D80–D87 *Nucleic Acids Research, 2007, Vol. 35, Database issue* *Published online 14 December 2006*
doi:10.1093/nar/gkl1013

TRDB—The Tandem Repeats Database

Yevgeniy Gelfand[1],*, Alfredo Rodriguez[4] and Gary Benson[1,2,3,]*

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

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

ABSTRACT

Tandem repeats in DNA have been under intensive study for many years, first, as a consequence of their usefulness as genomic markers and DNA fingerprints and more recently as their role in human disease and regulatory processes has become apparent. The Tandem Repeats Database (TRDB) is a public repository of information on tandem repeats in genomic DNA. It contains a variety of tools for repeat analysis, including the Tandem Repeats Finder program, query and filtering capabilities, repeat clustering, polymorphism prediction, PCR primer selection, data visualization and data download in a variety of formats. In addition, TRDB serves as a centralized research workbench. It provides user storage space and permits collaborators to privately share their data and analysis. TRDB is available at https://tandem.bu.edu/cgi-bin/trdb/trdb.exe.

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 trinucloetide 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); Friedreich'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 16–17 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

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 trinucloetide 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 (*Fugu*, *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.

Table 1. Tandem repeats in the public database genomes for pattern sizes from 1 to 2000 nt

Genome	Repeats
Anopheles gambiae (February 2003)	84 076
Apis mellifera (January 2005)	178 586
Arabidopsis thaliana (March 2004)	26 769
Caenorhabditis briggsae (July 2002)	45 054
Caenorhabditis elegans (March 2004)	40 331
Canis familiaris (July 2004)	899 154
Danio rerio (June 2004)	906 234
Drosophila melanogaster (April 2004)	29 915
Drosophila mojavensis (August 2004)	181 843
Drosophila persimilis (October 2005)	99 895
Drosophila pseudoobscura (August 2003)	76 354
Drosophila yakuba (April 2004)	73 939
Escherichia coli (October 2001)	58
Fugu rubripes (August 2002)	112 902
Gallus gallus (February 2004)	136 691
Homo sapiens (March 2006)	947 696
Mus musculus (March 2006)	1 556 231
Oryza sativa (Build 4)	134 927
Pan troglodyte (November 2003)	726 005
Rattus norvegicus (June 2003)	1 363 390
Saccharomyces cerevisiae (October 2003)	2075
Tetraodon nigroviridis (February 2004)	66 600

Because of repeat reporting criteria in TRF, some duplication of repeat loci at diffierent pattern sizes are included.

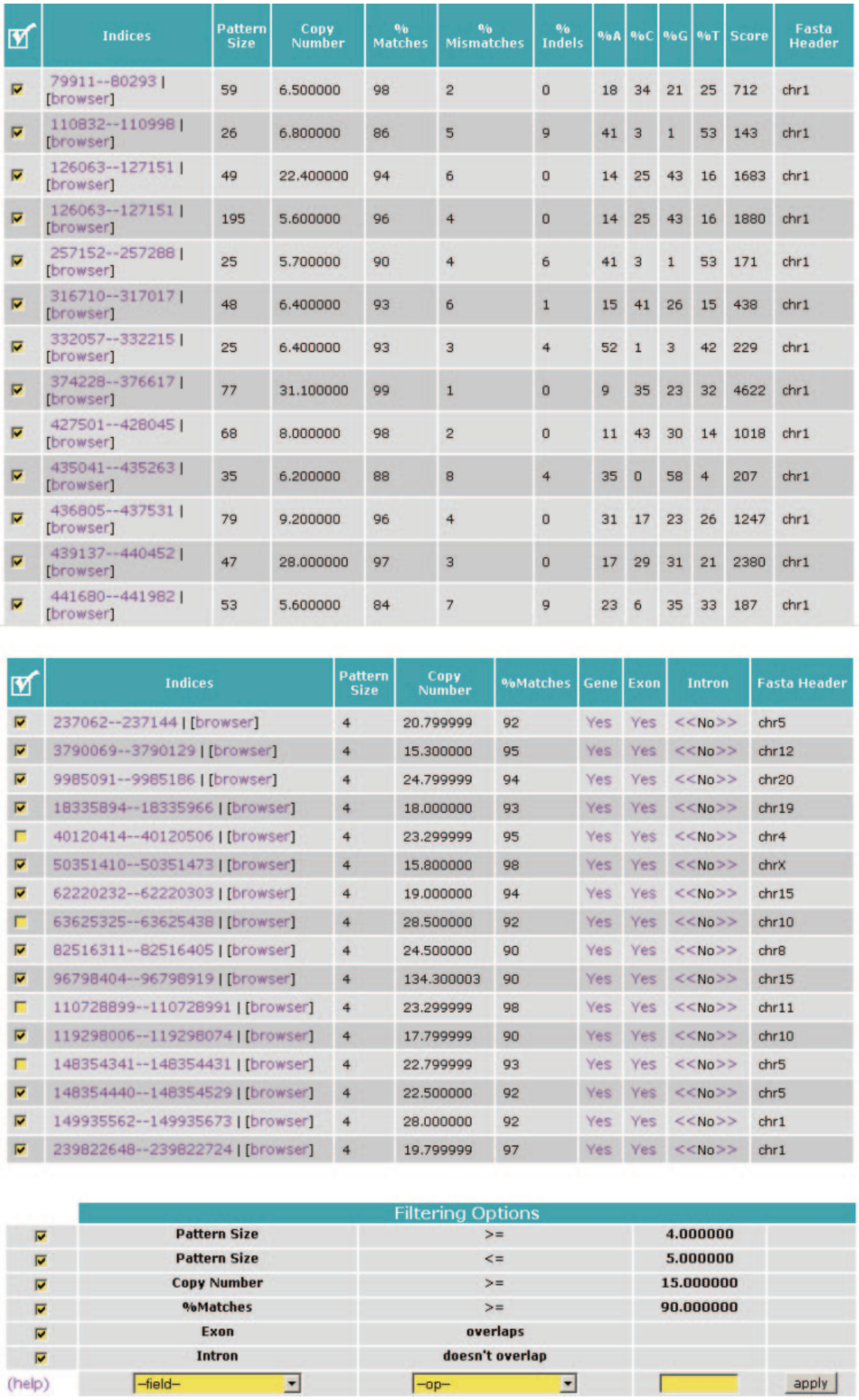
The major design feature of the workbench is the user workspace which is a centralized storage space for user data and the results of analysis. The workspace permits users to collect public information, upload and analyze their own sequences, add sequence annotations, and store the results of analysis 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



☑	Indices	Pattern Size	Copy Number	% Matches	% Mismatches	% Indels	%A	%C	%G	%T	Score	Fasta Header
☑	79911--80293 | [browser]	59	6.500000	98	2	0	18	34	21	25	712	chr1
☑	110832--110998 | [browser]	26	6.800000	86	5	9	41	3	1	53	143	chr1
☑	126063--127151 | [browser]	49	22.400000	94	6	0	14	25	43	16	1683	chr1
☑	126063--127151 | [browser]	195	5.600000	96	4	0	14	25	43	16	1880	chr1
☑	257152--257288 | [browser]	25	5.700000	90	4	6	41	3	1	53	171	chr1
☑	316710--317017 | [browser]	48	6.400000	93	6	1	15	41	26	15	438	chr1
☑	332057--332215 | [browser]	25	6.400000	93	3	4	52	1	3	42	229	chr1
☑	374228--376617 | [browser]	77	31.100000	99	1	0	9	35	23	32	4622	chr1
☑	427501--428045 | [browser]	68	8.000000	98	2	0	11	43	30	14	1018	chr1
☑	435041--435263 | [browser]	35	6.200000	88	8	4	35	0	58	4	207	chr1
☑	436805--437531 | [browser]	79	9.200000	96	4	0	31	17	23	26	1247	chr1
☑	439137--440452 | [browser]	47	28.000000	97	3	0	17	29	31	21	2380	chr1
☑	441680--441982 | [browser]	53	5.600000	84	7	9	23	6	35	33	187	chr1

☑	Indices	Pattern Size	Copy Number	%Matches	Gene	Exon	Intron	Fasta Header
☑	237062--237144 | [browser]	4	20.799999	92	Yes	Yes	<<No>>	chr5
☑	3790069--3790129 | [browser]	4	15.300000	95	Yes	Yes	<<No>>	chr12
☑	9985091--9985186 | [browser]	4	24.799999	94	Yes	Yes	<<No>>	chr20
☑	18335894--18335966 | [browser]	4	18.000000	93	Yes	Yes	<<No>>	chr19
☐	40120414--40120506 | [browser]	4	23.299999	95	Yes	Yes	<<No>>	chr4
☑	50351410--50351473 | [browser]	4	15.800000	98	Yes	Yes	<<No>>	chrX
☑	62220232--62220303 | [browser]	4	19.000000	94	Yes	Yes	<<No>>	chr15
☐	63625325--63625438 | [browser]	4	28.500000	92	Yes	Yes	<<No>>	chr10
☑	82516311--82516405 | [browser]	4	24.500000	90	Yes	Yes	<<No>>	chr8
☑	96798404--96798919 | [browser]	4	134.300003	90	Yes	Yes	<<No>>	chr15
☐	110728899--110728991 | [browser]	4	23.299999	98	Yes	Yes	<<No>>	chr11
☑	119298006--119298074 | [browser]	4	17.799999	90	Yes	Yes	<<No>>	chr10
☐	148354341--148354431 | [browser]	4	22.799999	93	Yes	Yes	<<No>>	chr5
☑	148354440--148354529 | [browser]	4	22.500000	92	Yes	Yes	<<No>>	chr5
☑	149935562--149935673 | [browser]	4	28.000000	92	Yes	Yes	<<No>>	chr1
☑	239822648--239822724 | [browser]	4	19.799999	97	Yes	Yes	<<No>>	chr1

	Filtering Options			
☑	Pattern Size	>=	4.000000	
☑	Pattern Size	<=	5.000000	
☑	Copy Number	>=	15.000000	
☑	%Matches	>=	90.000000	
☑	Exon	overlaps		
☑	Intron	doesn't overlap		
(help)	–field– ▾	–op– ▾		apply

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 $\geqslant 25$, copy number $\geqslant 5.0$. 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

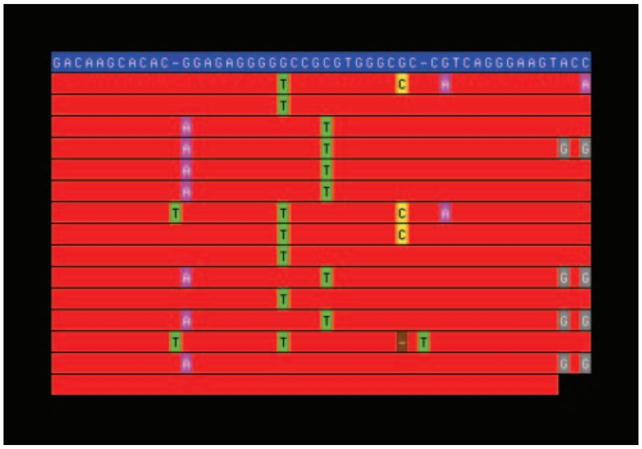


Figure 2. View of a tandem array aligned with its consensus pattern. This repeat is from human chromosome 5 (hg18, indices 720 890–721 608). The pattern size is 48 and the array contains 14.9 copies. The top line is the consensus. Dashes in the top line indicate an insertion relative to the consensus in one or more of the copies. Remaining bars represent consecutive copies in the tandem array. Only substitutions, insertions and deletions relative to the consensus are displayed as separate characters. When the repeat has a fractional copy, the final bar shows the point where the repeat terminates. This repeat is predicted to be polymorphic. Note that there are eight columns where repetitive point mutations occur and at least three groupings of these mutations suggest that the repeat has undergone multiple rounds of expansion. These are the A,T,G,G grouping that appears four times (one of which is in copy 4), the related A,T grouping that appears three times (one in copy 3) and the T,C,A grouping that appears twice (one in copy 1).

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 [*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

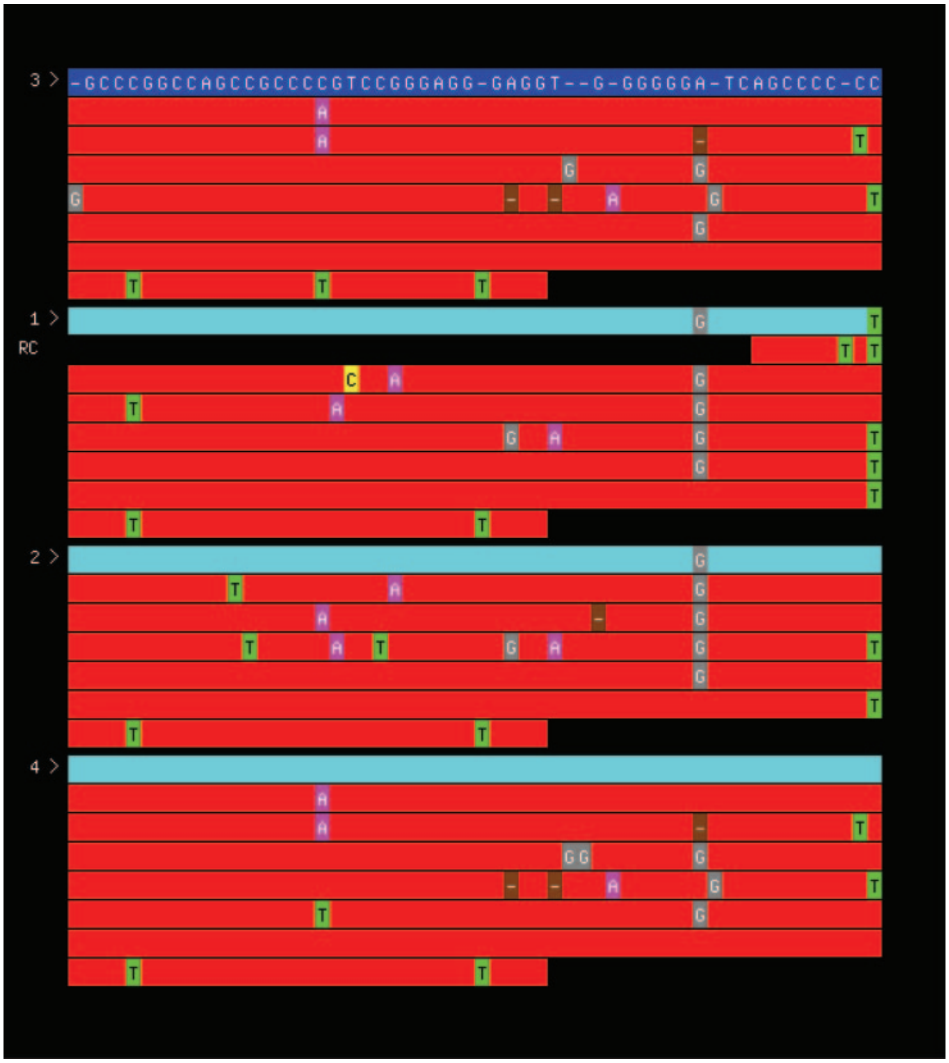


Figure 3. View of four related repeats found by the clustering tool, shown as a multiple alignment, from a cluster containing 19 repeats discovered in human chromosome 1 (hg18). These repeats exhibit minor variations, including differences in copy number and are widely spaced along the chromosome (1: 25, 032, 377–25, 032, 662; 2: 63, 711, 700–63, 711, 975; 3: 112, 638, 002–112, 638, 328; 4: 147, 399, 764–147, 400, 090). Note that repeat 1 is present in a reverse complement orientation (RC notation). The top repeat is considered the ''master'' and all alignments are to its consensus (top line). The master may be picked manually, or TRDB will choose it as the repeat with the smallest combined alignment distance to the remaining repeats. For every other repeat, a consensus bar is shown and the initial and final repeat bars show where the repeat starts and ends relative to the master consensus. Note that repeat 1 starts at a different position than the other three.

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.

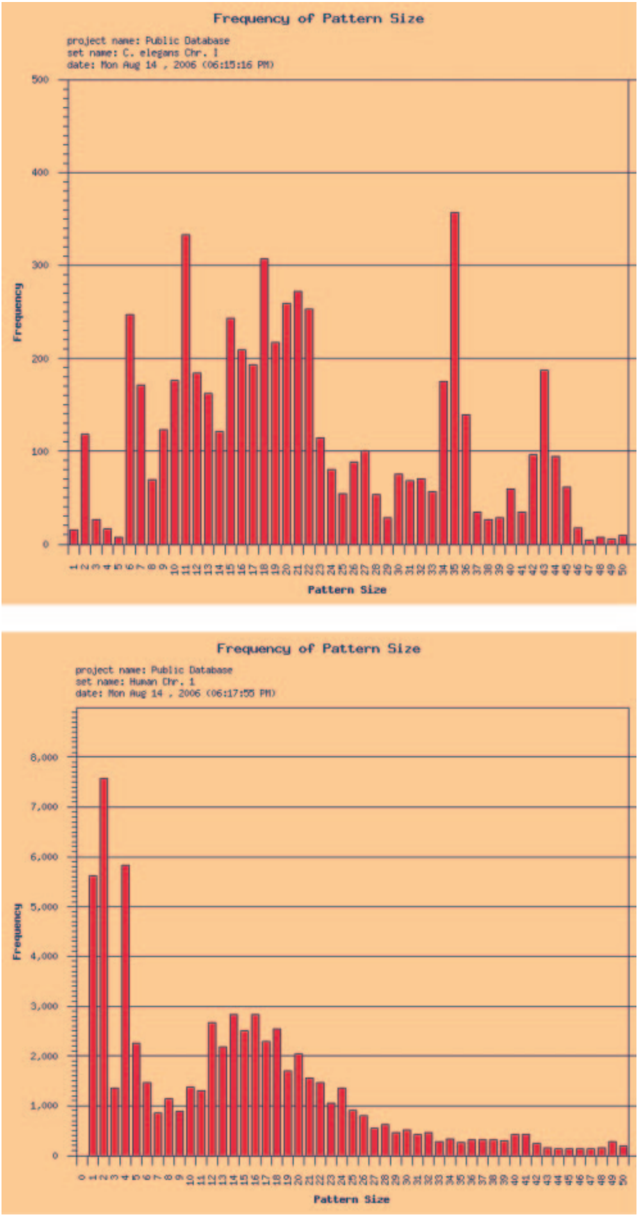


Figure 4. Histograms illustrating distinctly different distributions of tandem repeat pattern sizes in *C.elegans* (upper panel) and human (lower panel). Note the significant overrepresentation in humans of microsatellite repeats with periods 1, 2 and 4.

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

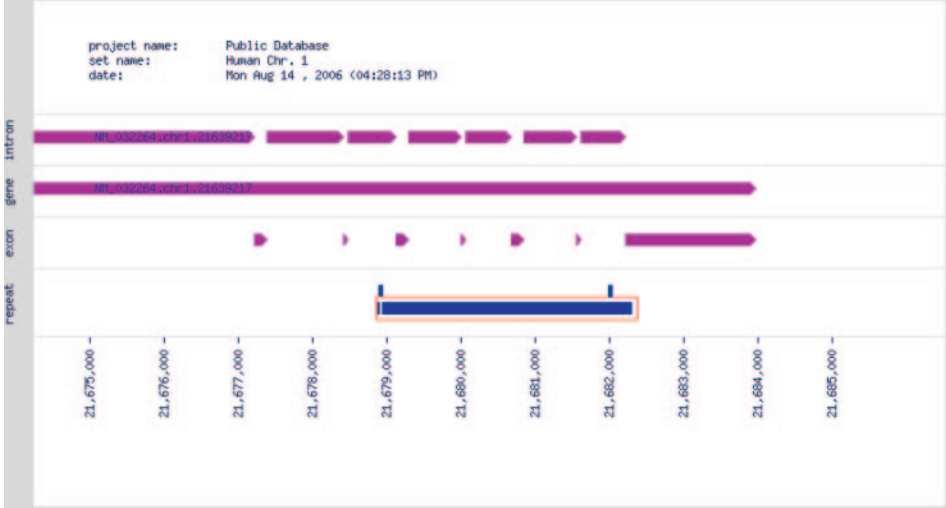
Data download

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



project name: Public Database
set name: Human Chr. 1
date: Mon Aug 14 , 2006 (04:28:13 PM)

Figure 5. A view from the browser. Here a single tandem repeat (hg 18, chromosome 1: 21, 678, 941–21, 682, 295), boxed, covers both introns and exons of a gene. The repeat has 2.2 copies of a pattern of 1583 nt. The periodic nature of the introns and exons can be clearly seen. This repeat clusters with six others on chromosome 1, some of which do not overlap any genes.

converted to a percent similarity through the formula

$$1 - \frac{(\text{alignment distance})}{(\text{maximum possible alignment distance})}.$$

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, $\%G + \%C \geqslant 0.48$ and $\text{HistoryR} \geqslant 0.54$. The 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 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 ⩾17 bp, (ii) copy number ⩾10, (iii) total length ⩾350 bp and (iv) percent matches ⩾70%. No data on the effectiveness of the prediction method for other repeats is currently available.

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 'HistoryR' and '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

1. Kit,S. (1961) Equilibrium sedimentation in density gradients of DNA preparations from animal tissues. *J. Mol. Biol.*, **3**, 711–716.
2. La Spada,A.R., Wilson,E.M., Lubahn,D.B., Harding,A.E. and Fischbeck,K.H. (1991) Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy. *Nature*, **352**, 77–79.
3. Huntington's disease collaborative research group (1993) A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell*, **72**, 971–983.
4. Verkerk,A.J., Pieretti,M., Sutcliffe,J.S., Fu,Y.H., Kuhl,D.P., Pizzuti,A., Reiner,O., Richards,S., Victoria,M.F., Zhang,F.P. *et al*. (1991) Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. *Cell*, **65**, 905–914.
5. Campuzano,V., Montermini,L., Molto,M.D., Pianese,L., Cossee,M., Cavalcanti,F., Monros,E., Rodius,F., Duclos,F., Monticelli,A. *et al*. (1996) Friedreich's ataxia: Autosomal recessive disease caused by an intronic GAA triplet repeat expansion. *Science*, **271**, 1423–1427.
6. Fu,Y.-H., Pizzuti,A., Fenwick,R.G.,Jr, King,J., Rajnarayan,S., Dunne,P.W., Dubel,J., Nasser,G.A., Ashizawa,T., DeJong,P. *et al*. (1992) An unstable triplet repeat in a gene related to myotonic muscular dystrophy. *Science*, **255**, 1256–1258.
7. Koob,M.D., Moseley,M.L., Schut,L.J., Benzow,K.A., Bird,T.D., Day,J.W. and Ranum,L.P. (1999) An untranslated CTG expansion causes a novel form of spinocerebellar ataxia (SCA8). *Nature Genet.*, **21**, 379–384.
8. Holmes,S.E., O'Hearn,E.E., McInnis,M.G., Gorelick-Feldman,D.A., Kleiderlein,J.J., Callahan,C., Kwak,N.G., Ingersoll-Ashworth,R.G., Sherr,M., Sumner,A.J. *et al*. (1999) Expansion of a novel CAG trinucleotide repeat in the 5′ region of PPP2R2B is associated with SCA12. *Nature Genet.*, **23**, 391–392.
9. Vandenbergh,D., Persico,A.M. and Uhl,G.R. (1992) A human dopamine transporter cDNA predicts reduced glycosylation, displays a novel repetitive element and provides racially-dimorphic *Taq*I RFLPs. *Mol. Brain Res.*, **15**, 161–166.
10. Cook,E.H.,Jr, Stein,M.A., Krasowski,M.D., Cox,N.J., Olkon,D.M., Kieffer,J.E. and Leventhal,B.L. (1995) Association of attention-deficit disorder and the dopamine transporter gene. *Am. J. Hum. Genet.*, **56**, 993–998.
11. Gilbert,D.L., Wang,Z., Sallee,F.R., Ridel,K.R., Merhar,S., Zhang,J., Lipps,T.D., White,C., Badreldin,N. and Wassermann,E.M. (2006) Dopamine transporter genotype influences the physiological response to medication in ADHD. *Brain*, **129**, 2038–2046.
12. Lott,D., Kim,S.J., Cook,E.H.,Jr and de Wit,H. (2005) Dopamine transporter gene associated with diminished subjective response to amphetamine. *Neuropsychopharmacology*, **30**, 602–609.
13. Guindalini,C., Howard,M., Haddley,K., Laranjeira,R., Collier,D., Ammar,N., Craig,I., O'Garag,C., Bubb,V.J., Greenwood,T. *et al*. (2006) A dopamine transporter gene functional variant associated with cocaine abuse in a Brazilian sample. *Proc. Natl Acad. Sci. USA*, **103**, 4552–4557.
14. Klenova,E., Scott,A.C., Roberts,J., Shamsuddin,S., Lovejoy,E.A., Bergmann,S., Bubb,V.J., Royer,H.-D. and Quinn,J.P. (2004) YB-1 and CTCF differentially regulate the 5-HTT polymorphic intron 2 enhancer which predisposes to a variety of neurological disorders. *J. Neurosci.*, **24**, 5966–5973.
15. Cook,E.H.,Jr, Courchesne,R., Lord,C., Cox,N.J., Yan,S., Lincoln,A., Haas,R., Courchesne,E. and Leventhal,B.L. (1997) Evidence of linkage between the serotonin transporter and autistic disorder. *Mol. Psychiatry*, **2**, 247–250.
16. Murphy,G.,Jr, Hollander,S.B., Rodrigues,H.E., Kremer,C. and Schatzberg,A.F. (2004) Effects of the serotonin transporter gene promoter polymorphism on mirtazapine and paroxetine efficacy and adverse events in geriatric major depression. *Arch. Gen. Psychiatry*, **61**, 1163–1169.
17. Owerbach,D. and Gabbay,K.H. (1993) Localization of a type 1 diabetes susceptibility locus to the variable tandem repeat region flanking the insulin gene. *Diabetes*, **42**, 1708–1714.
18. Bennett,S.T., Lucassen,A.M., Gough,S.C., Powell,E.E., Undlien,D.E., Pritchard,L.E., Merriman,M.E., Kawaguchi,Y., Dronsfield,M.J., Pociot,F. *et al*. (1995) Susceptibility to human type 1 diabetes at IDDM2 is determined by tandem repeat variation at the insulin gene minisatellite locus. *Nature Genetics*, **9**, 284–292.
19. Stam,M., Belele,C., Dorweiler,J.E. and Chandler,V.L. (2002) Differential chromatin structure within a tandem array 100 kb upstream of the maize b1 locus is associated with paramutation. *Genes Dev.*, **16**, 1906–1918.
20. Alleman,M., Sidorenko,L., McGinnis,K., Seshadri,V., Dorweiler,J.E., White,J., Sikkink,K. and Chandler,V.L. (2006) An RNA-dependent RNA polymerase is required for paramutation in maize. *Nature*, **442**, 295–298.
21. Fondon,J.W.,III and Garner,H.R. (2004) Molecular origins of rapid and continuous morphological evolution. *Proc. Natl Acad. Sci. USA*, **101**, 18058–18063.
22. Benson,G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*, **27**, 573–580.
23. Rozen,S. and Skaletsky,H (2000) Primer3 on the WWW for general users and for biologist programmers. In Krawetz,S. and Misener,S. (eds), *Bioinformatics Methods and Protocols: Methods in Molecular Biology*. Humana Press, pp. 365–386.
24. Gribskov,M., McLachlan,A.D. and Eisenberg,D. (1987) Profile analysis: Detection of distantly related proteins. *Proc. Natl Acad. Sci. USA*, **84**, 4355–4358.
25. Maes,M. (1990) On a cyclic string-to-string correction problem. *Information Processing Letters*, **35**, 73–78.
26. Rao,S., Rodriguez,A. and Benson,G. (2005) Evaluating distance functions for clustering tandem repeats. *Genome Inform.*, **16**, 3–12.
27. Kaufman,L. and Rousseeuw,P.J. (1990) *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley and Sons, New York.
28. Denoeud,F., Vergnaud,G. and Benson,G. (2003) Predicting human minisatellite polymorphism. *Genome Res.*, **13**, 856–867.
29. Benson,G. and Dong,L. (1999) Reconstructing the duplication history of a tandem repeat. In *Seventh International Conference on Intelligent Systems for Molecular Biology—ISMB99,* pp. 44–53.
30. Kolpakov,R., Bana,G. and Kucherov,G. (2003) mreps: efficient and flexible detection of tandem repeats in DNA. *Nucleic Acids Res.*, **31**, 3672–3678.
31. Delgrange,O. and Rivals,E. (2004) STAR: an algorithm to search for tandem approximate repeats. *Bioinformatics*, **20**, 2812–2820.
32. Naslund,K., Saetre,P., von Salome,J., Bergstrom,T.F., Jareborg,N. and Jazin,E. (2005) Genome-wide prediction of human VNTRs. *Genomics*, **85**, 24–35.
33. Wren,J., Forgacs,E., Fondon,J., Pertsemlidis,A., Cheng,S., Gallardo,T., Williams,R., Shohet,R., Minna,J. and Garner,H. (2000) Repeat polymorphisms within gene regions: phenotypic and evolutionary implications. *Am. J. Hum. Genet.*, **67**, 345–56.
34. Fondon,J.W.,III, Mele,G.M., Brezinschek,R.I., Cummings,D., Pande,A., Wren,J., O'Brien,K.M., Kupper,K.C., Wei,M.H., Lerman,M. *et al*. (1998) Computerized polymorphic marker identification: experimental validation and a predicted human polymorphism catalog. *Proc. Natl Acad. Sci. USA*, **95**, 7514–7519.