Methods for Repeat Detection In Nucleotide Sequences

Gary Benson

Computer Science, Biology, Bioinformatics
Boston University

gbenson@bu.edu
Outline

• Classes of Repeats in DNA
• Tandem Repeats
• Techniques for finding repetitive sequence
• Tandem Repeats Database
• Variant Tandem Repeats
Why Look at Repeats in DNA?

• Repeats make up the largest portion of DNA.
  – coding sequence (~5% of human DNA)
  – repetitive sequence (>50% of human DNA)
Classes of Repeats in DNA

• Interspersed repeats:
  – Retrotransposons
    • Sines:
    • Lines:
    • LTRs
  – Transposons
Classes of Repeats in DNA

• Inverted repeats
• Tandem repeats
  – Satellite repeats
  – microsatellites
  – minisatellites
  – VNTR (variable number of tandem repeats)
Tandem Repeats

A tandem repeat (TR) is any pattern of nucleotides that has been duplicated so that it appears several times in succession.

For example, the sequence fragment below contains a tandem repeat of the trinucleotide cgt:

tcgctggtcata cgt cgt cgt cgt cgt tacaacgttccggt
Approximate Tandem Repeats

More typically, the tandem copies are only approximate due to mutations. Here is an alignment of copies from a human TR from Chromosome 5.

<table>
<thead>
<tr>
<th>Indices</th>
<th>Period Size</th>
<th>Copy Number</th>
<th>%Matches</th>
<th>Array Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>105535--106695</td>
<td>49</td>
<td>23.700001</td>
<td>95</td>
<td>1161</td>
</tr>
</tbody>
</table>

Parameters: 2 7 7 80 10 50 2000
Organism: Homo sapiens
sequence: Human Chr. 5 (more info)
FASTA header: chr5

Shown are a consensus pattern and 23.7 copies
Why are tandem repeats interesting?

• They are associated with **human disease:**
  Fragile-X mental retardation
  Myotonic dystrophy
  Huntington’s disease
  Friedreich’s ataxia
  Epilepsy
  Diabetes
  Ovarian cancer

• They are often polymorphic, making them valuable **genomic markers.** Also, they may cause significant variation in the human population.

• They are involved in **gene regulation** and often may contain **transcription factor binding sites.**
LOCUS RATIGCA 4461 bp DNA ROD 18-APR-1994

DEFINITION Rat Ig germline epsilon H-chain gene C-region, 3' end.

```
2881 cgccccaagt aggcttcatac atgctcttttg gttaagcaat agcccaaaagc aagctatgca
2941 tccatctcag gcccagaggg atgaggagac cagaatcaag acataccccac gcccaccccac
3001 cgcccaacca ccaaccacca gcacatcagg ttcacacacc tgagaccagt ggcctccccac
3061 acacacacac acacacacac acacacacac acacacacac acacacacac ccgtacacatc
3121 ccacccatatc cagagacaag tgtcttgagtc tgagataacct ctcagagata ccaatggcag
3181 agtccgccccag cacctcagcc tccagcccc caacattatat tcgccccact gcagggccatc
3241 agagatggag gagggtggagg cctgagctgt ggaaaaccag agacagagag atgtctgtga
3301 ctcccagccca atcctttatac tttggccccac gcagggccatc gagagatgga ggagggtggag
3361 gcctgagctg tgtgaaaacct gagacagggaa gatggtctgtg atggagagag tagtaaaacca
3421 gattataggg agactgaggg cggagtagag cttcctacaag gccagttagtc taccttagag
3481 tctctataagtt ctgggcttggg agtccatgtg tcctgaccttg tcctcctgat atcacaacca
3541 agatccctgg agcagaggtg tgcatgcagc cccctagaaga aatgtggagcc tttaggccccct
3601 tctctggaggc cctggccacac tctgaacaaa cgcctgaccca gagcaattct gtaggtgtgta tagaggcac
3661 tggctcagata cacacacacca tgcacacacca tacacacacca gagacacacca cacaacacca
3721 tgccacacac cactgcatac cacaacacac cacaacacac cacaacacac cacacacacac
3781 catgcacccaca cacatgcatac cacaacatgca tgcacacacca cacaacacac tacacatca
3841 cacacacaca cacacccccac aggtagcctt catcgtgctg tctagcagata gcccctgctgga
3901 ggggtggaga tactggttca tgtggtggccac cggagtagaa agaggaatg agcagctcagg
3961 gtctccgagaa aagagcatct gcctccaggg ctaaagacag acttggagca gtcctccagac
4021 aagtgggatg gggagctcttg ccactccagtc ttcaccagga ctcgctgaga ccagtgaggg
```
LOCUS RATIGCA 4461 bp DNA ROD 18-APR-1994

DEFINITION Rat Ig germline epsilon H-chain gene C-region, 3' end.

2881 cgcccccaagt aggcttccatc atgctcttttg gtttagcaat agcccaaagc aagctatgca
2941 tccatctcag gcccagaggg atgagggac cagaatcaag acatacaccac gcccatccca
3001 acacacacac ccacacaccaca gcacatcagg ttcacacacc tggagcagt ggctccccatc
3061 ccacacacac acacacacac acacacacac acacacacac acacacacag ccgtacacat
3121 ccaccatatc cagagacagagc tgtctgagtc tgtgatacct ctgagagatca ccaatgacag
3181 agtcgggccag cacctcagcc tccaggccaa tctttatatc ttggcccaact gcaggccatg
3241 agagatggag gaggtggagg cctgcagctgt ggaaaccag agacaggaag atggtctgtga
3301 ctccaggcaca atccttatata tttgccccac tcgaggccact gagagatgga ggaggtgag
3361 gcctgagcttg tgtgaaaaccag gagacaggaag gatgggtctgt atggagagag tagtaaacca
3421 gattataggg agactgaggg cggagtagag tctctacaaag gccagtagtc taccttagag
3481 tccttataagt cttggcctggg agtccatgtg tctctgacttg tctctcagat atcacaacca
3541 agattcctgg agccagagtg tgcctgcagg tcctgactgca aatgtggagg cctgagctccct
3601 tcctggaggg ccctggccac tctgaaaacca aggcaattct tgtgaggtga tagaggcatac
3661 ctgagctgata cccacacaca tcgacacaca tacacacaca gagacacagac cacacacaca
3721 tgtggccacaca catgcataca cacacgacacac acacacacac acacacacag cagacacaca
3781 ccacatgcata cccagagagc ttgctgcctg tctagctgga gctcctgactga
3841 cagacacacac cccaccccccac aggtagcctt cgctagcctc tctctgagta gccctgagctga
3901 ggcgtggagaga tctggtggtca tgtgggaccc cggagtagaa agagggaatg agcagctcagg
3961 gctcaggggaa aagagcaatct gcctgcctgggt ctagagcagag acttggagcca ctgcccagacg
4021 aaggtggagatg ggagagctctg ccactccagct ttcaccagga ctgctgagttc caggtgagg
Tandem Repeats Finder

An online sequence analysis tool.

OR

A program to download and run locally.

Data from TRF is listed as “simple repeats” at the UCSC genome browser website.
Similarity Models

Match/mismatch model – Sequences differ only by mismatches:

Sequence 1: AAAGCTTTCGGAGTGCCCAG
Sequence 2: AATGCAATCGGGGTGCTCTAG
Similarity Models

Match/mismatch model – Sequences differ only by mismatches:

Sequence 1: \texttt{AAAGCTTCCGGAGTGCCCCGA}
Sequence 2: \texttt{AATGCAATCGGGGTGCCCTGA}

\begin{center}
\begin{tabular}{c}
11011011110111110111110111
\end{tabular}
\end{center}

Alignments of similar sequences can be represented by bit strings (zeros and ones).
Similiarity Models

Match/mismatch model – Sequences differ only by mismatches:

One model parameter required:

\[ p = \text{probability of matching letters in a column} \]
\[ = \text{probability of a 1 in the bit string} \]

Sometimes known as a Bernoulli (coin toss) model.
Similarity Models

Match/mismatch/indel model – adds indels to sequence differences:

Sequence 1:  AAAGCTTTCGG–AGT--GCCCGA
Sequence 2:  AA–GCATCGGGAGTTAGCCTGA
Similarity Models

Match/mismatch/Indel model – adds indels to sequence differences:

Sequence 1: \textcolor{red}{AAAGCTTTCGG–AGT--GCCC} \textcolor{red}{GA}
Sequence 2: \textcolor{red}{AA–GCATCGGGAGTTAGCCTG} \textcolor{red}{A}

\underline{1121101111211122111011}

Alignments of sequences can be represented by strings of numbers in \([0,1,2]\).
Similarity Models

Match/mismatch/Indel model – adds indels to sequence differences:

Sequence 1:   AAAGCTTCGG–AGT--GCCCGA
Sequence 2:   AA–GCATCGGGAGTAGCCCTGA

\[
\begin{array}{ccccccccccc}
1121101111311133111011 \\
\downarrow & \uparrow & \uparrow & \uparrow
\end{array}
\]

If the direction of insertion or deletion matters, we use strings of numbers in \([0,1,2,3]\).
Similarity Models

Match/mismatch/indel model – adds indels to sequence differences:

At least two model parameters required:

\( p \) = probability of matching letters in a column
\( = \) probability of a 1 in the numerical string

\( r \) = probability of an insertion or deletion
\( = \) probability of a 2 or 3 in the numerical string
Detecting Similar Sequences

Methods for similarity detection involve some form of scanning the input sequences, usually, with a window of fixed size. Information about the contents of the window is stored. This is called indexing.
Indexing

The index is a list of all possible window contents together with a list, for each content, of where it occurs:

AAA, AAC, ..., ACG, ..., CAG, ..., CGT, ..., GAC, ..., GTT, ..., TGC, ..., TTG, ... TTT
Building an Index

First sequence:

AAA, AAC, ..., ACG, ..., CAG, ..., CGT, ..., GAC, ..., GTT, ..., TGC, ..., TTG, ..., TTT
Building an Index

| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| A | C | G | T | T | G | C | A | G | T | T | G | A | C | T | G | A | C | G |

AAA, AAC, ..., ACG, ..., CAG, ..., CGT, ..., GAC, ..., GTT, ..., TGC, ..., TTG, ... TTT

↓

0
Building an Index

AAA, AAC, ..., ACG, ..., CAG, ..., CGT, ..., GAC, ..., GTT, ..., TGC, ..., TTG, ..., TTT

0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8
A C G T T G C A G T T G A C T G A C G

↓  ↓
0 1
Building an Index

AAA, AAC, ..., ACG, ..., CAG, ..., CGT, ..., GAC, ..., GTT, ..., TGC, ..., TTG, ..., TTT

0 1 2
Building an Index

\[
\begin{array}{cccccccccccccccc}
0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\
A & C & G & T & T & T & G & C & A & G & T & T & G & A & C & T & G & A & C & G
\end{array}
\]

AAA, AAC, ..., ACG, ..., CAG, ..., CGT, ..., GAC, ..., GTT, ..., TGC, ..., TTG, ..., TTT

\[
\begin{array}{cccc}
0 & 1 & 2 & 3 \\
\end{array}
\]
Building an Index

AAA, AAC, ..., ACG, ..., CAG, ..., CGT, ..., GAC, ..., GTT, ..., TGC, ..., TTG, ... TTT

0 1 2 3 4 5 6 7 8 0 1 2 3 4 5 6 7 8
A C G T T G C A G T T G A C T G A C G

AAA, AAC, ..., ACG, ..., CAG, ..., CGT, ..., GAC, ..., GTT, ..., TGC, ..., TTG, ... TTT

0 1 2 4 3
Building an Index

AAA, AAC, ..., ACG, ..., CAG, ..., CGT, ..., GAC, ..., GTT, ..., TGC, ..., TTG, ..., TTT

0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8
A C G T T G C A G T T G A C T G A C G

0 1 2 4 3

AAA, AAC, ..., ACG, ..., CAG, ..., CGT, ..., GAC, ..., GTT, ..., TGC, ..., TTG, ..., TTT
Building an Index

AAA, AAC, ..., ACG, ..., CAG, ..., CGT, ..., GAC, ..., GTT, ..., TGC, ..., TTG, ... TTT

0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8
A C G T T G C A G T T G A C T G A C G

AAA, AAC, ..., ACG, ..., CAG, ..., CGT, ..., GAC, ..., GTT, ..., TGC, ..., TTG, ... TTT

0 1 2 3 4 5 6 7 8 9

Bioinformatic and Comparative Genome Analysis course, Institut Pasteur, July 5 - July 17, 2010
Scanning a new sequence

AAA, AAC, ..., ACG, ..., CAG, ..., CGT, ..., GAC, ..., GTT, ..., TGC, ..., TTG, ..., TTT

0 \quad 1 \quad 2 \quad 4 \quad 3 \quad 8 \quad 9

Second sequence:

\begin{array}{c}
\text{T} \\
\text{G} \\
\text{C} \\
\text{A} \\
\text{G} \\
\text{T} \\
\text{T} \\
\text{T} \\
\text{G} \\
\end{array}
Scanning a new sequence

AAA, AAC, ..., ACG, ..., CAG, ..., CGT, ..., GAC, ..., GTT, ..., TGC, ..., TTG, ..., TTT

0 1 2 4 3 8 9

Second sequence:

T G C A G T T G . . .
Scanning a new sequence

AAA, AAC, ..., ACG, ..., **CAG**, ..., CGT, ... GAC, ..., GTT, ..., TGC, ... TTG, ... TTT

0 1 2 4 3 8 9

Second sequence:

T G C A G T T G ...
Scanning a new sequence

AAA, AAC, ACG, CAG, CGT, GAC, GTT, TGC, TTG, TTT

0 1 2 4 3
8 9

Second sequence:

T G C A G T T G . . .
Scanning a new sequence

AAA, AAC, ..., ACG, ..., CAG, ..., CGT, ... GAC, ..., **GTT**, ..., TGC, ... TTG, ... TTT

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>4</th>
<th>3</th>
</tr>
</thead>
</table>

Second sequence:

T G C A **G T T** G . . .
Interaction between the similarity model and the index

Once a model is chosen and the index is built, two questions arise:

1. Is it possible to find a match using the window size chosen?
2. How many character matches are likely to be detected with the window size chosen?
Q1: Is it possible to find a match?

This is known as the waiting time problem.

Waiting Time: How many consecutive positions must be examined until a run of k ones occurs.
Q1: Is it possible to find a match?

This is known as the waiting time problem.

Waiting Time: How many consecutive positions must be examined until a run of k ones occurs.

Specific sequence example:

Sequence 1:  AAAGCTTCGGAGTGCCCAGA
Sequence 2:  AATGCATCGGGGTGCCTGA

1101101111011111011
Waiting Time Specific Example

Sequence 1:  AAAGCTTTCGGAGTGCCCAGA
Sequence 2:  AATGCAATCGGGGTGCTTGA

1101101111011111011
↑↑↑↑↑↑↑↑↑

<table>
<thead>
<tr>
<th>k</th>
<th>waiting time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>16</td>
</tr>
<tr>
<td>6</td>
<td>-</td>
</tr>
</tbody>
</table>
Q1: Is it possible to find a match?

Waiting Time: Given a Bernoulli sequence with generating probability \( p \) and length \( n \), what is the probability that a run of \( k \) ones occurs?

Randomly generated Bernoulli sequence using \( p \):

\[
1110101111011011010
\]
Waiting Time Formulas

These calculate the probability of a first occurrence of a run of k ones at every sequence length from 1 to n.

for n ≥ 3, k = 3

\[
F(111:n) = 
\begin{align*}
& P(1)^3 - F(111: n - 1) \cdot P(1) - F(111: n - 2) \cdot P(1)^2 \\
& - \sum_{k=3}^{n-3} F(111: k) \cdot P(1)^3
\end{align*}
\]

where:

F(111:n) is the probability of a first occurrence of 3 ones in a row at position n,

P(1) is the model probability of a match.
Waiting Time Formulas

Predictions: If $k = 3$, $p = .5$, $n = 12$

1. In what position $[0..12]$ is it most likely to get a first occurrence of 3 ones in a row?
2. By what position will there be a cumulative probability of 30% to see a first occurrence of 3 ones in a row?
3. What is the likely cumulative probability of getting 3 ones in a row anywhere up to position 12?
## Waiting Time Formulas

### Calculated probabilities:

**Probabilities of first occurrence of patterns in coin toss sequences**

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Position</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>HHH</td>
<td>1</td>
<td>0.000000</td>
</tr>
<tr>
<td>P(111)</td>
<td>2</td>
<td>0.000000</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.125000</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.062500</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.062500</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>0.062500</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>0.054688</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>0.050781</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>0.046875</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.042969</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>0.039551</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>0.036377</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Cumulative</th>
</tr>
</thead>
<tbody>
<tr>
<td>HHH</td>
<td>0.000000</td>
</tr>
<tr>
<td>P(111)</td>
<td>0.000000</td>
</tr>
<tr>
<td></td>
<td>0.125000</td>
</tr>
<tr>
<td></td>
<td>0.187500</td>
</tr>
<tr>
<td></td>
<td>0.250000</td>
</tr>
<tr>
<td></td>
<td>0.312500</td>
</tr>
<tr>
<td></td>
<td>0.367188</td>
</tr>
<tr>
<td></td>
<td>0.417969</td>
</tr>
<tr>
<td></td>
<td>0.464844</td>
</tr>
<tr>
<td></td>
<td>0.507813</td>
</tr>
<tr>
<td></td>
<td>0.547363</td>
</tr>
<tr>
<td></td>
<td>0.583740</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pattern</th>
<th>P(1)</th>
<th>P(0)</th>
<th>P(1)</th>
<th>P(1)^2</th>
<th>P(1)^3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.25</td>
<td>0.125</td>
</tr>
</tbody>
</table>

- P(HHH) = 0.5^2 = 0.25
- P(P(111)) = 0.5^3 = 0.125

\[
\begin{align*}
P(1) & = 0.5 \\
P(0) & = 0.5 \\
P(1)^2 & = 0.25 \\
P(1)^3 & = 0.125
\end{align*}
\]
Q2: How many character matches will be detected?

This is known as the coverage problem.

Coverage: Given a Bernoulli sequence with generating probability $p$ and length $n$, what is the probability distribution for number of ones contained in runs of $k$ or more ones?
Q2: How many character matches will be detected?

Specific sequence example: Let $k = 3$, $n = 19$

Sequence 1: AAAGC TTCCGA GTGCCCGA

Sequence 2: AATGC ACTCGGGGTGCTTGA

1101101111011111011

$k$ $n$
Q2: How many character matches will be detected?

Specific sequence example: Let $k = 3$, $n = 19$

Sequence 1: $\text{AAAGCTTCGGAGTGCCC} \text{CGA}$

Sequence 2: $\text{AAATGCAATCGGGGTGCTCT} \text{GA}$

Total character matches detected is 9.
Data Structure – modified Aho Corasick Tree

Seed is 1*1**1.

Tree represents all patterns obtained by replacing each * by either 0 or 1.

Fail links in AC tree go to longest match between a string suffix and a prefix of a pattern.
Recurrence Formula

\[
\forall i \in [0, n], j \in [0, i], b_x \in B_x \\
X[n, i, j, b_x] = \sum_{b_y \in BF_y(b_x)} Y[n - dd_{xy}, i - od_{xy}, j - cv(b_y), b_y] \\
+ \sum_{Z_l, l \in \{1, 2, \ldots\}} \left( \sum_{b_{zl} \in BD_{zl}(b_x)} Z_l[n, i, j, b_{zl}] \right).
\]
Q2: How many character matches will be detected?
Basic Assumption

We assume that two, mutated, adjacent copies of a pattern will contain runs of exact matches.

T A T A C G T C

T C C A C G G A

\( d \)
Basic Assumption

We assume that two, mutated, adjacent copies of a pattern will contain runs of exact matches.

We identify the runs with seeds.
Basic Assumption

We assume that two, mutated, adjacent copies of a pattern will contain runs of exact matches.
The TRF Algorithm Outline

\[ k \text{ length window scan} \]

\[ i \]

\[ \text{ACGCT} \]

Window contents

Probes: AAAAA \ldots ACGCT \ldots TTTTT

\[ i \]

History List:

Indices of \[ k \]-tuple occurrence

\[ i - j_1 = d_1 \]

Possible tandem repeats:

\[ d_1 \]

\[ d_1 \]

Test Criteria

\[ j_1 \]

\[ j_1 \]

\[ D_1 \]

\[ i \]

\[ i - j_m = d_m \]

\[ d_m \]

\[ d_m \]

Test Criteria

\[ j_m \]

\[ j_m \]

\[ D_{j_m} \]

\[ i \]
Criteria for Recognition

- Are there enough matches at a common distance?
- Are there enough matches if nearby distances are included (random walk)?
- Do the matches start close enough to the left end?
Part 2
The Tandem Repeats Database
Log in using the [guest account]

Not yet a user? [register here]

Forgot your password? [password reminder]
Selecting a Data Set

This page lets you view sets. Select a project, then select a set. (help)
### Viewing a Data Set

This page lets you view sets. Select a project, then select a set. (help)

<table>
<thead>
<tr>
<th>Projects</th>
<th>Sets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Public Database</td>
<td>Human Chr. 1</td>
</tr>
<tr>
<td>(ca)n sites</td>
<td>Human Chr. 2</td>
</tr>
<tr>
<td>43bp repeats C. elegans</td>
<td>Human Chr. 3</td>
</tr>
<tr>
<td>Bams alleles</td>
<td>Human Chr. 4</td>
</tr>
<tr>
<td>BENSON_NEW_CODE</td>
<td>Human Chr. 5</td>
</tr>
<tr>
<td>Bug2003 fix Project</td>
<td>Human Chr. 6</td>
</tr>
<tr>
<td>default</td>
<td>Human Chr. 7</td>
</tr>
<tr>
<td>default</td>
<td>Human Chr. 8</td>
</tr>
<tr>
<td>family_search</td>
<td>Human Chr. 9</td>
</tr>
<tr>
<td>fractional copies</td>
<td>Human Chr. 10</td>
</tr>
</tbody>
</table>

**Multiple Sequence Summary**

- **name:** Human Chr. 1
- **description:** processed using LBI DAS batch system
- **status:** Done
- **created by:** Administrator
- **created on:** 2005-06-28
- **number of repeats:** 71536
- **project:** Public Database
### TRF Characteristics Table

<table>
<thead>
<tr>
<th>Indices</th>
<th>Pattern Size</th>
<th>Copy Number</th>
<th>% Matches</th>
<th>% Mismatches</th>
<th>% Indels</th>
<th>%A</th>
<th>%C</th>
<th>%G</th>
<th>%T</th>
<th>Score</th>
<th>Fasta Header</th>
</tr>
</thead>
<tbody>
<tr>
<td>79911--80293</td>
<td>59</td>
<td>6.500000</td>
<td>98</td>
<td>2</td>
<td>0</td>
<td>18</td>
<td>34</td>
<td>21</td>
<td>25</td>
<td>712</td>
<td>chr1</td>
</tr>
<tr>
<td>126063--127151</td>
<td>49</td>
<td>22.400000</td>
<td>94</td>
<td>6</td>
<td>0</td>
<td>14</td>
<td>25</td>
<td>43</td>
<td>16</td>
<td>1683</td>
<td>chr1</td>
</tr>
<tr>
<td>126063--127151</td>
<td>195</td>
<td>5.600000</td>
<td>96</td>
<td>4</td>
<td>0</td>
<td>14</td>
<td>25</td>
<td>43</td>
<td>16</td>
<td>1880</td>
<td>chr1</td>
</tr>
<tr>
<td>366710--367017</td>
<td>49</td>
<td>6.400000</td>
<td>93</td>
<td>5</td>
<td>2</td>
<td>15</td>
<td>41</td>
<td>26</td>
<td>15</td>
<td>439</td>
<td>chr1</td>
</tr>
<tr>
<td>382057--382215</td>
<td>25</td>
<td>6.400000</td>
<td>93</td>
<td>3</td>
<td>4</td>
<td>52</td>
<td>1</td>
<td>3</td>
<td>42</td>
<td>229</td>
<td>chr1</td>
</tr>
<tr>
<td>424228--426617</td>
<td>77</td>
<td>31.100000</td>
<td>99</td>
<td>1</td>
<td>0</td>
<td>9</td>
<td>35</td>
<td>23</td>
<td>32</td>
<td>4622</td>
<td>chr1</td>
</tr>
<tr>
<td>477501--478045</td>
<td>58</td>
<td>8.000000</td>
<td>98</td>
<td>2</td>
<td>0</td>
<td>11</td>
<td>43</td>
<td>30</td>
<td>14</td>
<td>1018</td>
<td>chr1</td>
</tr>
<tr>
<td>486805--487531</td>
<td>79</td>
<td>9.200000</td>
<td>96</td>
<td>4</td>
<td>0</td>
<td>31</td>
<td>17</td>
<td>23</td>
<td>26</td>
<td>1247</td>
<td>chr1</td>
</tr>
<tr>
<td>489137--490452</td>
<td>47</td>
<td>28.000000</td>
<td>97</td>
<td>3</td>
<td>0</td>
<td>17</td>
<td>29</td>
<td>31</td>
<td>21</td>
<td>2380</td>
<td>chr1</td>
</tr>
<tr>
<td>491680--491982</td>
<td>53</td>
<td>5.600000</td>
<td>94</td>
<td>7</td>
<td>9</td>
<td>23</td>
<td>6</td>
<td>35</td>
<td>33</td>
<td>187</td>
<td>chr1</td>
</tr>
<tr>
<td>573038--573977</td>
<td>47</td>
<td>20.000000</td>
<td>99</td>
<td>1</td>
<td>0</td>
<td>19</td>
<td>31</td>
<td>31</td>
<td>17</td>
<td>1835</td>
<td>chr1</td>
</tr>
<tr>
<td>575586--576075</td>
<td>79</td>
<td>6.200000</td>
<td>96</td>
<td>4</td>
<td>0</td>
<td>26</td>
<td>23</td>
<td>17</td>
<td>32</td>
<td>818</td>
<td>chr1</td>
</tr>
</tbody>
</table>
Filter for large patterns

Tandem Repeats Database

Filtering Options

- **Pattern Size**: > = 25.000000
- **Copy Number**: > = 10.000000

# repeats (after filters): **815**

Goto page 1 out of 9
### More information about a single repeat

<table>
<thead>
<tr>
<th>Indices</th>
<th>Pattern Size</th>
<th>Copy Number</th>
<th>% Matches</th>
<th>% Mismatches</th>
<th>% Indels</th>
<th>%A</th>
<th>%C</th>
<th>%G</th>
<th>%T</th>
<th>Score</th>
<th>Fasta Header</th>
</tr>
</thead>
<tbody>
<tr>
<td>79911--80293</td>
<td>59</td>
<td>6.500000</td>
<td>98</td>
<td>2</td>
<td>0</td>
<td>18</td>
<td>34</td>
<td>21</td>
<td>25</td>
<td>712</td>
<td>chr1</td>
</tr>
<tr>
<td>126063--127151</td>
<td>49</td>
<td>22.400000</td>
<td>94</td>
<td>6</td>
<td>0</td>
<td>14</td>
<td>25</td>
<td>43</td>
<td>16</td>
<td>1603</td>
<td>chr1</td>
</tr>
<tr>
<td>126063--127151</td>
<td>195</td>
<td>5.600000</td>
<td>96</td>
<td>4</td>
<td>0</td>
<td>14</td>
<td>25</td>
<td>43</td>
<td>16</td>
<td>1880</td>
<td>chr1</td>
</tr>
<tr>
<td>366710--367017</td>
<td>49</td>
<td>6.400000</td>
<td>93</td>
<td>5</td>
<td>2</td>
<td>15</td>
<td>41</td>
<td>26</td>
<td>15</td>
<td>439</td>
<td>chr1</td>
</tr>
<tr>
<td>382057--382215</td>
<td>25</td>
<td>6.400000</td>
<td>93</td>
<td>3</td>
<td>4</td>
<td>52</td>
<td>1</td>
<td>3</td>
<td>42</td>
<td>229</td>
<td>chr1</td>
</tr>
<tr>
<td>424228--426617</td>
<td>77</td>
<td>31.100000</td>
<td>99</td>
<td>1</td>
<td>0</td>
<td>9</td>
<td>38</td>
<td>25</td>
<td>32</td>
<td>4622</td>
<td>chr1</td>
</tr>
<tr>
<td>477501--478045</td>
<td>68</td>
<td>8.000000</td>
<td>96</td>
<td>2</td>
<td>0</td>
<td>11</td>
<td>43</td>
<td>30</td>
<td>14</td>
<td>1018</td>
<td>chr1</td>
</tr>
<tr>
<td>486805--487531</td>
<td>79</td>
<td>9.200000</td>
<td>96</td>
<td>4</td>
<td>0</td>
<td>31</td>
<td>17</td>
<td>23</td>
<td>26</td>
<td>1247</td>
<td>chr1</td>
</tr>
<tr>
<td>489137--490452</td>
<td>47</td>
<td>29.000000</td>
<td>97</td>
<td>3</td>
<td>0</td>
<td>17</td>
<td>29</td>
<td>31</td>
<td>21</td>
<td>2380</td>
<td>chr1</td>
</tr>
<tr>
<td>491680--491982</td>
<td>53</td>
<td>5.600000</td>
<td>84</td>
<td>7</td>
<td>9</td>
<td>23</td>
<td>6</td>
<td>35</td>
<td>33</td>
<td>187</td>
<td>chr1</td>
</tr>
<tr>
<td>573036--573977</td>
<td>47</td>
<td>20.000000</td>
<td>99</td>
<td>1</td>
<td>0</td>
<td>19</td>
<td>31</td>
<td>31</td>
<td>17</td>
<td>1835</td>
<td>chr1</td>
</tr>
<tr>
<td>575586--576075</td>
<td>79</td>
<td>6.200000</td>
<td>96</td>
<td>4</td>
<td>0</td>
<td>28</td>
<td>23</td>
<td>17</td>
<td>32</td>
<td>818</td>
<td>chr1</td>
</tr>
</tbody>
</table>
Single repeat view

Parameters: 27780 1050 2000
Organism: Homo sapiens (May 2004)
sequence: Human Chr. 1 (more info)
FASTA header: chr1

<table>
<thead>
<tr>
<th>Indices</th>
<th>Period Size</th>
<th>Copy Number</th>
<th>%Matches</th>
<th>%Indels</th>
<th>Gene</th>
<th>Exon</th>
<th>Intron</th>
<th>RepeatID</th>
<th>Fasta Header</th>
</tr>
</thead>
<tbody>
<tr>
<td>125053--127151</td>
<td>49</td>
<td>22400000</td>
<td>94</td>
<td>0</td>
<td>&lt;No&gt;</td>
<td>&lt;No&gt;</td>
<td>&lt;No&gt;</td>
<td>17141297</td>
<td>chr1</td>
</tr>
</tbody>
</table>
Filter for Gene Overlap or Proximity

<table>
<thead>
<tr>
<th>Filtering Options</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>First Index</td>
<td>&gt; 10000000000.000000</td>
</tr>
<tr>
<td>First Index</td>
<td>&lt;= 2000000000.000000</td>
</tr>
<tr>
<td>Copy Number</td>
<td>&gt; 15000000</td>
</tr>
<tr>
<td>Pattern Size</td>
<td>= 3000000</td>
</tr>
</tbody>
</table>

Gene: search for overlap or proximity

# repeats (after filters): 117
## Filters for Triplets in Genes

### Filtering Options

<table>
<thead>
<tr>
<th>Field</th>
<th>Operator</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>First Index</td>
<td></td>
<td>&gt; 1000000000.000000</td>
</tr>
<tr>
<td>First Index</td>
<td></td>
<td>&lt; 200000000.000000</td>
</tr>
<tr>
<td>Copy Number</td>
<td></td>
<td>&gt; 15.000000</td>
</tr>
<tr>
<td>Pattern Size</td>
<td></td>
<td>= 3.000000</td>
</tr>
<tr>
<td>Gene</td>
<td></td>
<td>overlaps</td>
</tr>
</tbody>
</table>

# repeats (after filters): **39**  
[(help)]
## Triplets in Genes

<table>
<thead>
<tr>
<th>Indices</th>
<th>Pattern Size</th>
<th>Copy Number</th>
<th>% Matches</th>
<th>% Indels</th>
<th>Gene</th>
<th>Exon</th>
<th>Intron</th>
<th>Repeatid</th>
<th>Fasta Header</th>
</tr>
</thead>
<tbody>
<tr>
<td>100345356--100345403</td>
<td>3</td>
<td>16.000000</td>
<td>100</td>
<td>0</td>
<td>Yes</td>
<td>&lt;&lt;No&gt;&gt;</td>
<td>Yes</td>
<td>17174963</td>
<td>chr1</td>
</tr>
<tr>
<td>108139065--108139116</td>
<td>3</td>
<td>17.000000</td>
<td>98</td>
<td>1</td>
<td>Yes</td>
<td>&lt;&lt;No&gt;&gt;</td>
<td>Yes</td>
<td>17177132</td>
<td>chr1</td>
</tr>
<tr>
<td>111874531--111874583</td>
<td>3</td>
<td>17.299999</td>
<td>92</td>
<td>1</td>
<td>Yes</td>
<td>Yes</td>
<td>&lt;&lt;No&gt;&gt;</td>
<td>17178151</td>
<td>chr1</td>
</tr>
<tr>
<td>112093326--112093375</td>
<td>3</td>
<td>16.700001</td>
<td>90</td>
<td>3</td>
<td>Yes</td>
<td>&lt;&lt;No&gt;&gt;</td>
<td>Yes</td>
<td>17178205</td>
<td>chr1</td>
</tr>
<tr>
<td>113692381--113692627</td>
<td>3</td>
<td>81.000000</td>
<td>88</td>
<td>4</td>
<td>Yes</td>
<td>&lt;&lt;No&gt;&gt;</td>
<td>Yes</td>
<td>17178742</td>
<td>chr1</td>
</tr>
<tr>
<td>115955908--115955908</td>
<td>3</td>
<td>25.700001</td>
<td>85</td>
<td>2</td>
<td>Yes</td>
<td>Yes</td>
<td>&lt;&lt;No&gt;&gt;</td>
<td>17179336</td>
<td>chr1</td>
</tr>
<tr>
<td>116016066--116016193</td>
<td>3</td>
<td>42.299999</td>
<td>96</td>
<td>0</td>
<td>Yes</td>
<td>&lt;&lt;No&gt;&gt;</td>
<td>Yes</td>
<td>17179349</td>
<td>chr1</td>
</tr>
<tr>
<td>118134131--118134177</td>
<td>3</td>
<td>15.700000</td>
<td>91</td>
<td>0</td>
<td>Yes</td>
<td>&lt;&lt;No&gt;&gt;</td>
<td>Yes</td>
<td>17179922</td>
<td>chr1</td>
</tr>
<tr>
<td>119673283--119673332</td>
<td>3</td>
<td>16.700001</td>
<td>98</td>
<td>0</td>
<td>Yes</td>
<td>&lt;&lt;No&gt;&gt;</td>
<td>Yes</td>
<td>17180371</td>
<td>chr1</td>
</tr>
</tbody>
</table>
Changing Visible Columns

### Filtering Options

<table>
<thead>
<tr>
<th>Pattern Size</th>
<th>Copy Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;= 25.000000</td>
<td>&gt;= 10.000000</td>
</tr>
</tbody>
</table>

### Table

<table>
<thead>
<tr>
<th>Indices</th>
<th>Pattern Size</th>
<th>Copy Number</th>
<th>%Matches</th>
<th>%Mismatches</th>
<th>%Indels</th>
<th>%A</th>
<th>%C</th>
<th>%G</th>
<th>%T</th>
</tr>
</thead>
<tbody>
<tr>
<td>3242935--3243679</td>
<td>42</td>
<td>17.799999</td>
<td>85</td>
<td>8</td>
<td>7</td>
<td>18</td>
<td>14</td>
<td>31</td>
<td>34</td>
</tr>
<tr>
<td>3268133--3268469</td>
<td>33</td>
<td>10.000000</td>
<td>92</td>
<td>6</td>
<td>2</td>
<td>21</td>
<td>16</td>
<td>49</td>
<td>13</td>
</tr>
</tbody>
</table>

# repeats (after filters): **815** (help)
Changing Visible Columns

This page allows you to select which data columns you want to be present when viewing a set of repeats. Warning: selecting a large number of columns will increase the amount of data sent to you per page and therefore slow down page loads. (help)
<table>
<thead>
<tr>
<th>V</th>
<th>Indices</th>
<th>Pattern Size</th>
<th>Copy Number</th>
<th>% Matches</th>
<th>% Indels</th>
<th>Gene</th>
<th>Exon</th>
<th>Intron</th>
<th>Repeatid</th>
<th>Fasta Header</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td>100345356--100345403</td>
<td>3</td>
<td>16.000000</td>
<td>100</td>
<td>0</td>
<td>Yes</td>
<td>&lt;&lt;No&gt;&gt;</td>
<td>Yes</td>
<td>17174963</td>
<td>chr1</td>
</tr>
<tr>
<td>✗</td>
<td>108139065--108139116</td>
<td>3</td>
<td>17.000000</td>
<td>98</td>
<td>1</td>
<td>Yes</td>
<td>&lt;&lt;No&gt;&gt;</td>
<td>Yes</td>
<td>17177132</td>
<td>chr1</td>
</tr>
<tr>
<td>✓</td>
<td>111874531--111874583</td>
<td>3</td>
<td>17.299999</td>
<td>92</td>
<td>1</td>
<td>Yes</td>
<td>Yes</td>
<td>&lt;&lt;No&gt;&gt;</td>
<td>17178151</td>
<td>chr1</td>
</tr>
<tr>
<td>✓</td>
<td>112093326--112093375</td>
<td>3</td>
<td>16.700000</td>
<td>90</td>
<td>3</td>
<td>Yes</td>
<td>&lt;&lt;No&gt;&gt;</td>
<td>Yes</td>
<td>17178205</td>
<td>chr1</td>
</tr>
<tr>
<td>✓</td>
<td>113692381--113692627</td>
<td>3</td>
<td>16.700000</td>
<td>88</td>
<td>4</td>
<td>Yes</td>
<td>&lt;&lt;No&gt;&gt;</td>
<td>Yes</td>
<td>17178742</td>
<td>chr1</td>
</tr>
<tr>
<td>✓</td>
<td>115955908--115955964</td>
<td>3</td>
<td>25.700001</td>
<td>85</td>
<td>2</td>
<td>Yes</td>
<td>Yes</td>
<td>&lt;&lt;No&gt;&gt;</td>
<td>17179336</td>
<td>chr1</td>
</tr>
<tr>
<td>✓</td>
<td>116016066--116016193</td>
<td>3</td>
<td>42.299999</td>
<td>96</td>
<td>0</td>
<td>Yes</td>
<td>&lt;&lt;No&gt;&gt;</td>
<td>Yes</td>
<td>17179349</td>
<td>chr1</td>
</tr>
<tr>
<td>✓</td>
<td>118134131--118134177</td>
<td>3</td>
<td>15.700000</td>
<td>91</td>
<td>0</td>
<td>Yes</td>
<td>&lt;&lt;No&gt;&gt;</td>
<td>Yes</td>
<td>17179922</td>
<td>chr1</td>
</tr>
<tr>
<td>✓</td>
<td>119673283--119673332</td>
<td>3</td>
<td>16.700001</td>
<td>98</td>
<td>0</td>
<td>Yes</td>
<td>&lt;&lt;No&gt;&gt;</td>
<td>Yes</td>
<td>17180371</td>
<td>chr1</td>
</tr>
</tbody>
</table>
## Annotations

### Annotation (Gene)

<table>
<thead>
<tr>
<th>Indices</th>
<th>Period Size</th>
<th>Copy Number</th>
<th>%Matches</th>
<th>%Indels</th>
<th>Gene</th>
<th>Exon</th>
<th>Intron</th>
<th>Repeatid</th>
<th>Fasta Header</th>
</tr>
</thead>
<tbody>
<tr>
<td>111874447--111874593</td>
<td>3</td>
<td>17.299999</td>
<td>92</td>
<td>1</td>
<td>Yes</td>
<td>Yes</td>
<td>&lt;&lt;No&gt;&gt;</td>
<td>17170151</td>
<td>chr1</td>
</tr>
</tbody>
</table>

The repeat overlaps the following Gene(s):

<table>
<thead>
<tr>
<th>Sequence Name</th>
<th>Source</th>
<th>Feature</th>
<th>Start</th>
<th>End</th>
<th>Score</th>
<th>Strand</th>
<th>Frame</th>
<th>Attribute</th>
<th>Comment/Link</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>UCSC</td>
<td>gene</td>
<td>111874447</td>
<td>111968142</td>
<td>0.00</td>
<td>+</td>
<td>0</td>
<td>NM_001010935.chr1.111874446</td>
<td>More Information</td>
</tr>
<tr>
<td>1</td>
<td>UCSC</td>
<td>gene</td>
<td>111874447</td>
<td>111968142</td>
<td>0.00</td>
<td>+</td>
<td>0</td>
<td>NM_002984.chr1.111874446</td>
<td>More Information</td>
</tr>
</tbody>
</table>

The repeat also overlaps the following intergenic features:

<table>
<thead>
<tr>
<th>Sequence Name</th>
<th>Source</th>
<th>Feature</th>
<th>Start</th>
<th>End</th>
<th>Score</th>
<th>Strand</th>
<th>Frame</th>
<th>Attribute</th>
<th>Comment/Link</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>UCSC</td>
<td>exon</td>
<td>111874447</td>
<td>111874598</td>
<td>0.00</td>
<td>+</td>
<td>0</td>
<td>NM_001010935.chr1.111874446</td>
<td>More Information</td>
</tr>
<tr>
<td>1</td>
<td>UCSC</td>
<td>exon</td>
<td>111874447</td>
<td>111874598</td>
<td>0.00</td>
<td>+</td>
<td>0</td>
<td>NM_002884.chr1.111874446</td>
<td>More Information</td>
</tr>
</tbody>
</table>
Information link to the Source Database

**Annotation (Gene)**

<table>
<thead>
<tr>
<th>Indices</th>
<th>Period Size</th>
<th>Copy Number</th>
<th>%Matches</th>
<th>%Indels</th>
<th>Gene</th>
<th>Exon</th>
<th>Intron</th>
<th>RepeatID</th>
<th>Fasta Header</th>
</tr>
</thead>
<tbody>
<tr>
<td>111874531-111874583</td>
<td>3</td>
<td>17.299999</td>
<td>92</td>
<td>1</td>
<td>Yes</td>
<td>Yes</td>
<td>&lt;&lt;No&gt;&gt;</td>
<td>17178151</td>
<td>chr1</td>
</tr>
</tbody>
</table>

The repeat overlaps the following **Gene(s):**

<table>
<thead>
<tr>
<th>Sequence Name</th>
<th>Source</th>
<th>Feature</th>
<th>Start</th>
<th>End</th>
<th>Score</th>
<th>Strand</th>
<th>Frame</th>
<th>Attribute</th>
<th>Comment/Link</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>UCSC</td>
<td>gene</td>
<td>111874447</td>
<td>111968142</td>
<td>0.00</td>
<td>+</td>
<td>0</td>
<td>NM_001010935.chr1.111874445</td>
<td>More Information</td>
</tr>
<tr>
<td>1</td>
<td>UCSC</td>
<td>gene</td>
<td>111874447</td>
<td>111968142</td>
<td>0.00</td>
<td>+</td>
<td>0</td>
<td>NM_002884.chr1.111874446</td>
<td>More Information</td>
</tr>
</tbody>
</table>

The repeat also overlaps the following **intergenic** features:

<table>
<thead>
<tr>
<th>Sequence Name</th>
<th>Source</th>
<th>Feature</th>
<th>Start</th>
<th>End</th>
<th>Score</th>
<th>Strand</th>
<th>Frame</th>
<th>Attribute</th>
<th>Comment/Link</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>UCSC</td>
<td>exon</td>
<td>111874447</td>
<td>111874598</td>
<td>0.00</td>
<td>+</td>
<td>0</td>
<td>NM_001010935.chr1.111874446</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>UCSC</td>
<td>exon</td>
<td>111874447</td>
<td>111874598</td>
<td>0.00</td>
<td>+</td>
<td>0</td>
<td>NM_002884.chr1.111874446</td>
<td></td>
</tr>
</tbody>
</table>
Following the Information link to the UCSC Browser

UCSC Genome Browser on Human May 2004 Assembly

move <<< << < > > > > > > zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x

position/search chr1:111,674,446-111,968,142 jump clear size 93,697 bp. configure

Bioinformatic and Comparative Genome Analysis course, Institut Pasteur, July 5 - July 17, 2010
## The TRDB Browser link

<table>
<thead>
<tr>
<th>Indices</th>
<th>Pattern Size</th>
<th>Copy Number</th>
<th>% Matches</th>
<th>% Indels</th>
<th>Gene</th>
<th>Exon</th>
<th>Intron</th>
<th>Repeatid</th>
<th>Fasta Header</th>
</tr>
</thead>
<tbody>
<tr>
<td>100345356-100345403</td>
<td>3</td>
<td>16.000000</td>
<td>100</td>
<td>0</td>
<td>Yes</td>
<td>&lt;&lt;No&gt;&gt;</td>
<td>Yes</td>
<td>17174963</td>
<td>chr1</td>
</tr>
<tr>
<td>108139065-108139116</td>
<td>3</td>
<td>17.000000</td>
<td>98</td>
<td>1</td>
<td>Yes</td>
<td>&lt;&lt;No&gt;&gt;</td>
<td>Yes</td>
<td>17177132</td>
<td>chr1</td>
</tr>
<tr>
<td>111874531-111874583</td>
<td>3</td>
<td>17.299999</td>
<td>92</td>
<td>1</td>
<td>Yes</td>
<td>Yes</td>
<td>&lt;&lt;No&gt;&gt;</td>
<td>17178151</td>
<td>chr1</td>
</tr>
<tr>
<td>112093526-112093375</td>
<td>3</td>
<td>16.700001</td>
<td>90</td>
<td>3</td>
<td>Yes</td>
<td>&lt;&lt;No&gt;&gt;</td>
<td>Yes</td>
<td>17178205</td>
<td>chr1</td>
</tr>
<tr>
<td>113692381-113692627</td>
<td>3</td>
<td>81.000000</td>
<td>88</td>
<td>4</td>
<td>Yes</td>
<td>&lt;&lt;No&gt;&gt;</td>
<td>Yes</td>
<td>17178742</td>
<td>chr1</td>
</tr>
<tr>
<td>115955908-115955984</td>
<td>3</td>
<td>25.700001</td>
<td>85</td>
<td>2</td>
<td>Yes</td>
<td>Yes</td>
<td>&lt;&lt;No&gt;&gt;</td>
<td>17179336</td>
<td>chr1</td>
</tr>
<tr>
<td>116016066-116016193</td>
<td>3</td>
<td>42.299999</td>
<td>96</td>
<td>0</td>
<td>Yes</td>
<td>&lt;&lt;No&gt;&gt;</td>
<td>Yes</td>
<td>17179349</td>
<td>chr1</td>
</tr>
<tr>
<td>118134131-118134177</td>
<td>3</td>
<td>15.700000</td>
<td>91</td>
<td>0</td>
<td>Yes</td>
<td>&lt;&lt;No&gt;&gt;</td>
<td>Yes</td>
<td>17179922</td>
<td>chr1</td>
</tr>
<tr>
<td>119673283-119673332</td>
<td>3</td>
<td>16.700001</td>
<td>98</td>
<td>0</td>
<td>Yes</td>
<td>&lt;&lt;No&gt;&gt;</td>
<td>Yes</td>
<td>17180371</td>
<td>chr1</td>
</tr>
</tbody>
</table>
The TRDB Browser

- Project name: Public Database
- Set name: Human Chr. 1
- Date: Wed Nov 23, 2005 (10:58:28 AM)

- Introns
- Genes
- Exons
- Repeats

Click on a feature for more info. Click on base position to center/zoom in around cursor.
Distributions for a Data Set

This page lets you view sets. Select a project, then select a set.

Projects

- Public Database
- (cag)n sites
- 43bp repeats C. elegans
- Bams alleles
- BENS0N_NEW_CODE
- Bug2003 fix Project
default
default
family_search
fractional copies

Sets

- Human Chr. 1
- Human Chr. 2
- Human Chr. 3
- Human Chr. 4
- Human Chr. 5
- Human Chr. 6
- Human Chr. 7
- Human Chr. 8
- Human Chr. 9
- Human Chr. 10

Multiple Sequence Summary

name: Human Chr. 1
description: processed using LBI DAS batch system
status: Done
created by: Administrator
created on: 2005-06-28
number of repeats: 71536
project: Public Database
Pattern size distribution
Human chr. 1: sizes 60 - 120
Pattern size distribution

*Drosophila* chr. 2R: size 1 - 60
Clustering repeats
Clustering repeats

Tandem Repeats Database

FAQs | Help | About TRDB | Privacy Statement | Log Out | Contact Us | News | IRDB | LBI

> partitions

This page lets you view partitions. Select a project, then select a partition. Partitions are created by running clustering algorithms on sets. You can access clustering through the TOOLS menu. (help)

Projects

- Public Database default

Partitions

- S. cerev. (Oct 2003)
Data Download

Tandem Repeats Database

FAQs | Help | About TRDB | Privacy Statement | Log Out | Contact Us | News | TRDB | LBI

> tools

- TRF Tandem Repeats Finding Program
- Flanking Sequence and Primers Extraction
- Clustering
- Transcription Factors Binding Sites Prediction
- mreps Alternative Tandem Repeats Finding Program
- Polymorphism Prediction
- Sub-Pattern Strength
- Set Comparison
- Search for Exact Pattern Within a Set
- Import Set
- Redundancy Elimination

Bioinformatic and Comparative Genome Analysis course, Institut Pasteur, July 5 - July 17, 2010
Data Download

This form will allow you to download data from [C. elegans Chr. I] set (info.)
Select the columns you want to be present and press DOWNLOAD. (help)

### Available Columns
- Repeatid
- First Index
- Last Index
- Pattern Size
- Copy Number
- %Matches

### Selected Columns

### Sorting Options
- **Order by:** First Index (ascending)
- **Group by Sequence:**
- **Destination:** Windows

### Output Options
- **ASCII**
  - delimiter
  - make first row the column heading
- **XML**
- **FASTA** (DNA fields only, field order is preset) (help)
  - short ids
  - reverse comp too
- **GFF** (these fields are preselected) (help)
  - Print UCSC Custom Track Header
- **LEB36**

[DOWNLOAD] [CANCEL]
Using TRF on your own data
Uploading a Sequence

> sequences > upload sequence

Your data must be a DNA sequence in\n\n**FASTA** format. (Multiple sequences in\na single file are permitted with a FASTA header before each sequence.)

(help)

**Sequence Source**

* Choose one of the following ways to send your data:

- Get the sequence from GENBANK.

Enter the GenBank accession number (put multiple numbers separated by commas if using the Automatic Genbank Sequence pull to get a multiple sequence FASTA file):

OR Select a file:

OR Cut and paste sequence (don’t forget the FASTA header):
Running TRF on a Sequence

This form will process the [bams7] sequence (info) with the TRF algorithm and create a set out of extracted repeats. Select the necessary parameters, enter the name and description of your resulting set and press PROCESS.

Parameters:

Alignment Parameters (match, mismatch, indels)

Minimum Alignment Score To Report Repeat

Set Information:

* Select which project you want your set to belong to. If the box below is empty, then you have not yet created/joined any projects.

* Enter the name of your set.

Provide a brief description of your set (500 chars maximum.)