TRDB—The Tandem Repeats Database

Gelfand, Yevgeniy

Oxford University Press


http://hdl.handle.net/2144/3128

Boston University
TRDB—The Tandem Repeats Database

Yevgeniy Gelfand1,*, Alfredo Rodriguez4 and Gary Benson1,2,3,*

1Lab for Biocomputing and Informatics, 2Department of Computer Science and 3Department of Biology, Boston University, Boston, MA 02215, USA and 4Department of Neuroscience, Mount Sinai School of Medicine, New York, NY 10029, USA

Received August 15, 2006; Revised October 30, 2006; Accepted October 31, 2006

INTRODUCTION

Our understanding of the role of tandem repeats in DNA has grown significantly over the past 40 years. The discovery of satellite DNA in 1961 (1) prompted research into the properties of repetitive DNA and this eventually led to an understanding of the wide range of sizes and genomic locations of tandem repeats. One class, the microsatellites, was recognized early on as useful genomic markers and today they form the basis of DNA fingerprints in forensics. Even in the face of strong competition from the more numerous single nucleotide polymorphisms (SNPs), polymorphic tandem repeats including microsatellites, and also the longer patterned minisatellites or VNTRs (variable number of tandem repeats) remain as important tools in genetic testing and linkage analysis because, unlike SNPs, they frequently exhibit more than two high frequency copy-number-variant alleles and thus can have high heterozygosity rates.

Starting 15 or so years ago, it became widely recognized that tandem repeats are causally associated with human disease. Perhaps the most well-known disease-associated repeats are the trinucleotide tandem repeats which cause severe neurological syndromes including those associated with polyglutamine (CAG)n expansion, such as Spinobulbar muscular atrophy (2); Huntington’s disease (3); and Spinocerebellar ataxias types 1, 2, 3, 6 and 7; and those associated with expansion in non-coding regions, such as Fragile X mental retardation (4); Friedrech’s ataxia (5); Myotonic dystrophy (6); and Spinocerebellar ataxias types 8 and 12 (7,8).

Other, more common, affective disorders and addictive behaviors have been associated with longer unit tandem repeats. For example, variations in a 40 bp VNTR at the 3’ end of the dopamine transporter gene (DAT1) (9) have been linked to attention deficit hyperactivity disorder (ADHD) (10), medication response to that disorder in children (11), and response to amphetamine in adults (12). A 30 bp VNTR in intron 8 of the same gene has been linked to cocaine dependence (13). In the serotonin transporter gene (5-HTT), variations in a 20–23 bp VNTR in intron 2 have been associated with bipolar disorder. The transcription factors, YB-1 and CTCF, have been shown to interact with the VNTR and modulate differences in gene expression in different copy number variants (14). Numerous studies have linked common polymorphisms in a 20–23 bp VNTR in the promoter of the same gene with various affective disorders, including autism (15), and response to medication for depression (16). A common, non-neurological disease associated with tandem repeat polymorphism is type 1 diabetes which is linked to allelic variation in a 14–15 bp VNTR at the IDDM2 locus situated ~600 bp 5’ to the insulin gene (17,18).

Some or all of the effects of intronic and non-coding polymorphic tandem repeats are presumably mediated by changes in cis-regulation of gene expression. A non-human example occurs in maize where a large tandem repeat is required for paramutational suppression of the b1 transcription factor gene which affects plant pigmentation. The paramutagenic region, 100 kb upstream of the gene, contains seven tandem copies of an 853 bp motif, while alleles with fewer copies have decreased or no paramutational effect (19). The mechanism, which involves differential cytosine methylation within the repeat region, requires an RNA-dependent RNA polymerase (20) and thus may be related to RNA interference through repeat mediated formation of double stranded RNA.

*To whom correspondence should be addressed. Tel: +1 617 358 2965; Fax:+1 617 353 4814; Email: gbenson@bu.edu

© 2006 The Author(s).
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
Due to a variety of mechanisms that affect their stability, including slippage replication and unequal crossing over, tandem repeats can exhibit high-mutation rates and this property may yield plasticity in a species. In dogs, variations in copy number for trinucleotide tandem repeats found in the coding regions of developmental genes have been quantitatively associated with morphological variations in the foot and skull among different domestic breeds (21). The implication is that plasticity in these repeats has enabled the selective breeding of dogs to achieve widely divergent morphologies.

The foregoing examples of known functional roles for tandem repeats are suggestive of future discoveries. They also highlight the need for readily available computational resources to study repeats. The growing interest in tandem repeats in the late 1990s led one of us (Benson) to develop the Tandem Repeats Finder program in 1999 (22), one of several now used to rapidly identify approximate tandem repeats in genomic DNA. Despite that program’s usefulness and heavy usage (100 citations in 2005), what has been lacking is a more comprehensive computational resource. The Tandem Repeats Database (TRDB), described here, has been designed to fill that void. It consists of two parts: the first is a web accessible, public repository of information on the presence and characteristics of tandem repeats in a variety of genomes; the second is a research workbench which (hopefully) will serve as a model for future biological database development.

Currently, the public database contains 22 genomes, including six land vertebrates (human, chimpanzee, mouse, rat, dog, chicken), three fish (Tetraodon, zebrafish), seven insects (five Drosophila species, honeybee, mosquito), two roundworms (Caenorhabditis species), two plants (Arabidopsis, rice), Saccharomyces cerevisiae and Escherichia coli (see also Table 1). In addition, archival copies of some of these genomes are maintained. Other species are being added as they become available and as interest warrants. A variety of tools, built into TRDB, simplify the study of repeats. These include query and filtering capabilities for finding particular repeats of interest, repeat clustering algorithms based on sequence similarity, polymorphism prediction based on common patterns of mutation, an interface for PCR primer selection using the Primer3 software (23), and data download in a variety of formats. Along with the tools, TRDB provides data visualization features including dynamically generated histograms and scatterplots of repeat characteristics, a browser for visualizing repeats in the context of other sequence features, and alignment views which accentuate both patterns of mutations and sequence similarity among repeats.

The major design feature of the workbench is the user workspace which is a centralized storage space for user data and the results of analysis. The workspace permits users to collect public information, upload and analyze their own sequences, add sequence annotations, and store the results of analysis in projects and reports so that work may extend over multiple sessions. All the tools provided for the public data are available for use with private data as well. Most features of TRDB are available for anonymous use, and data stored anonymously in the workspace is generally available for a limited time (currently 7 days). Users have the option of registering with TRDB which gives access to several tools that require high-computational resources, such as repeat clustering and polymorphism prediction, and eliminates the time limit for data stored in the workspace. In addition, for registered users, TRDB facilitates sharing and exchange of information through a collaboration protocol. Collaborators may be added simply by supplying their user names in the system (email addresses) and can then share data and independently work on and view joint projects. Collaboration as implemented in TRDB eliminates the need for back-and-forth data transfer between colleagues and permits simultaneous multi-party viewing and analysis.

### DATA STORED IN TRDB

#### Repeats

The primary data stored in TRDB are tandem repeats as detected by the Tandem Repeats Finder (TRF) program (22). TRDB currently uses TRF version 4.0. Repeats stored in the Public Database are detected with default TRF parameter values. For repeat detection in user supplied sequences, other parameter settings are available (see Supplementary Data). Tandem repeats are organized into groups called *sets*. For most public genomes, TRDB maintains one set per chromosome and one set for mitochondrial repeats. All the repeats for a genome are additionally combined into a single set. Some genomes are incomplete and for these TRDB stores only what is currently available. Table 1 gives the total number of repeats stored for each of the public genomes in TRDB.

Sets of repeats are presented to the user in a table format. For each repeat, various descriptive characteristics are displayed. These are conceptually grouped into four categories described below. Figure 1 shows two partial tables from

<table>
<thead>
<tr>
<th>Genome</th>
<th>Repeats</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anopheles gambiae</td>
<td>84076</td>
</tr>
<tr>
<td>Apis mellifera</td>
<td>178586</td>
</tr>
<tr>
<td>Arabidopsis thaliana</td>
<td>26769</td>
</tr>
<tr>
<td>Caenorhabditis briggsae</td>
<td>45054</td>
</tr>
<tr>
<td>Caenorhabditis elegans</td>
<td>40331</td>
</tr>
<tr>
<td>Canis familiaris</td>
<td>899154</td>
</tr>
<tr>
<td>Danio rerio</td>
<td>906234</td>
</tr>
<tr>
<td>Drosophila melanogaster</td>
<td>29915</td>
</tr>
<tr>
<td>Drosophila mojavensis</td>
<td>181943</td>
</tr>
<tr>
<td>Drosophila persimilis</td>
<td>99895</td>
</tr>
<tr>
<td>Drosophila pseudobscura</td>
<td>76554</td>
</tr>
<tr>
<td>Drosophila yakuba</td>
<td>73939</td>
</tr>
<tr>
<td>Escherichia coli</td>
<td>58</td>
</tr>
<tr>
<td>Fugu rubripes</td>
<td>112902</td>
</tr>
<tr>
<td>Gallus gallus</td>
<td>136691</td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>947696</td>
</tr>
<tr>
<td>Mus musculus</td>
<td>1556231</td>
</tr>
<tr>
<td>Oryza sativa</td>
<td>134927</td>
</tr>
<tr>
<td>Pan troglodyte</td>
<td>726005</td>
</tr>
<tr>
<td>Rattus norvegicus</td>
<td>1363590</td>
</tr>
<tr>
<td>Saccharomyces cerevisiae</td>
<td>2075</td>
</tr>
</tbody>
</table>

Because of repeat reporting criteria in TRF, some duplication of repeat loci at different pattern sizes are included.
Figure 1. Repeat tables for the human genome (hg18 obtained from the UCSC genome browser website). Upper panel: TRF computed characteristics for repeats from chromosome 1. Filters applied were pattern size > 25, copy number > 5.0. Note that the third and fourth repeats redundantly report the same locus because of TRF reporting criteria (the larger pattern gives at least a 10% better alignment score). Note also that repeats with pattern sizes 25 and 26 are essentially AT repeats with some slight variations. This can be inferred from the percentage nucleotide columns and quickly checked by clicking on the repeat indices which opens an alignment window (see Figure 2 for an example). Middle panel: TRDB computed characteristics for repeats drawn from the entire genome. The repeats in this table have the potential to cause frame shift mutations. They have periods of 4 or 5 (rather than a multiple of 3), are contained exclusively in exons, and have a high percent matching which is typical of microsatellites that undergo replication slippage. The 4 unchecked repeats contain at least 14 exact copies in a row (as determined by visual inspection of their alignments). Lower panel: The filter used to obtain the middle table.
human chromosome I. The first contains characteristics primarily determined by TRF analysis and the second contains characteristics primarily determined by additional processing within TRDB. Users may select any combination of characteristics to view in a table. A complete description of all characteristics is given in the Supplementary Data.

- **Sequence characteristics** are based on the tandem array and the consensus pattern. The array is the entire sequence of the repeat. The consensus is estimated by TRDB to be the best pattern to align to the tandem array. The consensus pattern is not displayed in the repeats table, but may be obtained through data download.
- **Annotation characteristics** are obtained from annotation data which can be uploaded to TRDB. The characteristics table contains an indicator (yes or no) for each feature class (e.g. genes) indicating whether the repeat overlaps a member of the class. For those that do overlap, a hyperlink points to a description of the feature and a link to the external source database. For those repeats that do not overlap a member of the feature class, hyperlinks point to descriptions of the nearest features upstream and downstream and these descriptions include the distance in nucleotides from the repeat to the feature. This distance may be used in filtering, permitting queries that can, e.g. find all repeats within 10,000 nt of any gene.
- **Tool generated characteristics** are obtained from analysis by TRDB tools.
- **Identifier characteristics** help identify the source of the repeat and are useful when repeats from different sources are mixed in a single set.

**User data**

Three components make up the persistent data stored by a user: sequences, projects and reports. For TRF/TRDB analysis, a sequence must first be uploaded to the user workspace, either (i) as a FASTA file, (ii) by entering a GenBank accession number (for direct upload from GenBank), or (iii) by cutting and pasting. Multiple sequences in a single FASTA file are permitted, as are sequences with masked characters (Ns, upper case, lower case) or ambiguous characters (R, Y, etc.). Once stored, the following operations can be performed on a sequence:

- **TRF processing.** Repeats detected in the sequence are stored as a new set in a user project.
- **Annotations.** Locations of other features within a sequence may be uploaded as a file in General Feature Format (GFF), or by cutting and pasting. Annotated features can be used to filter a set of repeats by proximity to the features (see Filtering, Sorting and Merging) and their locations can be visualized in the browser tool (see Data visualization).
- **Sequence download.** The sequence or any single contiguous part of the sequence (specified by the starting and ending positions) may be retrieved as a FASTA format file. Repeats detected within the sequence can be masked (as Ns, upper case or lower case).
- **PCR primer selection.** Flanking sequence bordering any set of repeats may be retrieved for upload into primer selection software. Additionally an interface to the Primer3 software (23) is built directly into TRDB.

Every set of tandem repeats, whether detected in a user supplied sequence or selected and saved from the public data, is stored in a user project which forms the core for ownership and data sharing. TRDB produces a variety of visual and tabular data and any of these may be stored as static images in a report and supplemented with descriptive text. As with projects, reports are owned and can be shared with collaborators.

**FILTERING, SORTING AND MERGING**

A repeat set derived from a chromosome or other large sequence will typically contain thousands of repeats. By default, they are presented in order of occurrence along the sequence but may be sorted on any single characteristic in either ascending or descending order. To further tailor a set to the specifics of the research problem, TRDB provides filtering capabilities based on repeat characteristics. Using drop down menus and a text box, the user creates a collection of filter conditions and applies them to the set. Those repeats that meet all the conditions pass through the filter and can be saved as a new set. A distinctive property of TRDB is its ability to filter by proximity to annotated sequence features. This is accomplished by selecting a class of annotation features and requiring that the repeats either overlap one of the features or occur nearby, where nearby is expressed as a user-selected nucleotide distance upstream, downstream or in either direction (e.g. gene upstream within 10,000 nt). Repeats can also be selected manually for inclusion or exclusion in combination with other filters by checking or...
unchecking repeat label boxes. Figure 1 (lower panel) shows the expressions for a filter that finds short period repeats that could cause frameshift mutations: they are located in exons, have high percent matching which is typical of microsatellites that undergo replication slippage, and their unit sizes are not multiples of 3. The four repeats unchecked in the middle of Figure 1 contain at least 14 exact copies in a row (as determined by visual inspection of their alignments).

A new set of repeats can be produced by merging existing sets. For example, to create a set for the entire human genome, we merge the sets for the individual chromosomes using a union operation (i.e. \( A \cup B \)). Other allowed binary operations are intersection (\( A \cap B \)), complement of intersection [\( \text{not}(A \cap B) \)] and set difference (\( A - B \)). Set merging is possible in two modes. By default it is based on the repeat id, an internal TRDB identifier. In this mode, equality of repeats means equality of the identifiers, i.e. the repeats are actually the same, from the same run of TRF. The alternative is to merge based on tandem array position. In this case, two repeats are considered the ‘same’ if their tandem arrays are identical or they overlap by a user-specified percentage. This is useful in cases where the repeats come from different runs of TRF or the repeats are redundant. Associated with each merged set is an interactive tree diagram called the **history** which records and can display the merging conditions.

**DATA VISUALIZATION**

TRDB produces a variety of data visualizations, in .PNG format, which may be stored as static images in a report. Figure 2 shows TRDB’s visualization of the alignment of a repeat to its consensus pattern. This view is accessed by...
clicking the repeat indices in a repeats table or a repeat image in the browser. Figure 3 shows the multiple alignment of a set of related repeats. Up to 20 repeats may be displayed in this way. Mutiple alignments are appropriate for repeats related by sequence similarity and can be accessed from the ‘view repeats’ page.

For a repeat set, TRDB can produce a distribution histogram for any single numeric characteristic. The histogram can be presented as a graph or a table. In the case of a table, three values are returned per accumulation interval (bucket), the low and high ends of the interval range and the count for the interval. For any pair of characteristics, TRDB can produce a scatterplot of the ordered data points.

Histograms and scatterplots can be accessed from the ‘sets’ page. Figure 4 shows two histograms produced by TRDB.

The TRDB browser visualizes the occurrence of repeats along a source sequence in combination with the positions of other annotated features contained in the sequence. It was inspired by the UCSC Human Genome Browser but has more limited capability. Repeats and annotation features are displayed in separate horizontal strips. Within a strip, features are stacked if they would otherwise overlap. Each feature and repeat image contains a hyperlink. Resting the cursor on the image brings up a small text box with the feature name/id number. Clicking brings up a new window containing the feature description and an additional hyperlink for annotations to an external source database if available. Figure 5 shows a typical browser image. The browser can be accessed from the ‘sets’ page or from the entry for a single repeat in a repeats table.

**TRDB TOOLS**

**Data download**

TRDB provides datafile output for repeat sets in several formats. Each repeat is described by a collection of characteristics which can be modified by the user. Additionally, sequence information can be provided, including the tandem array (subsequence), the consensus (pattern), the repeat profile (24) (a summary of the alignment of the tandem array to its consensus in terms of the A, C, G, T and indel content of each alignment column) and flanking sequence on either side of the repeat (in several prespecified lengths from 50 to 1000 bp). Repeats are sorted, ascending or descending, based on any single numeric characteristic and may be grouped by source sequence for a multi-sequence set. The output format is one of four possibilities: (i) ASCII, either tab or comma delimited, for use in spreadsheet programs; (ii) XML; (iii) FASTA for sequence information only (subsequence, pattern, flanking sequence); and (iv) GFF or UCSC custom track (see Supplementary Data for additional details).

**Clustering**

This tool clusters repeats by sequence similarity, thereby identifying repeats that are evolutionarily related within a single genome, or across genomes, or which may have common functional or structural properties. The output is a partition of the original repeat set into a group of clusters, each containing at least two related repeats. Those repeats unrelated to any others are omitted from the partition. Clusters can be viewed from the ‘partitions’ page, by selecting a partition and then ‘view clusters’. Clusters are numbered arbitrarily and a table reports for each cluster the number of repeats it contains and the range of their consensus sizes. Each cluster is treated as a set and can be filtered, renamed and saved.

The clustering algorithm works with repeat profiles. Each element of the profile is the nucleotide and indel composition of one column in the alignment. Every pair of profiles is compared using a cyclic alignment algorithm (25) to produce a distance type alignment score for the pair. Several weighting functions for composition-to-composition scoring are available (26) and are still being tested. Alignment distance is
converted to a percent similarity through the formula
\begin{equation}
1 - \frac{\text{alignment distance}}{\text{maximum possible alignment distance}}.
\end{equation}

Connected components clustering is used to produce initial clusters with a percent similarity cut-off value (default = 85%). Clusters may be refined with the slower Partition Around Medoids (PAM) algorithm (26,27) which is a \(k\)-means approach. Figure 3 shows an example of related repeats detected by clustering.

**Polymorphism prediction**

As discussed in the Introduction, polymorphic repeats are useful as genomic markers and can cause differential gene expression. The prediction method used in TRDB is based on the method validated in (28). A minisatellite repeat is predicted to be polymorphic based on two factors, \(\%G + \%C \geq 0.48\) and \(\text{HistoryR} \geq 0.54\). The \(\text{HistoryR}\) value (a real number between 0 and 1) measures the levels of redundant mutations in the repeat (mutations that appear in the same position in several copies of the repeat) and redundant mutation motifs (the same or similar sets of mutations that appear in several copies of the repeat, see Figure 2). A larger number means more redundancy. The \(\text{HistoryR}\) value is computed by a parsimony-based duplication history reconstruction algorithm (29).

In the validation study (28), various sequence characteristics were tested as predictors of polymorphism and heterozygosity in 127 repeats from human chromosomes 21 and 22. The highest predictive values were obtained with the pair of factors stated above. Validation was done on minisatellites with the following characteristics (i) unit length \(\geq 17\) bp, (ii) copy number \(\geq 10\), (iii) total length \(\geq 350\) bp and (iv) percent matches \(\geq 70\)%.

The Polymorphism Prediction tool is run on a set of repeats. Only the set owner can run this tool, as it modifies some fields in the source repeats. Once complete, the results are stored in the ‘\text{HistoryR}’ and ‘\text{Predicted Polymorphism}’ characteristics. These must be added to the repeat table (with the ‘change columns’ button) in order to use them for filtering or sorting.

**FUTURE ENHANCEMENTS**

In the coming months, we will add enhancements to TRDB. These are expected to include the following:

- Pre-computed clusters of all repeats in the public database. Clustering will be performed within and across genomes. It is expected that a consensus or representative repeat will be selected for each cluster so that newly deposited repeats may be compared quickly to existing clusters.

- Inclusion of other repeat detection programs. These will allow search for tandem repeats by alternate methods. One program, mreps (30), is already available for detecting longer repeats than are possible with TRF. Another, STAR (31) will allow search for repeats with a particular motif. A function ‘import a set’ in the tools section has been implemented to allow external file upload of a repeat set detected by any means. It will be given more flexibility in terms of the allowed data file formats.

- Extended polymorphism prediction and annotation. Several other methods for computational polymorphism prediction have been published, both for microsatellites and minisatellites (32–34). We will add these methods to the polymorphism prediction tool already available in TRDB. In addition, we will cooperate with laboratory groups conducting polymorphism typing to include annotation data on known polymorphic tandem repeats.

**CONCLUSION**

TRDB is intended as a central resource for comprehensive information on tandem repeats in sequenced genomes and as a workspace providing essential computational tools for...
tandem repeat analysis. Our goal is to make TRDB an informative and innovative database. We thank those who have helped in the past through their suggestions which have improved the functionality of the database and we welcome new suggestions, even wildly ambitious ones, that will simplify or extend data analysis.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

ACKNOWLEDGEMENTS

This research was partially supported by National Science Foundation grants DBI-0090789, CCR-0073081, DBI-0413462 and IIS-0612153. Funding to pay the Open Access publication charges for this article was provided by Boston University.

Conflict of interest statement. None declared.

REFERENCES