat regions	http://repeatalign.cgb.ki.se/
Protein Repeats	Detects internal sequence repeats	
REPRO		http://ibivu.cs.vu.nl/programs/reprowww/
Internal Repeat Finder		http://www.doe-mbi.ucla.edu/Services/Repeats/

For instance, the mouse LTR elements MaLR and intracisternal-A particles (IAP) are still active and expanding. Approximately 10 per cent of spontaneous mouse mutants are attributed to IAP insertions.[8] While most of the data underlying potential repeat-associated biological effects are part of larger genome databases they are not presented in an obvious functional context.

Defining repeat context information

A few specialist repeat databases (Table 2) offer excellent repeat classification (HERVd[9]), easy-to-use sequence analysis and retrieval services for human Alu sequences from exons or introns (*Alu*Gene[10]) or simple tandem repeats (GRID[11]), but no or little information on potential effects of repeats in a particular gene or on its gene product. On the other hand, the curated collection of vertebrate transposable elements is rich in functional information but not searchable. The situation for simple repeats is comparable.

Functional REPeats

Simple repeats were originally defined as reiterated dinucleotide or tandem sequences (eg $(TG)_n$) with 10–60 repeating units.[12] As the simple repeats are polymorphic they are also called variable number tandem repeats or microsatellites. Microsatellites are an important source of genetic variation[13] that is exploited in DNA typing and linkage studies. Several databases and programs that are dedicated to microsatellite mapping and identification are listed in Tables 2 and 3.

Semi-automated knowledge discovery support system for inferring FREPs

Edwards and coworkers[14] expanded the simple repeat definition to triplet and tetrameric repeats. Triplet repeats of protein-coding region sequence may result in loss or gain of protein function, depending on the number of repeating units and therefore cause a number of genetic diseases. For example, the expansion of polyglutamine-translated CAG triplet repeats has been associated with increased cell toxicity and neurodegenerative diseases.[13] Associations of simple repeat sequences with inherited disease genes are reported in RSDB.[14] However, its scope is restricted to human tandem repeats in genes with OMIM[15] records.

FUNCTIONAL REPEAT CANDIDATES

In the light of the existing repeat analyses and databases, we need to establish the context that permits biologists to retrieve and easily explore data and supporting evidence on genes or transcripts with potential repeat-function associations. Here, context is defined as information extracted from genomic, cDNA and protein sequence data, annotations and controlled vocabulary in MEDLINE abstracts that enable the user to infer potential effects on promoter function, transcription (eg splicing or polyadenylation), modulated protein functions or pathologies with suspected repeat etiology contributions. Repeats that satisfy one or more of the above context conditions are designated Functional REPeat (FREP) candidates.[16]

The computational inference of FREPs is a multi-step process that involves (1) information extraction from multiple data types and sources, (2) filtering, (3) integration and (4) curation or manual checking of the reported information. The first step requires *a priori* knowledge of the data sources tools and their limitations. Steps 1–3 are only a means to an end, while step 4 is the beginning of the validation process or a new round of accumulating more complex knowledge. This strategy has already been put into practice in a knowledge discovery support system (FACTS)[17] that was used to infer molecular interactions and disease gene association for 60,770 curated RIKEN mouse full-length cDNA sequences.[18]

Therefore components of FACTS were modified to build a system for inferring functional repeats from the source sequences – 60,770 curated RIKEN mouse full-length cDNA sequences and 44,106 non-EST (expressed sequence tag) mouse cDNAs (GenBank release 131) – of the representative transcript and protein set.[18] The value of the representative transcript set (RTS) to infer

functional repeat candidates lies in its curated clusters of 33,409 representative transcripts (transcriptional units), easy access to cross-accessions and simpler identification of variant repeats.

FREP SYSTEM

The FREP system (Fig. 1) consists of a production pipeline and a relational database. The production pipeline comprises sequence-based analysis and textual information extraction components. The former include RepeatMasker, BLAT,[19] SIM4[20] (cDNA-genome alignment) and InterProScan[21] (motif search) programs. Various Perl scripts were used to (1) determine the positions of the repeats in relation to the coding sequence (CDS) start and end positions; (2) extract information from the RepeatMasker outputs (eg repeat annotation, alignments and GC content); (3) translate the repeat if it resides in the CDS; (4) determine if repeats of RTS cluster member show variation and (5) identify poly(A) signals in 3′ untranslated region (UTR) located repeats.

Extraction of accessory text and sequence information

The components for extracting accessions, gene names and disease-related information from OMIM or MEDLINE abstracts are described in detail in FACTS.[17] In brief, OMIM titles were extracted from the OMIM Morbidmap files by matching the associated gene symbols/names to the corresponding gene name/symbol fields of FREP candidate sequences. Disease MeSH (Medical Subject Headings) were extracted from MEDLINE abstracts which were retrieved with semi-automatically constructed queries using the gene names and/or symbols of repeat-containing cDNAs. Disease MeSH derived from the MeSH tree were matched to MeSH terms in the MH field of the extracted MEDLINE abstracts. When the substance

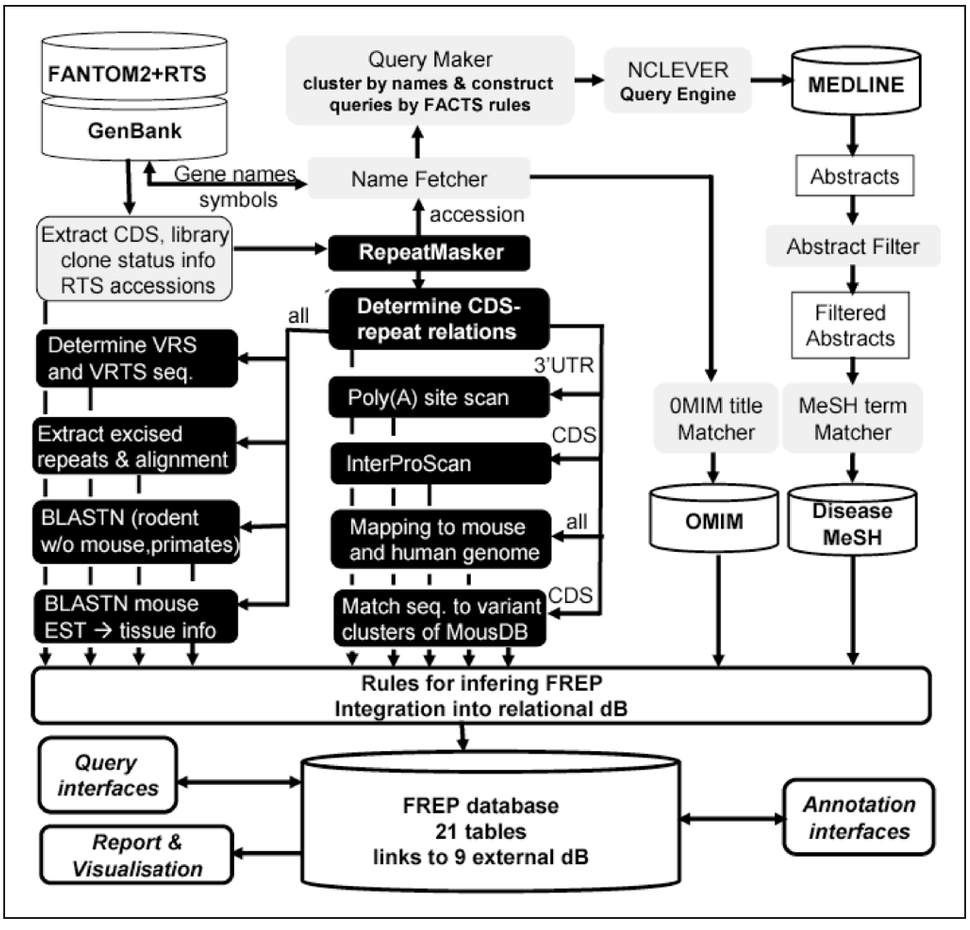

Figure 1: FREP production system and database. Black boxes symbolise programs and scripts for sequence-based analyses. Programs for text information extraction and processing are shown in grey boxes. Arrows indicate the information flow. Abbreviations: RTS, representative transcript set; VRS, variant repeat set; VRTS, variant representative transcript set; OMIM, Online Mendelian Inheritance in Man; MeSH, Medical Subject Headings

(RN) field of the abstract contained the words 'protein', 'proteins', 'DNA', 'RNA', 'gene' or 'nucleotide' the relation between the gene (name) and disease (MeSH) was often found to be indirect. If the RN field words (partially) matched the gene name (eg 'ABC protein' matches 'ABC') or symbols, the extracted disease MeSH term is probably directly associated with the gene name or symbol (label: direct). In the following sections reasoning and results of key steps are explained.

REPEAT IDENTIFICATION

With the exception of internal protein repeats,[22] the role and effect of nucleotide sequence repeats in cDNAs with protein-coding potential have not been systematically investigated. In consequence, only 74,031 of 104,028 RTS source sequences that had CDS information were searched with RepeatMasker (-mus -a -xm -excln -pa 16) against Repbase 7.2. The caveat of using RepeatMasker is that it detects only major simple di- to hexameric repeats having more than 20 nucleotides. A small number of simple repeats that could have been detected with specialised programs such as Tandem Repeat Finder[23] were missed in a general repeat detection strategy. Likewise, a small faction of potential new mouse repeats that do not have a consensus in the repeat library of RepeatMasker may not be or may be incorrectly identified and positioned.[24] With the Smith–Waterman score threshold set to ⩾225, 40,701 repeats were identified and excised from 21,808 repeat-containing sequences with CDS information, corresponding to 14,229 repeats in 7,027 representative transcripts.

RepeatMasker and its caveats

Ancestral and species-specific repeats

To identify repeats with variable length, substitutions or insertions/deletions among members (eg from different tissues) a Perl program was used to count the number of differences in repeats of the same type among representative cluster members. If several cluster members contained identical variant repeats, the longest cDNA was listed as variant repeat transcript set (VRTS) member and its variant repeat sequence as VRS member. All repeat-containing singletons (cDNAs that did not cluster) were included in the VRS and VRTS. The VRS contains 31,396 repeats derived from 16,527 VRTS members. Using position information of the repeat and CDS it turned out that 12.2 per cent (3,819) of VRS were located in or overlap with the CDS region of 18.7 per cent (3,089) VRTS members. About 50 per cent of VRS are transposon-derived elements, whereas 33 per cent belonged to simple repeats. The majority of repeats (75 per cent VRS) were located in the 3′UTR of 80.8 per cent VRTS members. Nearly half of the 23,552 UTR repeats consist of SINEs, raising the chance of introducing additional poly(A) signals. Depending on the location and distance of the insertion in the context of the neighbouring gene, SINE B2 may affect the transcription of adjacent gene by doubling as a pol II promoter.[25]

Variant repeats

REPEAT CONSERVATION

BLAT mapping of repeat-containing mouse cDNA sequences to the mouse and human genome assemblies and BLASTN[26] searches against primate and rodent (except mouse) cDNA sequences of GenBank establishes potential orthologous or homologous relationships, including absence or conservation of repeats in other species. Some 2,818 VRTS sequences encompassing 3,364 variant repeats aligned to both mouse and human (>60 per cent aligned cDNA length; <3 gaps and 1 mismatch per tile) genome assemblies. As expected simple repeats (78.7 per cent) prevailed among orthologues harbouring ancestral repeat candidates. Similarly, 48.8 per cent of 20,995 variant repeats in 11,584 VRTS with BLASTN hits (E-50; excluding repeat sequence matches of 1–5 bp) to human homologues comprise simple repeats. The small percentage of conserved interspersed repeats (eg 13 per cent LINE) reflects their species-specific

expansion and diversification. In contrast 54 per of 4,688 variant repeats of 2534 VRTS members that matched by BLASTN to rat cDNA sequences belong to the SINE and 20 per cent to LINE repeat classes. Since the information was integrated, the FREP database users can assess individual repeat-containing cDNAs for the biological significance of an absent or present repeat in other species.

REPEATS AND SPLICING

Comparison of the genomic exon positions with repeat positions in the cDNA can reveal spliced repeats. If the repeat sequence terminates at the splice site of the genomic exon the repeat in the cDNA is truncated. If the repeat spans the splice junction (exon–intron–exon) the repeat in the cDNA consists of two adjacent truncated repeat stretches. In total 1,401 repeats in 1,253 (40.6 per cent) VRTS members were found to be spliced. About 75 per cent of spliced repeats are SINEs (33.8 per cent), LTRs (25.2 per cent) and simple repeats (16.3 per cent). Some 615 (43.8 per cent) spliced repeats were located in the CDS of 579 (18.7 per cent) VRTS members. Most of them were spliced in-frame, revealing the selection pressure on insertion or exon shuffling events. The actual number of spliced repeats is likely to be higher. A repeat sequence that resides mostly in an intron and covers only a few bases of the adjacent exon(s) may not be detected by RepeatMasker in the cDNA sequence. To capture all spliced repeats it would be necessary to identify the repeat start and end in the entire genomic region of the cDNA-genome alignment.

Majority of spliced repeats are SINEs

Repeats contribute poly(A) sites

REPEATS AND ALTERNATIVE SPLICING

Insertion of interspersed repeats or expansion of simple repeats may generate alternative splice sites. The resulting alternate exon may affect the length of protein-coding sequence, abrogate protein functions or give rise to proteins with new functions and properties. One can identify repeats in alternative splice sites either from variant genomic exon positions of genome-aligned cDNAs or by extracting variant exons from existing databases such as MousDB.[27] Pairwise alignment of repeat-containing cDNAs to the variant exon sequences of 4,750 variant clusters of MousDB showed that 239 repeats of 229 (7.4 per cent) VRTS contributed to 4.6 per cent (219 of 4,750) of variant clusters.

Alternate exons generated by repeats

REPEATS AND POLYADENYLATION

Repeats, particularly transposable elements, can generate additional poly(A) signals in the 3′UTR that can alter the expression pattern of a genes. Alternate poly(A) signals may influence post-transcriptional processing of precursor RNA to mature mRNA by generating multiple, alternatively polyadenylated mRNA species.[28] AATAAA is the most frequently occurring poly(A) signal in the 3′UTR. Alternative signals ATTAAA, AATTAA, AAATAA, AGTAAA, AATATA, CATAAA, TAATAA and AATAAT are less frequently used. Poly(A) sites of repeats in the 3′UTR were identified with a Perl script that matches poly(A) hexamers stored in a hash table, using a sliding window of 6 bp size. Some 2,902 repeats located in the 3′UTR of 23,552 VRTS sequences included 1,884 conserved AATAAA poly(A) sites. In total 5,540 poly(A) sites in 4,355 VRTS sequences were supplied by repeats, with 70 per cent derived from LTRs (1,100) and SINEs (2,823). Since the LTR and SINE content varies among species, the contribution of additional poly(A) signals by transposable elements may result in transcriptional differences that require attention when extrapolating biological effects from mouse to human.

DISEASE MESH TERMS AND OMIM MORBIDMAP TITLES

The presence of a gene name in the OMIM title or of disease MeSH terms in an abstract containing the query gene

Inferring repeat-disease associations

name would be misleading towards the involvement of repeats in disease. To avoid a large number of false positive assignments three provisions must be met: (1) the repeat is conserved in human and mouse genome mapping results, (2) OMIM Morbidmap and MEDLINE queries with gene name or symbol yielded an OMIM title or a disease MeSH term in the MH field of the retrieved abstracts, and (3) one of the other functional associations is positive. In fact, 1,195 cDNAs would have an OMIM-derived disease candidate association. When applying the above conditions, the number of candidates with repeat-disease association candidates decreased to 65. Some 235 (8.2 per cent) of the orthologous VRTS comprise 271 computationally inferred human disease-associated VRS repeats. The majority of repeat-disease candidates is associated with simple triplet repeats, including established human disease associations of CAG repeats (eg androgen receptor) .[29] Among the new candidates is the polycomb-group gene early developmental regulator 1. *Edr1* (NM_007905) contains a CAG repeat which is translated into 21 glutamine residues in the mouse and 15 in the human. Depending on the length, polyglutamine tracts are known to interfere with certain transcription factors. Since there are no experimental data available, it remains to be seen whether Edr1 poly(Q) expands and affects developmental processes such as organogenesis.[30]

Protein motif modification by repeats

Triplet repeat disease associations

REPEATS AND PROTEIN FUNCTION

Interspersed repeats that are translated may alter the function of a protein in mouse compared with the same protein without a repeat in human. Simple repeats may change the protein function by differences in length and base composition between mouse and human. CDS and repeat position information was used to categorise repeats in relation to the CDS. Some 2,881 VRS in 2,331 VRTS were designated as potentially translated. Of these, 1,160 (40.2 per cent) were spliced.

Translated repeat

Rule-based funtional repeat assignment

REPEATS AND PROTEIN MOTIFS

Some 460 (16 per cent of 2,881) potentially translated VRS in 418 VRTS members harbour 216 non-redundant protein motifs detected by InterProScan. Translated repeats without matches to InterPro[31] entries may include either new domain/motif candidates or partially overlapping known domains that were not detected because of the threshold setting. The majority of translated repeats with InterPro motifs were derived from triplet and tandem repeats. Proline-rich regions (IPR000694) and bipartite nuclear localisation signals (IPR001472) occurred in 45 per cent (188) VRTS members. More interesting are repeat-motif overlaps or motif truncations by transposable elements. For example, flavin containing monooxygenase-1 (AK042457) has an IAP element of the LTR ERV-K family. The IAP provided an alternative splice site which led to a splice variant, causing a premature stop codon in the flavin-containing monooxygenase (FMO) domain (IPR IPR000960). The shortened FMO may alter FMO-dependent drug metabolism.[32]

FREP DATABASE

The various outputs of the production pipeline were integrated into FREP database[16] which serves as a decision basis for selecting true candidates that justify experimental validation. To achieve query functionalities and a comprehensive report that satisfy biological decision-making, specifications for integration and decision rules for functional repeats in cDNAs were designed by biomedical experts. The rules require (1) the occurrence of genomic exon–exon boundaries in repeats, (2) the presence of polyadenylation sites in 3′UTR-located repeats, (3) effect on translation, (4) position in the protein-coding region or protein domains or (5) the conditional

association of repeats with OMIM titles and disease MeSH terms extracted from MEDLINE abstracts. At present the FREP database contains 9,261 non-redundant computationally inferred functional repeats derived from 6,861 mouse cDNAs.

FREP as hypothesis generator

Key factors that aid biological interpretation and curation of the repeat information are interfaces with multiple intuitive query options. FREP offers up to 18 options for broad or highly specific querying and filtering queries by combining repeat classification, length, Smith–Waterman score, repeat-CDS relation, FREP definitions, chromosomal location or repeat conservation among primates and rodents and query reports with an integrated view of prioritised biological information.

FREP REPORT – STARTING POINT FOR NEW HYPOTHESES

The panels in Fig. 2 show an example of the FREP report for tumour necrosis factor superfamily member 13b (*Tnfsf13b*), commonly called *Baff*. *Tnfsf13b* encodes a type II transmembrane protein belonging to the TNF superfamily of cytokines. *Tnfsf13b* (NM_033622) carries in the CDS a 93 bp LTR MaLR repeat. The cDNA-genome alignment revealed that the repeat is spliced in-frame and contributes the entire exon 3. Pre-computed BLAST searches of primate and rodent (without mouse) GenBank cDNA sequences did not reveal the repeat in human and rat (Panel 2; BLAST). Manual BLAST searches with the excised repeat sequence and mouse *Tnfsf13b* confirmed the absence of the repeat in human *Tnfsf13b*. Disease MeSH extraction from MEDLINE abstracts correctly associated *Tnfsf13b* with rheumatoid arthritis and systemic lupus erythematosus-type diseases (panel 6). Although the computational inferred information is not detailed, it sparked sufficient interest to perform further comparative analyses of mouse and human *Tnfsf13b*.

Cross-species comparison of FREP-containing transcripts

Repeats as evolutionary strategy to create complexity

In humans, the functionally active form of TNFS13B is a trimer that binds to TNFRSF13B (TACI), TNFRSF13C (BAFF-R), TNFRSF17 (BCMA) and is involved in B-cell survival and maturation.[33–34] Since the LTR MaLR gives rise to a longer N-terminal sequence of the soluble Tnfsf13 in mouse than in human, it is possible that different conformation may affect trimerisation and/or receptor binding. As there is growing evidence of human-soluble TNFSF13B as a key molecule in human systemic lupus erythematosus-type autoimmune diseases it will be interesting to see whether the repeat in mouse has any protective effect towards lupus-like autoimmune diseases by affecting the signalling network of its receptors.

Mouse *Tnfsf13b* does not appear to be an odd exception of translated repeats. At least another seven immune-related transcripts (*Ilf3*, *Zbp1*, *Tgfbr1*, *Glrp1*, *Bcl-x-γ*, *Phlda1* and *Rfx4*) in FREP database contain repeat elements that cause alternative splicing and/or potential modulation of protein functions in mouse. These examples demonstrate that analyses of repeat-containing protein-coding transcripts with a focus on cross-species comparisons will add new insights into the plethora of subtle functional differences of gene products between mouse and human.

CONCLUSION

The effects of repeats on gene functions are often mediated by the combination of repeat type, composition and location. For instance, the $(GAA\text{-}TTC)_n$ triplet repeat expansion in the first intron of frataxin slows transcription and translation by changing the structure of the DNA template,[37] whereas the differential distribution of Alu repeats is known to facilitate somatic recombinations, leading to a number of cancer-causing genomic deletions or chromosomal rearrangements.[38] Considering the abundance and redundancy of repeats in mammalian genomes, our knowledge about their biological

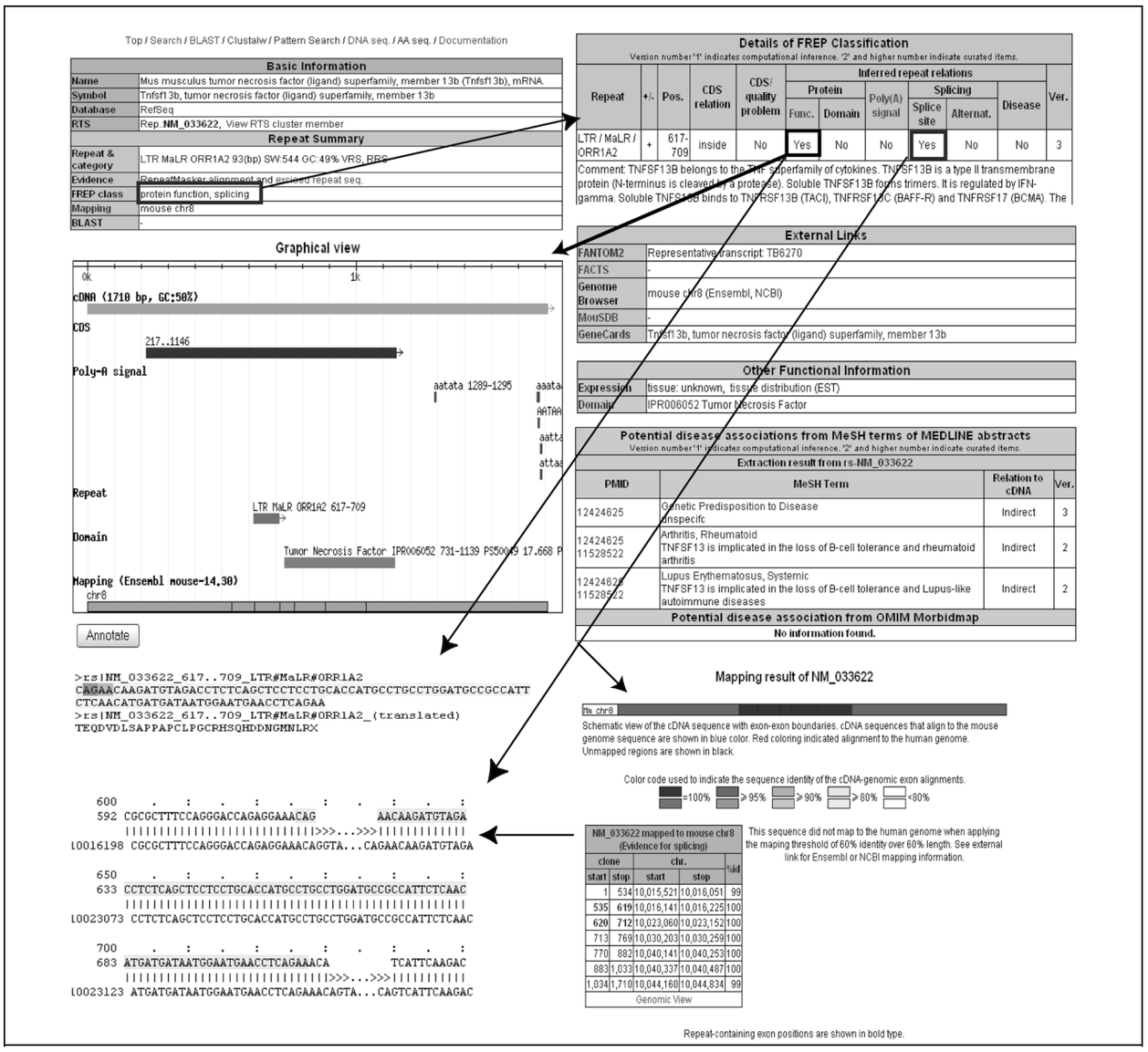

Figure 2: FREP report for mouse *Tnfsf13b* consists of 6 panels. Panels 1–3 are shown on the left and panels 4–6 on the right side. Arrows from the hyperlinks point to additional evidence for splicing and effect on protein functions. The entire exon 3 consists of an LTR/MaLR element. The complete report including curator comments can be viewed at http://facts.gsc.riken.go.jp/FREP/cgi/view.cgi?AC=NM_033622

meaning and their role as evolutionary strategy that generates complexity of genomic networks is still rudimentary. The current FREP strategy of placing repeats in biological relevant context information is neither perfect nor complete, but rather provides a foundation for building new multidimensional data analysis tools to dissect the relationships of repeat sequences with epigenetic data, differential use of promoters, or gene expression in context of genomic regulatory networks.

Acknowledgments

The author wishes to thank all FREP database members for their helpful comments.

References

1. Prak, E. T. and Kazazian H. H. Jr (2000), 'Mobile elements and the human genome', *Nature Rev. Genet.*, Vol. 1(2), pp. 134–144.
2. Epplen, J. T., Maueler, W. and Santos, E. J.

(1998), 'On GATAGATA and other "junk" in the barren stretch of genomic desert', *Cytogenet. Cell Genet.*, Vol. 80(1–4), pp. 75–82.

3. Richards, R. I. and Sutherland, G. R. (1994), 'Simple tandem repeats are not replicated simply', *Nature Genet.*, Vol. 6(2), pp. 114–116.
4. Lander, E. S., Linton, L. M., Birren, B. *et al.* (2001), 'Initial sequencing and analysis of the human genome', *Nature*, Vol. 409(6822), pp. 860–921.
5. Jurka J. (1998), 'Repeats in genomic DNA: Mining and meaning', *Curr. Opin. Struct. Biol.*, Vol. 8(3), pp. 333–337.
6. Jurka, J. (2000), Repbase update: A database and an electronic journal of repetitive elements', *Trends Genet.*, Vol. 16(9), pp. 418–420.
7. Waterston, R. H., Lindblad-Toh, K., Birney, E. *et al.* (2002) 'Initial sequencing and comparative analysis of the mouse genome', *Nature*, Vol. 420(6915), pp. 520–562.
8. Ostertag, E. M. and Kazazian, H. H. Jr (2001), 'Biology of mammalian L1 retrotransposons', *Annu. Rev. Genet.*, Vol. 35, pp. 501–538.
9. Paces, J., Pavlicek, A., Zika, R. *et al.* (2004), 'HERVd: The Human Endogenous RetroViruses Database: Update', *Nucleic Acids Res.*, Vol. 32, Database issue, p. D50.
10. Dagan, T., Sorek, R., Sharon, E. *et al.* (2004), 'AluGene: A database of Alu elements incorporated within protein-coding genes', *Nucleic Acids Res.*, Vol. 32, Database issue, pp. D489–492.
11. Collins, J. R., Stephens, R. M., Gold, B. *et al.* (2003). 'An exhaustive DNA micro-satellite map of the human genome using high performance computing', *Genomics,* Vol. 82(1), pp. 10–19.
12. Litt, M. and Luty, J. A. (1989), 'A hypervariable microsatellite revealed by *in vitro* amplification of a dinucleotide repeat within the cardiac muscle actin gene', *Amer. J. Hum. Genet.* Vol. 44(3), pp. 397–401.
13. Toth, G., Gaspari, Z. and Jurka, J. (2000), 'Microsatellites in different eukaryotic genomes: Survey and analysis', *Genome Res.*, Vol. 10(7), pp. 967–981.
14. Edwards, A., Civitello, A., Hammond, H. A. *et al.* (1991), 'DNA typing and genetic mapping with trimeric and tetrameric tandem repeats', *Amer. J. Hum. Genet.*, Vol. 49(4), pp. 746–756.
15. La Spada, A. R. and Taylor, J. P. (2003), 'Polyglutamines placed into context', *Neuron*, Vol. 38(5), pp. 681–684.
16. Horng, J. T., Lin, F. M., Lin, J. H. *et al.* (2003), 'Database of repetitive elements in complete genomes and data mining using transcription factor binding sites', *IEEE Trans Inf. Technol. Biomed.*, Vol. 7(2), pp. 93–100.
17. Hamosh, A., Scott, A. F., Amberger, J. *et al.* (2002), 'Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders', *Nucleic Acids Res.*, Vol. 30(1), pp. 52–55.
18. Nagashima, T., Matsuda, H., Silva, D. G. *et al.* (2004), 'FREP: A database of functional repeats in mouse cDNAs', *Nucleic Acids Res.*, Vol. 32, Database issue, pp. D471–475.
19. Nagashima, T., Silva, D. G., Petrovsky, N. *et al.* (2003), 'Inferring higher functional information for RIKEN mouse full-length cDNA clones with FACTS', *Genome Res.*, Vol. 13(6b), pp.1520–1533.
20. Okazaki, Y., Furuno, M., Kasukawa, T. *et al.* (2002), 'Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs', *Nature*, Vol. 420(6915), pp. 563–573.
21. Kent, W. J. (2002), 'BLAT – the BLAST-like alignment tool', *Genome Res.*, Vol. 12(4), pp. 656–664.
22. Florea, L., Hartzell, G., Zhang, Z. *et al.* (1998), 'A computer program for aligning a cDNA sequence with a genomic DNA sequence', *Genome Res.*, Vol. 8(9), pp. 967–974.
23. Zdobnov, E. M. and Apweiler, R. (2001), 'InterProScan – an integration platform for the signature-recognition methods in InterPro', *Bioinformatics*, Vol. 17(9), pp. 847–848.
24. Marcotte, E. M., Pellegrini, M., Yeates, T. O. *et al.* (1999), 'A census of protein repeats', *J. Mol. Biol.*, Vol. 293(1), pp. 151–160.
25. Benson, G. (1999), 'Tandem repeats finder – a program to analyze DNA sequences', *Nucleic Acids Res.*, Vol. 27(2), pp. 573–580.
26. Bao, Z. and Eddy, S. R. (2002), 'Automated *de novo* identification of repeat sequence families in sequenced genomes', *Genome Res.*, Vol. 12(8), pp. 1269-1276.
27. Ferrigno, O., Virolle, T., Djabari, Z. *et al.* (2001), 'Transposable B2 SINE elements can provide mobile RNA polymerase II promoters', *Nature Genet.*, Vol. 28(1), pp. 77–81.
28. Altschul, S. F., Madden, T. L., Schäffer, A. A. *et al.* (1997), 'Gapped BLAST and PSI-BLAST: A new generation of protein database search programs', *Nucleic Acids Res.*, Vol. 25(17), pp. 3389–3402.
29. Zavolan, M., Kondo, S., Schönbach, C. *et al.* (2003), 'Impact of alternative initiation, splicing, and termination on the diversity of the mRNA transcripts encoded by the mouse transcriptome', *Genome Res.*, Vol. 13(6b), pp. 1290–1300.

30. Edwalds-Gilbert, G., Veraldi, K. L. and Milcarek, C. (1997), 'Alternative poly(A) site selection in complex transcription units: Means to an end?', *Nucleic Acids Res.*, Vol. 25(13), pp. 2547–2561.

31. Ding, D., Xu, L., Menon, M. *et al.* (2004), 'Effect of a short CAG (glutamine) repeat on human androgen receptor function', *Prostate,* Vol. 58(1), pp. 23–32.

32. Shirai, M., Osugi, T., Koga, H. *et al.* (2002), 'The Polycomb-group gene Rae28 sustains Nkx2.5/Csx expression and is essential for cardiac morphogenesis', *J. Clin. Invest.*, Vol. 110(2), pp. 177–184.

33. Mulder, N. J., Apweiler, R., Attwood, T. K. *et al.* (2003), 'The InterPro Database, 2003 brings increased coverage and new features', *Nucleic Acids Res.*, Vol. 31(1), pp. 315–318.

34. Rodriguez, R. J. and Acosta, D. Jr (1997), 'Metabolism of ketoconazole and deacetylated ketoconazole by rat hepatic microsomes and flavin-containing monooxygenases', *Drug Metab. Dispos.*, Vol. 25(6), pp. 772–777.

35. Liu, Y., Xu, L., Opalka, N. *et al.* (2002) 'Crystal structure of sTALL-1 reveals a virus-like assembly of TNF family ligands', *Cell*, Vol. 108(3), pp. 383–394.

36. Gavin, A. L., Ait-Azzouzene, D., Ware, C. F. *et al.* (2003), 'DeltaBAFF, an alternate splice isoform that regulates receptor binding and biopresentation of the B cell survival cytokine, BAFF', *J. Biol. Chem.*, Vol. 278(40), pp. 38220–38228.

37. Sakamoto, N., Chastain, P. D., Parniewski, P. *et al.* (1999), 'Sticky DNA: Self-association properties of long GAA.TTC repeats in R.R.Y triplex structures from Friedreich's ataxia', *Mol. Cell.*, Vol. 3(4), pp. 465–475.

38. Kolomietz, E., Meyn, M. S., Pandita, A. *et al.* (2002), 'The role of Alu repeat clusters as mediators of recurrent chromosomal aberrations in tumors', *Genes Chromosomes Cancer*, Vol. 35(2), pp. 97–112.