Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences

Maja Tarailo-Graovac[1] **and Nansheng Chen**[1]

[1]Simon Fraser University, Burnaby, British Columbia, Canada

ABSTRACT

RepeatMasker is a popular software tool widely used in computational genomics to identify, classify, and mask repetitive elements, including low-complexity sequences and interspersed repeats. RepeatMasker searches for repetitive sequence by aligning the input genome sequence against a library of known repeats, such as Repbase. Here, we describe two Basic Protocols that provide detailed guidelines on how to use RepeatMasker, either via the Web interface or command-line Unix/Linux system, to analyze repetitive elements in genomic sequences. Sequence comparisons in RepeatMasker are usually performed by the alignment program cross_match, which requires significant processing time for larger sequences. An Alternate Protocol describes how to reduce the processing time using an alternative alignment program, such as WU-BLAST. Further, the advantages, limitations, and known bugs of the software are discussed. Finally, guidelines for understanding the results are provided. *Curr. Protoc. Bioinform.* 25:4.10.1-4.10.14. © 2009 by John Wiley & Sons, Inc.

Keywords: RepeatMasker • genome annotation • repetitive elements • repeat library • cross_match • WU-BLAST • RECON

INTRODUCTION

RepeatMasker (developed by A.F.A. Smit, R. Hubley, and P. Green; see *http://www.repeatmasker.org/*) was designed to identify and annotate repetitive elements in nucleotide sequences and mask them for further analysis. The repetitive elements, including low-complexity DNA sequences and interspersed repeats, are annotated and replaced by Ns, Xs, or lowercase letters (see below for options) in the corresponding positions of the DNA sequence. The new addition to the RepeatMasker package is a program that also identifies repetitive elements within protein sequences. Here, we focus on utilizing RepeatMasker to identify repetitive elements in genomic sequences. To run RepeatMasker, one needs to select the repeat library files, which contain repetitive elements consensus sequences. Currently, Repbase Update (Jurka, 2001; Jurka et al. 2005; *http://www.girinst.org/*) is the largest commercially available repeat library (free for academic use) and covers a number of organisms including human, rodent, zebrafish, *Drosophila*, and *Arabidopsis thaliana*. Library files for organisms that do not have Repbase Update library files can be generated ab initio using RECON (Bao and Eddy, 2002; *http://selab.janelia.org/recon.html*) or RepeatScout (*http://bix.ucsd.edu/repeatscout/*; Price et al., 2005). The newest version of RECON, v. 1.06, was released recently and is available from the RepeatModeler package at *http://www.repeatmasker.org/RepeatModeler.html*. Sequence comparisons in RepeatMasker are usually carried out by the program cross_match, developed by Phil Green (*http://www.phrap.org/consed/consed.html#howToGet*). One can also use WU-BLAST (*http://info.cchmc.org/help/wublast.html*; see Alternate Protocol) to replace cross_match for fast processing.

Current Protocols in Bioinformatics 4.10.1-4.10.14, March 2009
Published online March 2009 in Wiley Interscience (www.interscience.wiley.com).
DOI: 10.1002/0471250953.bi0410s25

BASIC PROTOCOL 1

USING RepeatMasker VIA THE WEB INTERFACE

RepeatMasker may be accessed through the Web at *http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker*. Unlike the command-line version of RepeatMasker (see Basic Protocol 2), Web RepeatMasker has a nucleotide sequence size limit of 100 kb. The attempt to analyze a sequence larger than 100 kb fails (whereupon a prompt is displayed in a message window, shown in Fig. 4.10.1). Sequences shorter than 100 kb are readily analyzed using the Web RepeatMasker, with the time needed for processing correlating with the length of the sequence. For faster service outside North America, there are RepeatMasker mirror sites in Germany, Israel, and Australia.

On the other hand, if one routinely submits large sequences for analysis, it may be better to download the command-line version and run RepeatMasker locally (see Basic Protocol 2). Importantly, if the query sequence exceeds the 100-kb limit, the only choice is to download RepeatMasker and run it locally.

Necessary Resources

Hardware

Any Internet-connected computer

Software

Web browser: e.g., Mozilla Firefox or Internet Explorer

Files

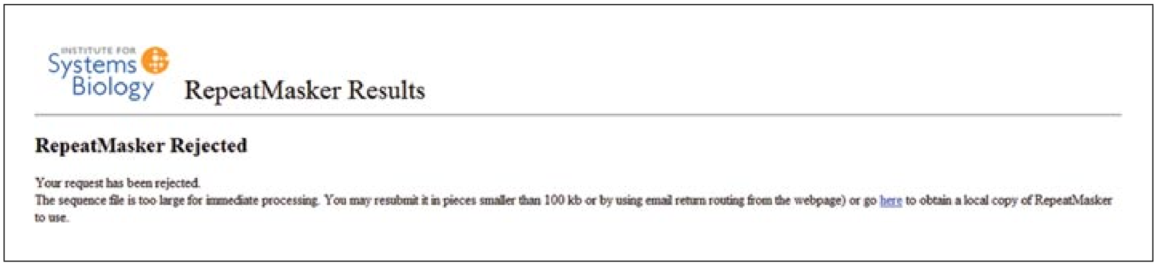
A FASTA file (*APPENDIX 1B*) or a collection of FASTA files can be processed via the Web interface. Note that the size limit is 100 kb for RepeatMasker via Web. The example file used in this protocol is a 22,539-bp human genomic DNA sequence from the UCSC Genome Browser (*http://genome.ucsc.edu/cgi-bin/hgGateway*). The coordinate is chr10:62743355-62765893.

1. Point the Web browser to *http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker*. Load the FASTA sequence file (maximum 100 kb) by entering the sequence name or browsing the file. Alternatively, paste the FASTA sequence (maximum 100 kb) into the indicated text field.

   RepeatMasker will return an error message if the input sequence contains non-DNA symbols or if the sequence is too long.

2. Select a format for results from the two radio buttons next to “return format”: “html” or “tar file.”

   If “html” is selected, the results will be written as an html file. If “tar file” is selected, the results will be packed into an archive using the Unix “tar” protocol. For the example here, select “html.”

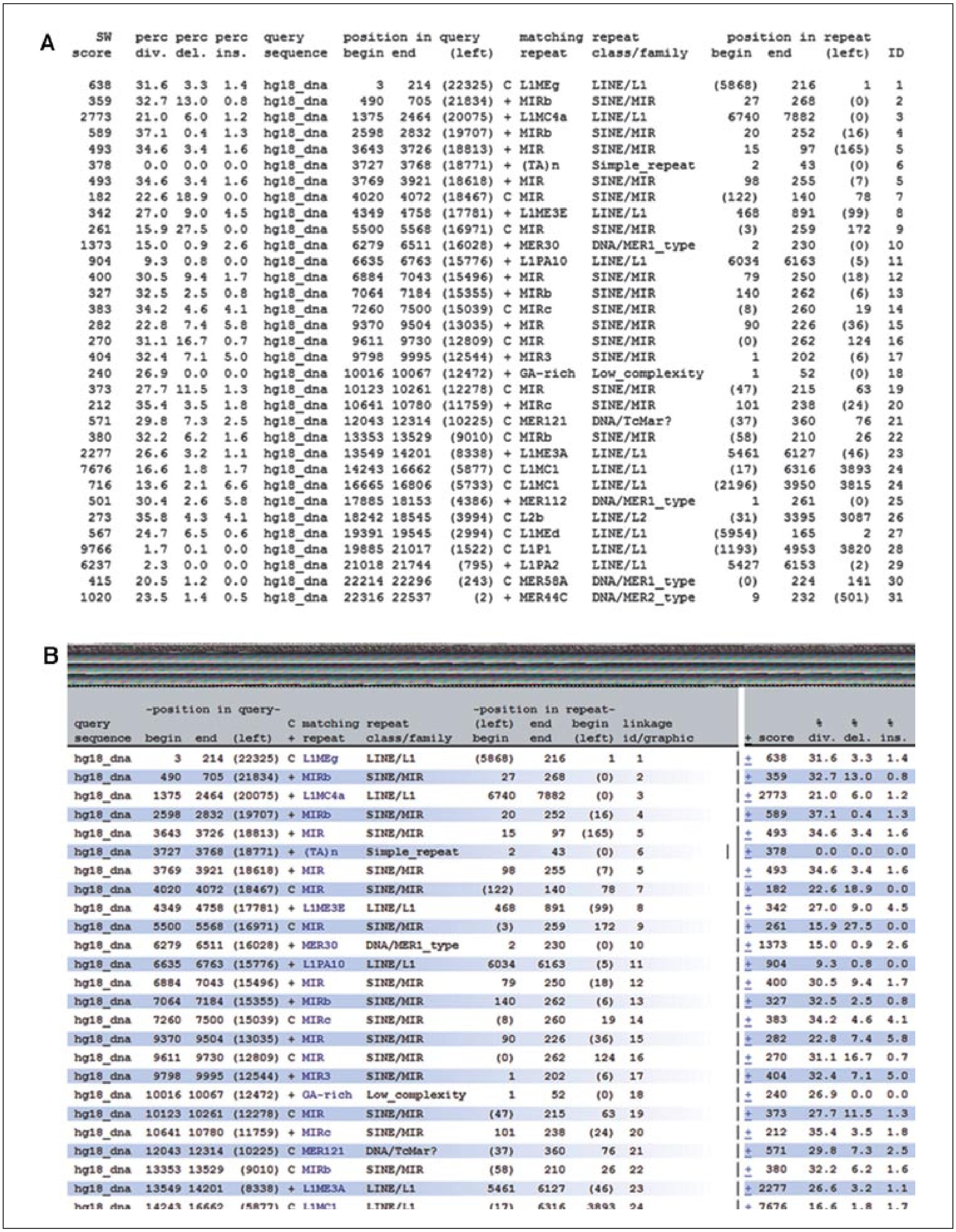


RepeatMasker Results

RepeatMasker Rejected

Your request has been rejected.
The sequence file is too large for immediate processing. You may resubmit it in pieces smaller than 100 kb or by using email return routing from the webpage) or go here to obtain a local copy of RepeatMasker to use.

Figure 4.10.1 Sequences with length >100 kb cannot be processed via the Web interface; user is informed by the RepeatMasker to consider alternate methods.



A

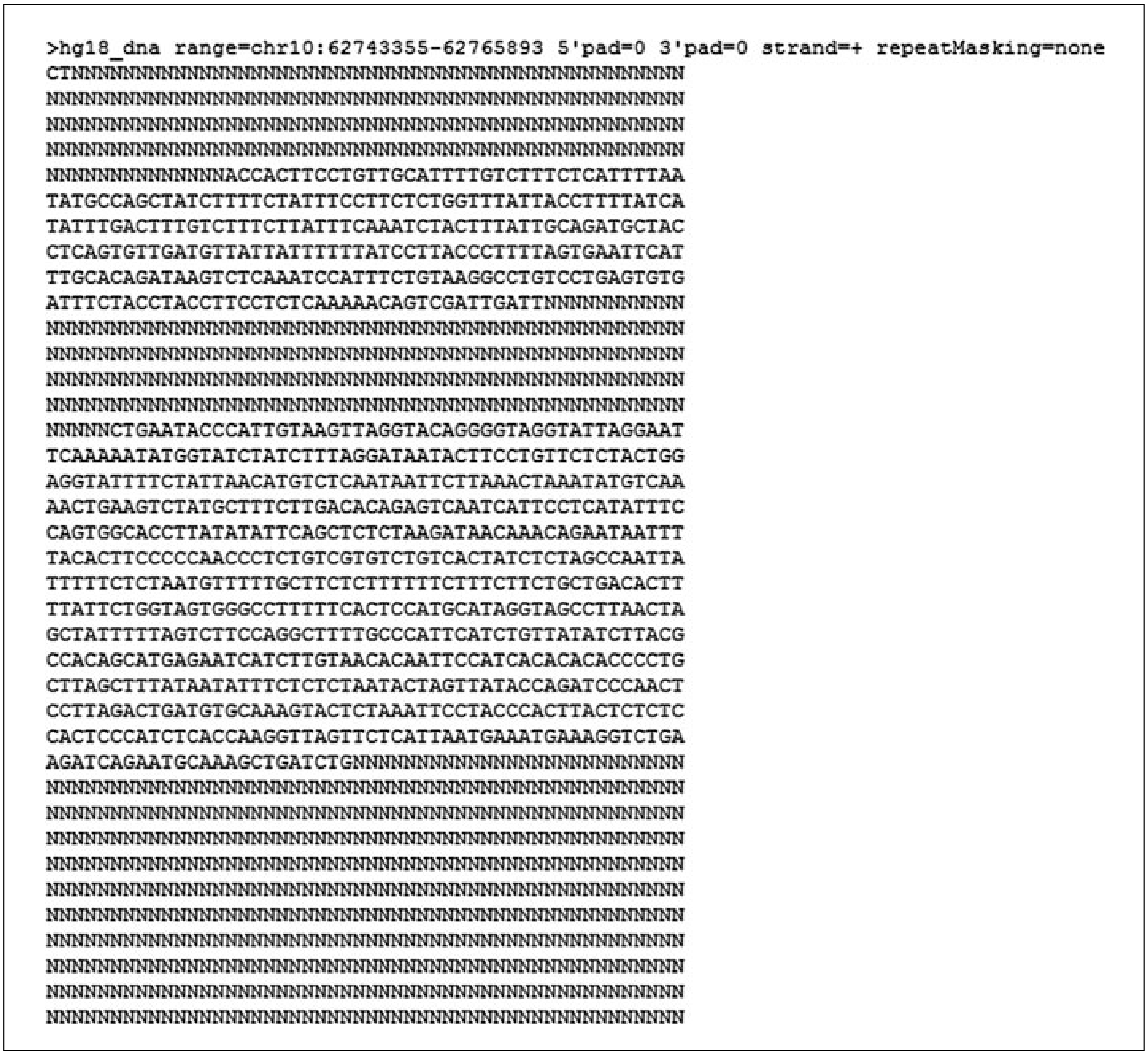
```
   SW   perc perc perc  query     position in query     matching repeat          position in repeat
score   div. del. ins.  sequence  begin end     (left)   repeat   class/family  begin  end     (left)  ID

  638   31.6  3.3  1.4  hg18_dna      3   214 (22325) C L1MEg     LINE/L1       (5868)    216       1   1
  359   32.7 13.0  0.8  hg18_dna    490   705 (21834) + MIRb      SINE/MIR          27    268     (0)   2
 2773   21.0  6.0  1.2  hg18_dna   1375  2464 (20075) + L1MC4a    LINE/L1         6740   7882     (0)   3
  589   37.1  0.4  1.3  hg18_dna   2598  2832 (19707) + MIRb      SINE/MIR          20    252    (16)   4
  493   34.6  3.4  1.6  hg18_dna   3643  3726 (18813) + MIR       SINE/MIR          15     97   (165)   5
  378    0.0  0.0  0.0  hg18_dna   3727  3768 (18771) + (TA)n     Simple_repeat      2     43     (0)   6
  493   34.6  3.4  1.6  hg18_dna   3769  3921 (18618) + MIR       SINE/MIR          98    255     (7)   5
  182   22.6 18.9  0.0  hg18_dna   4020  4072 (18467) C MIR       SINE/MIR        (122)    140      78   7
  342   27.0  9.0  4.5  hg18_dna   4349  4758 (17781) + L1ME3E    LINE/L1          468    891    (99)   8
  261   15.9 27.5  0.0  hg18_dna   5500  5568 (16971) C MIR       SINE/MIR          (3)    259     172   9
 1373   15.0  0.9  2.6  hg18_dna   6279  6511 (16028) + MER30     DNA/MER1_type      2    230     (0)  10
  904    9.3  0.8  0.0  hg18_dna   6635  6763 (15776) + L1PA10    LINE/L1         6034   6163     (5)  11
  400   30.5  9.4  1.7  hg18_dna   6884  7043 (15496) + MIR       SINE/MIR          79    250    (18)  12
  327   32.5  2.5  0.8  hg18_dna   7064  7184 (15355) + MIRb      SINE/MIR         140    262     (6)  13
  383   34.2  4.6  4.1  hg18_dna   7260  7500 (15039) C MIRc      SINE/MIR          (8)    260      19  14
  282   22.8  7.4  5.8  hg18_dna   9370  9504 (13035) + MIR       SINE/MIR          90    226    (36)  15
  270   31.1 16.7  0.7  hg18_dna   9611  9730 (12809) C MIR       SINE/MIR          (0)    262     124  16
  404   32.4  7.1  5.0  hg18_dna   9798  9995 (12544) + MIR3      SINE/MIR           1    202     (6)  17
  240   26.9  0.0  0.0  hg18_dna  10016 10067 (12472) + GA-rich   Low_complexity     1     52     (0)  18
  373   27.7 11.5  1.3  hg18_dna  10123 10261 (12278) C MIR       SINE/MIR         (47)    215      63  19
  212   35.4  3.5  1.8  hg18_dna  10641 10780 (11759) + MIRc      SINE/MIR         101    238    (24)  20
  571   29.8  7.3  2.5  hg18_dna  12043 12314 (10225) C MER121    DNA/TcMar?       (37)    360      76  21
  380   32.2  6.2  1.6  hg18_dna  13353 13529  (9010) C MIRb      SINE/MIR         (58)    210      26  22
 2277   26.6  3.2  1.1  hg18_dna  13549 14201  (8338) + L1ME3A    LINE/L1         5461   6127    (46)  23
 7676   16.6  1.8  1.7  hg18_dna  14243 16662  (5877) C L1MC1     LINE/L1         (17)   6316    3893  24
  716   13.6  2.1  6.6  hg18_dna  16665 16806  (5733) C L1MC1     LINE/L1       (2196)   3950    3815  24
  501   30.4  2.6  5.8  hg18_dna  17885 18153  (4386) + MER112    DNA/MER1_type      1    261     (0)  25
  273   35.8  4.3  4.1  hg18_dna  18242 18545  (3994) C L2b       LINE/L2          (31)   3395    3087  26
  567   24.7  6.5  0.6  hg18_dna  19391 19545  (2994) C L1MEd     LINE/L1       (5954)    165       2  27
 9766    1.7  0.1  0.0  hg18_dna  19885 21017  (1522) C L1P1      LINE/L1       (1193)   4953    3820  28
 6237    2.3  0.0  0.0  hg18_dna  21018 21744   (795) + L1PA2     LINE/L1         5427   6153     (2)  29
  415   20.5  1.2  0.0  hg18_dna  22214 22296   (243) C MER58A    DNA/MER1_type    (0)    224     141  30
 1020   23.5  1.4  0.5  hg18_dna  22316 22537     (2) + MER44C    DNA/MER2_type      9    232   (501)  31
```

B

query sequence	-position in query- begin	end	(left)	C +	matching repeat	repeat class/family	-position in repeat- (left) begin	end end	begin (left)	linkage id/graphic	+	score	% div.	% del.	% ins.
hg18_dna	3	214	(22325)	C	L1MEg	LINE/L1	(5868)	216	1	1	+	638	31.6	3.3	1.4
hg18_dna	490	705	(21834)	+	MIRb	SINE/MIR	27	268	(0)	2	+	359	32.7	13.0	0.8
hg18_dna	1375	2464	(20075)	+	L1MC4a	LINE/L1	6740	7882	(0)	3	+	2773	21.0	6.0	1.2
hg18_dna	2598	2832	(19707)	+	MIRb	SINE/MIR	20	252	(16)	4	+	589	37.1	0.4	1.3
hg18_dna	3643	3726	(18813)	+	MIR	SINE/MIR	15	97	(165)	5	+	493	34.6	3.4	1.6
hg18_dna	3727	3768	(18771)	+	(TA)n	Simple_repeat	2	43	(0)	6	+	378	0.0	0.0	0.0
hg18_dna	3769	3921	(18618)	+	MIR	SINE/MIR	98	255	(7)	5	+	493	34.6	3.4	1.6
hg18_dna	4020	4072	(18467)	C	MIR	SINE/MIR	(122)	140	78	7	+	182	22.6	18.9	0.0
hg18_dna	4349	4758	(17781)	+	L1ME3E	LINE/L1	468	891	(99)	8	+	342	27.0	9.0	4.5
hg18_dna	5500	5568	(16971)	C	MIR	SINE/MIR	(3)	259	172	9	+	261	15.9	27.5	0.0
hg18_dna	6279	6511	(16028)	+	MER30	DNA/MER1_type	2	230	(0)	10	+	1373	15.0	0.9	2.6
hg18_dna	6635	6763	(15776)	+	L1PA10	LINE/L1	6034	6163	(5)	11	+	904	9.3	0.8	0.0
hg18_dna	6884	7043	(15496)	+	MIR	SINE/MIR	79	250	(18)	12	+	400	30.5	9.4	1.7
hg18_dna	7064	7184	(15355)	+	MIRb	SINE/MIR	140	262	(6)	13	+	327	32.5	2.5	0.8
hg18_dna	7260	7500	(15039)	C	MIRc	SINE/MIR	(8)	260	19	14	+	383	34.2	4.6	4.1
hg18_dna	9370	9504	(13035)	+	MIR	SINE/MIR	90	226	(36)	15	+	282	22.8	7.4	5.8
hg18_dna	9611	9730	(12809)	C	MIR	SINE/MIR	(0)	262	124	16	+	270	31.1	16.7	0.7
hg18_dna	9798	9995	(12544)	+	MIR3	SINE/MIR	1	202	(6)	17	+	404	32.4	7.1	5.0
hg18_dna	10016	10067	(12472)	+	GA-rich	Low_complexity	1	52	(0)	18	+	240	26.9	0.0	0.0
hg18_dna	10123	10261	(12278)	C	MIR	SINE/MIR	(47)	215	63	19	+	373	27.7	11.5	1.3
hg18_dna	10641	10780	(11759)	+	MIRc	SINE/MIR	101	238	(24)	20	+	212	35.4	3.5	1.8
hg18_dna	12043	12314	(10225)	C	MER121	DNA/TcMar?	(37)	360	76	21	+	571	29.8	7.3	2.5
hg18_dna	13353	13529	(9010)	C	MIRb	SINE/MIR	(58)	210	26	22	+	380	32.2	6.2	1.6
hg18_dna	13549	14201	(8338)	+	L1ME3A	LINE/L1	5461	6127	(46)	23	+	2277	26.6	3.2	1.1
hg18_dna	14243	16662	(5877)	C	L1MC1	LINE/L1	(17)	6316	3893	24	+	7676	16.6	1.8	1.7

Figure 4.10.2 Web RepeatMasker result from an example run showing the repetitive elements annotations section, which lists cross_match summary lines; this result is available in Text File Format (**A**) and XHTML format (**B**). See Guidelines for Understanding Results and Table 4.10.1 for explanation.



```
>hg18_dna range=chr10:62743355-62765893 5'pad=0 3'pad=0 strand=+ repeatMasking=none
CTNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNACCACTTCCTGTTGCATTTTGTCTTTCTCATTTTAA
TATGCCAGCTATCTTTTCTATTTCCTTCTCTGGTTTATTACCTTTTATCA
TATTTGACTTTGTCTTTCTTATTTCAAATCTACTTTATTGCAGATGCTAC
CTCAGTGTTGATGTTATTATTTTTTATCCTTACCCTTTTAGTGAATTCAT
TTGCACAGATAAGTCTCAAATCCATTTCTGTAAGGCCTGTCCTGAGTGTG
ATTTCTACCTACCTTCCTCTCAAAAACAGTCGATTGATTNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNCTGAATACCCATTGTAAGTTAGGTACAGGGGTAGGTATTAGGAAT
TCAAAAATATGGTATCTATCTTTAGGATAATACTTCCTGTTCTCTACTGG
AGGTATTTTCTATTAACATGTCTCAATAATTCTTAAACTAAATATGTCAA
AACTGAAGTCTATGCTTTCTTGACACAGAGTCAATCATTCCTCATATTTC
CAGTGGCACCTTATATATTCAGCTCTCTAAGATAACAAACAGAATAATTT
TACACTTCCCCCAACCCTCTGTCGTGTCTGTCACTATCTCTAGCCAATTA
TTTTTCTCTAATGTTTTTGCTTCTCTTTTTTTCTTTCTTCTGCTGACACTT
TTATTCTGGTAGTGGGCCTTTTTCACTCCATGCATAGGTAGCCTTAACTA
GCTATTTTTAGTCTTCCAGGCTTTTGCCCATTCATCTGTTATATCTTACG
CCACAGCATGAGAATCATCTTGTAACACAATTCCATCACACACACCCCTG
CTTAGCTTTATAATATTTCTCTCTAATACTAGTTATACCAGATCCCAACT
CCTTAGACTGATGTGCAAAGTACTCTAAATTCCTACCCACTTACTCTCTC
CACTCCCATCTCACCAAGGTTAGTTCTCATTAATGAAATGAAAGGTCTGA
AGATCAGAATGCAAAGCTGATCTGNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
```

Figure 4.10.3 Web RepeatMasker result from an example run showing the Masked Sequence annotations section, which lists the repetitive elements masked sequences (replaced with Ns). See Guidelines for Understanding Results for explanation.

3. Select a method for returning results from the two radio buttons next to “return method”: “html” or “email.”

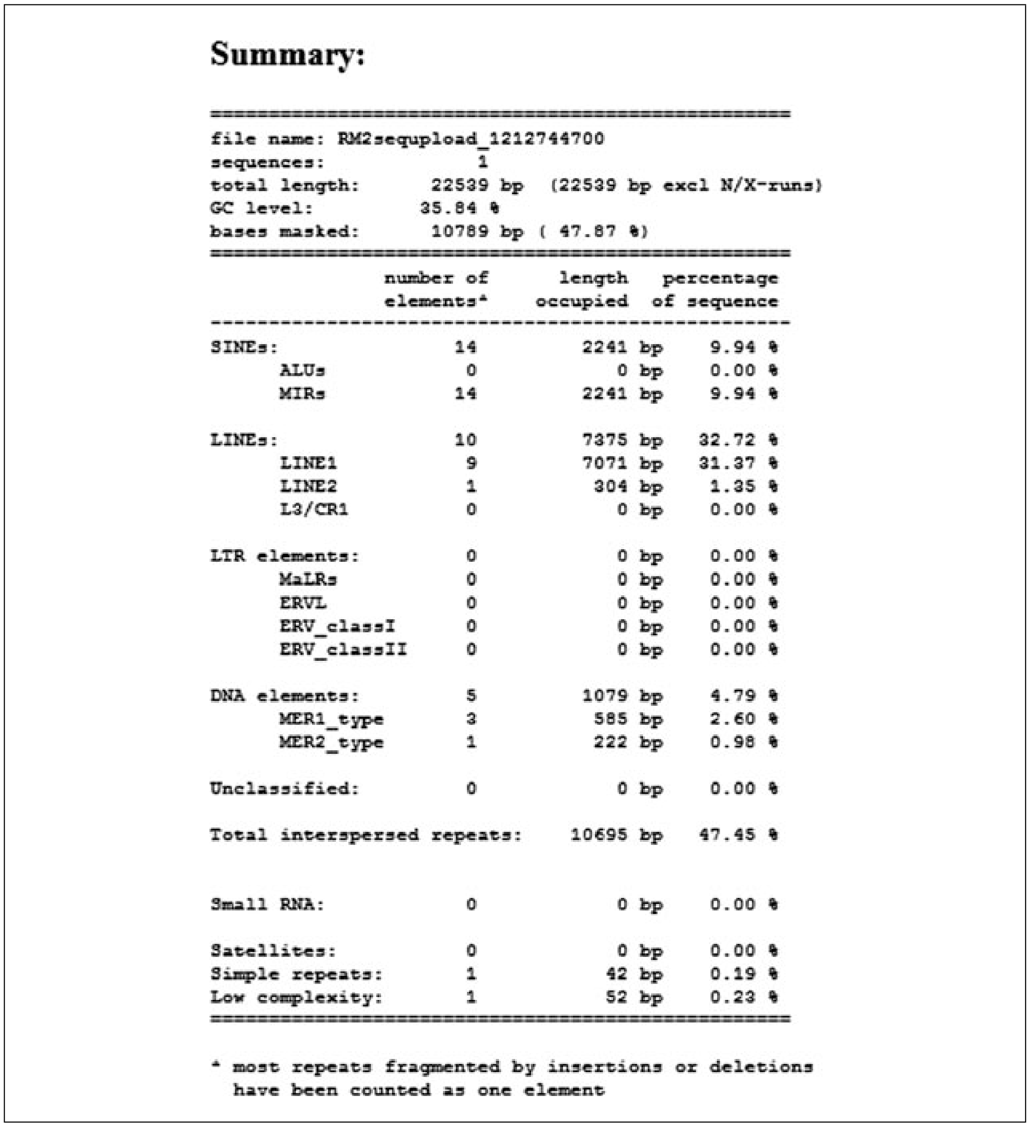
   If “html”is selected in this step and “html” format was selected in step 2, above, all of the results will be displayed in the browser. If “html”is selected in this step and “tar file” was selected in step 2, the results will be provided as links in the browser. If “email” is selected, one should enter one’s e-mail address so that the results can be sent via e-mail. For this example, select “html.”

4. At this stage, one can choose to click the Submit Sequence button to start running RepeatMasker with the other options set at default values. If the default settings do not satisfy one’s needs, continue with steps 5 to 8 and submit the sequence at step 9.

   For this example, click Submit Sequence with other options set at default values. The results that will be displayed on the browser are shown in Figures 4.10.2, 4.10.3, 4.10.4, and 4.10.5. See Guidelines for Understanding Results for details.

5. Adjust speed by selecting among the four radio buttons next to Speed/Sensitivity: “rush,” “quick,” “default,” or “slow.”

   Note that a faster speed is associated with a lower sensitivity. For example shown here, select “default” for Speed/Sensitivity. See Guidelines for Understanding Results for details.



Summary:

```
==================================================
file name: RM2sequpload_1212744700
sequences:            1
total length:      22539 bp  (22539 bp excl N/X-runs)
GC level:         35.84 %
bases masked:      10789 bp ( 47.87 %)
==================================================
               number of      length   percentage
               elements*    occupied  of sequence
--------------------------------------------------
SINEs:              14         2241 bp    9.94 %
      ALUs           0            0 bp    0.00 %
      MIRs          14         2241 bp    9.94 %

LINEs:              10         7375 bp   32.72 %
      LINE1          9         7071 bp   31.37 %
      LINE2          1          304 bp    1.35 %
      L3/CR1         0            0 bp    0.00 %

LTR elements:        0            0 bp    0.00 %
      MaLRs          0            0 bp    0.00 %
      ERVL           0            0 bp    0.00 %
      ERV_classI     0            0 bp    0.00 %
      ERV_classII    0            0 bp    0.00 %

DNA elements:        5         1079 bp    4.79 %
      MER1_type      3          585 bp    2.60 %
      MER2_type      1          222 bp    0.98 %

Unclassified:        0            0 bp    0.00 %

Total interspersed repeats:   10695 bp   47.45 %


Small RNA:           0            0 bp    0.00 %

Satellites:          0            0 bp    0.00 %
Simple repeats:      1           42 bp    0.19 %
Low complexity:      1           52 bp    0.23 %
==================================================

* most repeats fragmented by insertions or deletions
  have been counted as one element
```

Figure 4.10.4 Web RepeatMasker result from an example run showing the Summary section, which summarizes and categorizes repetitive elements found in the query DNA sequence. See Guidelines for Understanding Results for explanation.

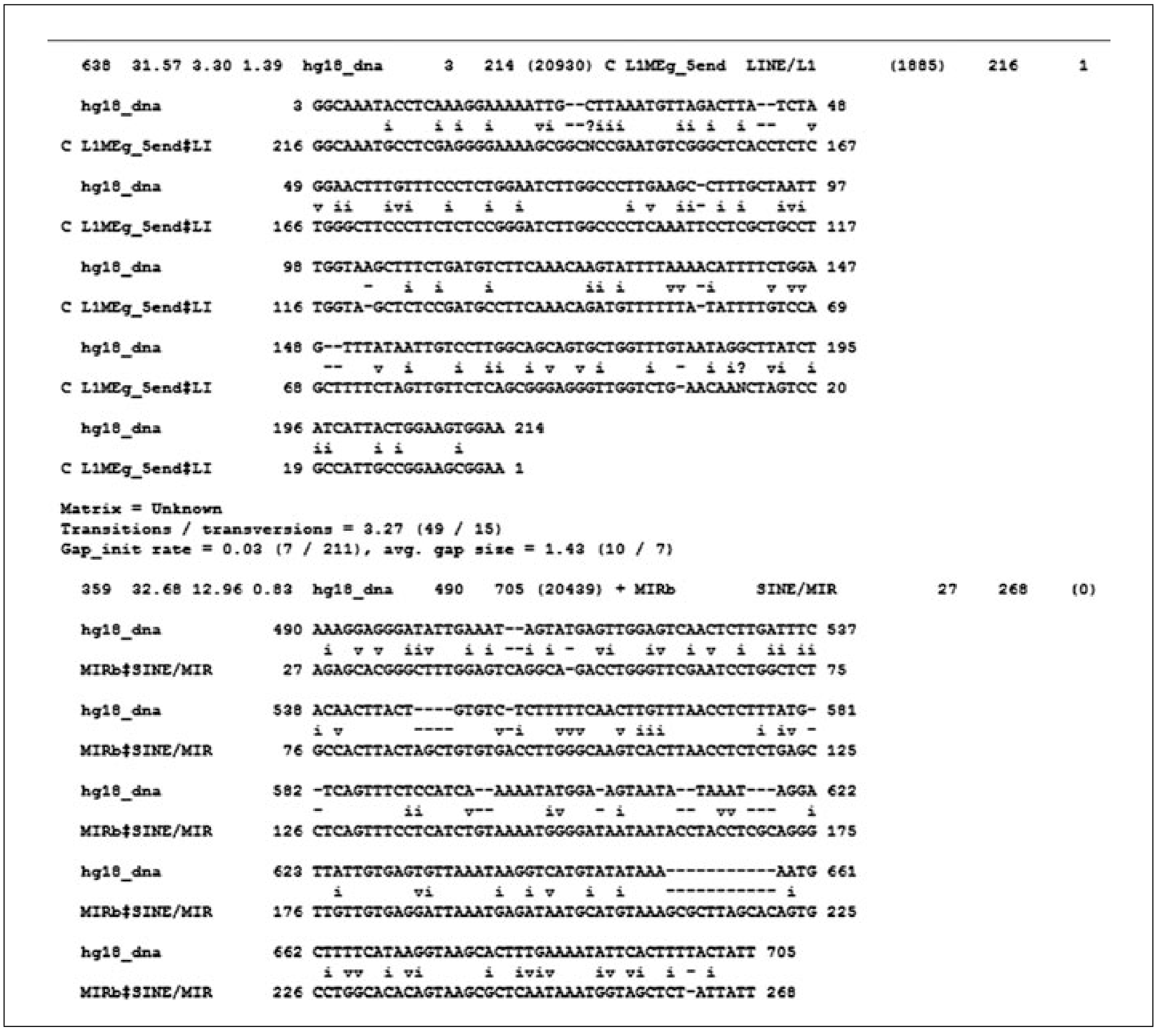
6. Select one of the entries from the pull-down menu next to "DNA source," each of which corresponds to a different repetitive element library.

   The default is Human. For the example here, select Human because the sequence is from the human genome.

   Note that if the query sequence is from an organism that is not listed here, the command-line version of RepeatMasker must be run locally (see Basic Protocol 2), and an appropriate repeat file from Repbase Update must be used, if there is one. If working with a genome for which Repbase does not have an appropriate repeat library, RECON (Bao and Eddy, 2002; Stein et al., 2003) or RepeatScout (http://bix.ucsd.edu/repeatscout/; Price et al., 2005) can be used to establish one from scratch.

7. In the series of pull-down menus, radio buttons, and check boxes under Lineage Annotation Options, select the appropriate options.

   These options are self-explanatory. For example, if Comparison Species is selected, the lineage-specific repeats are annotated with the RepeatMasker output with respect to the selected species.



```
  638  31.57 3.30 1.39  hg18_dna       3   214 (20930) C L1MEg_5end  LINE/L1       (1885)    216      1

  hg18_dna              3 GGCAAATACCTCAAAGGAAAAATTG--CTTAAATGTTAGACTTA--TCTA 48
                                 i    i i  i    vi --?iii     ii i  i --   v
C L1MEg_5end#LI       216 GGCAAATGCCTCGAGGGGAAAAGCGGCNCCGAATGTCGGGCTCACCTCTC 167

  hg18_dna             49 GGAACTTTGTTTCCCTCTGGAATCTTGGCCCTTGAAGC-CTTTGCTAATT 97
                          v ii   ivi   i   i  i          i v  ii- i i   ivi
C L1MEg_5end#LI       166 TGGGCTTCCCTTCTCTCCGGGATCTTGGCCCCTCAAATTCCTCGCTGCCT 117

  hg18_dna             98 TGGTAAGCTTTCTGATGTCTTCAAACAAGTATTTTAAAACATTTTCTGGA 147
                               -   i  i    i         ii i    vv -i     v vv
C L1MEg_5end#LI       116 TGGTA-GCTCTCCGATGCCTTCAAACAGATGTTTTTTA-TATTTTGTCCA 69

  hg18_dna            148 G--TTTATAATTGTCCTTGGCAGCAGTGCTGGTTTGTAATAGGCTTATCT 195
                           --   v  i    i  ii  i v  v i    i  -  i i?  vi  i
C L1MEg_5end#LI        68 GCTTTTCTAGTTGTTCTCAGCGGGAGGGTTGGTCTG-AACAANCTAGTCC 20

  hg18_dna            196 ATCATTACTGGAAGTGGAA 214
                          ii    i i     i
C L1MEg_5end#LI        19 GCCATTGCCGGAAGCGGAA 1

Matrix = Unknown
Transitions / transversions = 3.27 (49 / 15)
Gap_init rate = 0.03 (7 / 211), avg. gap size = 1.43 (10 / 7)

  359  32.68 12.96 0.83  hg18_dna    490   705 (20439) + MIRb        SINE/MIR          27    268    (0)

  hg18_dna            490 AAAGGAGGGATATTGAAAT--AGTATGAGTTGGAGTCAACTCTTGATTTC 537
                           i  v v  iiv   ii  --i i -  vi   iv  i v  i  ii ii
  MIRb#SINE/MIR        27 AGAGCACGGGCTTTGGAGTCAGGCA-GACCTGGGTTCGAATCCTGGCTCT 75

  hg18_dna            538 ACAACTTACT----GTGTC-TCTTTTTCAACTTGTTTAACCTCTTTATG- 581
                          i v       ----    v-i   vvv   v iii         i iv -
  MIRb#SINE/MIR        76 GCCACTTACTAGCTGTGTGACCTTGGGCAAGTCACTTAACCTCTCTGAGC 125

  hg18_dna            582 -TCAGTTTCTCCATCA--AAAATATGGA-AGTAATA--TAAAT---AGGA 622
                          -        ii    v--     iv   - i     --  vv ---   i
  MIRb#SINE/MIR       126 CTCAGTTTCCTCATCTGTAAAATGGGGATAATAATACCTACCTCGCAGGG 175

  hg18_dna            623 TTATTGTGAGTGTTAAATAAGGTCATGTATATAAA-----------AATG 661
                            i       vi      i  i v   i  i    ----------- i
  MIRb#SINE/MIR       176 TTGTTGTGAGGATTAAATGAGATAATGCATGTAAAGCGCTTAGCACAGTG 225

  hg18_dna            662 CTTTTCATAAGGTAAGCACTTTGAAAATATTCACTTTTACTATT 705
                           i vv  i vi      i  iviv    iv vi  i - i
  MIRb#SINE/MIR       226 CCTGGCACACAGTAAGCGCTCAATAAATGGTAGCTCT-ATTATT 268
```

Figure 4.10.5 Alignments between query sequence and consensus repetitive elements are shown if the option Show Alignments is selected.

8. In the series of pull-down menus under Advanced Options, select the appropriate Options.

   These options are straightforward as well. For example, if the user wants to make a choice between the Masking Options, users can either choose ambiguous characters, like "N" or "X" for masking, or lowercase letters, which may be more appropriate for subsequent alignments. Detailed explanation of these and additional options available can be accessed by clicking on the link to the right of each pull-down menu.

9. Click the Submit Sequence button to run RepeatMasker.

   The results displayed in the browser are shown in Figures 4.10.2, 4.10.3, 4.10.4, and 4.10.5. See Guidelines for Understanding Results for details.

BASIC PROTOCOL 2

USING THE COMMAND-LINE Unix/Linux VERSION OF RepeatMasker TO STUDY REPETITIVE ELEMENTS IN GENOMIC SEQUENCES

Command-line RepeatMasker provides users with more choices and does not have the 100-kb length limit for query sequences. To run RepeatMasker locally, one must obtain RepeatMasker, cross_match, and correct repetitive libraries from Repbase Update, as detailed below. It is also possible to run RepeatMasker with WU-BLAST (see Alternate Protocol) for faster processing.

NOTE: Investigators unfamiliar with the Unix environment should read *APPENDIX 1C* and *APPENDIX 1D*.

Necessary Resources

Hardware

Any Unix or Linux workstation

Software

RepeatMasker: The software is now licensed under the Open Source License v. 2.1, and can be downloaded from *http://www.repeatmasker.org/RMDownload.html*.

cross_match: This software is part of the Phred/Phrap/Consed (*http://www.phrap.org/consed/consed.html#howToGet*; also see *UNIT 11.2*) package. It is also free for academic use. Write to Phil Green (*phg@u.washington.edu*) and include the following information in the message: (a) name; (b) an acknowledgement of agreement to observe the licensing conditions described on the above Web site (state that cross_match is desired); (c) institution/department; (d) e-mail address for all future correspondence (ideally e-mail should be received through a Unix computer running a generic mail program, since several of the programs are sent as unencoded files which may be corrupted by some mail programs). Note that it takes up to 2 weeks for a license application to be processed.

Repbase Update: This database (*http://www.girinst.org/*; Jurka, 2001) manages a large selection of repetitive element libraries, which are required for running RepeatMasker. The library is free for download by academic users, who are required to set up accounts to access the database files by filling an online form (*http://www.girinst.org/accountservices/register.php*). Commercial users should contact Jolanta Walichiewicz (*jola@girinst.org*). Once again, if one's genome of interest does not have an appropriate repeat library file in Repbase Update, one can establish one with RECON (Bao and Eddy, 2002) or RepeatScout (*http://bix.ucsd.edu/repeatscout/*; Price et al., 2005). Stein et al. (2003) used RECON to establish a repeat library file for the round worms *C. elegans* and *C. briggsae*. RECON can also be obtained as part of a Repeat Modeler package, available for download from (*http://www.repeatmasker.org/RepeatModeler.html*). Alternatively, the RepeatScout software can also be used with RepeatMasker to identify and mask repeat family sequences from newly sequenced genomes.

Files

A FASTA file (*APPENDIX 1B*) or a collection of FASTA files can be processed via the command-line RepeatMasker. Note that there is essentially no size limit for query sequences for running RepeatMasker on the command line. The example file used in this protocol is the fully sequenced whole *Caenorhabditis elegans* genome, 102,287,094 bp in length, downloaded from the WormBase (*http://www.wormbase.org*) FTP site (*ftp://ftp.wormbase.org/pub/wormbase/genomes/elegans/sequences/dna/*).

Prepare system

1. Download and install programs—RepeatMasker, Tandem Repeat Finder (TRF), cross_match, and WU-BLAST, as well as Repbase library files. RepeatMasker is a Perl script and can be put in any desired directory.

   `cross_match` *will be e-mailed to users after contacting the authors. With an account properly set up, Repbase Update will assign a user name and password to download the repetitive library files.*

   For this example, make a directory called `repeat` *in the home directory and then copy RepeatMasker, TRF, and cross_match into this directory. For this example, type:*

```
[mta57@grouse ~]$ mkdir repeat
[mta57@grouse ~]$ cd repeat
```

2. Change the permission of the programs.

   For this example, type:

```
[mta57@grouse repeat]$ chmod u+x RepeatMasker
[mta57@grouse repeat]$ chmod u+x cross_match
[mta57@grouse repeat]$ ln -s trf321.linux.exe trf
```

3. Set the correct paths by running the Configure Script.

   First, find out where Perl is installed:

```
[mta57@grouse ~]which perl
/usr/bin/perl
```

   Then, after changing to the directory `repeat` *and the directory* `RepeatMasker`*, get the current directory path using the command* `pwd`*:*

```
[mta57@grouse RepeatMasker]$ pwd
/home/mta57/repeat/RepeatMasker
```

   Then, do the same for the `TRF` *and cross_match to get the paths to the directories.*

   To configure the program use the following script:

```
[mta57@grouse repeat]$ cd RepeatMasker
[mta57@grouse RepeatMasker]$ perl ./configure
```

   Enter the required paths; for example, to write the path to the Perl interpreter, enter:

```
Enter path: /usr/bin/perl
```

   To write the path to the location where the RepeatMasker program has been installed, enter:

```
Enter path: /home/mta57/repeat/RepeatMasker
```

   For the path to the location where the TRF program can be found, enter:

```
Enter path: /home/mta57/repeat
```

   To add a search engine, enter:

```
Enter path: /home/mta57/repeat/cross_match
```

4. Place repeat libraries in the correct directory (i.e., the same directory as the script `RepeatMasker`).

   Make sure that subdirectory `Libraries` *in the* `RepeatMasker` *directory contains* `RepeatMasker.lib` *and* `RepeatMaskerLib.embl` *files.*

5. Create a new directory for input and output files.

   Note that RepeatMasker output files will be written to the same directory as the input file resides.

   For this example, type the following:

```
[mta57@grouse repeat]$ mkdir RepeatMasker_file
[mta57@grouse repeat]$ cd RepeatMasker_file
[mta57@grouse RepeatMasker_file]$
```

   Next, download or copy the FASTA file (`current.dna.fa.gz`*) containing the sequence of C. elegans genome to the directory and unpack it:*

```
[mta57@grouse RepeatMasker_file]$ gunzip current.dna.
fa.gz
```

6. To get a brief description of the command-line parameters and options, type in the program name RepeatMasker on the command line.

   For this example:

   ```
   [mta57@grouse RepeatMasker_file]$ ../RepeatMasker/
   RepeatMasker
   ```

   The following contents will be returned:

   ```
   SYNOPSIS
   RepeatMasker [-options] <seqfiles(s) in fasta format>
   ...
   default settings are for masking all type of repeats
   in a primate sequence.
   ...
   ```

   Choose from a number of options:

   ```
   -q Quick search; 5-10% less sensitive, 2-5 times
   faster than default
   -nolow Do not mask low_complexity DNA or simple
   repeats
   -div [number] Mask only those repeats < x percent
   diverged from consensus seq
   ...
   -species <query species> Specify the species or
   clade of the input sequence (choose only one!)
   ...
   contamination options
   ...
   running options
   ...
   output options
   ...
   ```

   To get detailed help, type in:

   ```
   [mta57@grouse RepeatMasker_file]$../RepeatMasker/
   RepeatMasker -h
   ```

Run RepeatMasker

7. Run command-line version of RepeatMasker on the local system:

   ```
   % /path/to/RepeatMasker -el current.dna.fa
   ```

   For this example, run:

   ```
   [mta57@grouse RepeatMasker_file]$ ../RepeatMasker/
   RepeatMasker -species elegans current.dna.fa
   ```

   Because the example sequence is from C. elegans, the `-species elegans` *command is used, so that the C. elegans Repbase repetitive element library file is used.*

   The result files will be written into the directory `RepeatMasker_file`*, the same directory where the query sequence file(s) reside(s). For this example, the result files include:*

   ```
   current.dna.fa.masked
   current.dna.fa.log
   ```

```
current.dna.fa.dna.cat
current.dna.fa.dna.out
current.dna.fa.dna.tbl
```

The result files are explained in Guidelines for Understanding Results, below.

8. RepeatMasker provides users with a large array of options to meet the needs appropriate for different cases. Here, only commonly used ones are covered. For more advanced options, users are encouraged to read the help file `repeatmasker.help`, which comes with the RepeatMasker program package.

   Note that the order of the command-line options is not important when entering multiple commands.

   a. `Species options` and the `-lib` flag allow users to specify a particular library file for the corresponding organism. RepeatMasker provides common name flags for some species, like `-cat` or `-dog`, but not for all. For that reason, usage of Latin names as a species option is highly recommended. Users can also provide a repeat library file, especially if the library file is not from Repbase collection, to RepeatMasker using a `-lib` flag. The default repeat library is for primate.

   To establish one's own repeat library for RepeatMasker, use the format for IDs as recommended by `repeatmasker.help`*, e.g.:*

   ```
   >repeatname#class/subclass
   ```

   or, simply

   ```
   >repeatname#class
   ```

   b. Masking options are used for determining what kind of repeats are masked. Commonly used options within this category are: `-cutoff`, `-nolow`, and `-div`. The option `-cutoff` sets cutoff score for masking repeats when using `-lib`. The default cutoff score is 225. Lower scores give more false matches. A `-nolow` flag causes RepeatMasker not to mask low-complexity DNA or simple repeats.

   The `-div` *option sets the divergence level to limit the masking and annotation to a subset of less diverged (younger) repeats.*

   c. Some options are used to control processing speed and search parameters. Options that affect processing speed are:

`-q`	quick search; 5% to 10% less sensitive, 3× to 4× faster than default
`-qq`	rush job; ∼10% less sensitive
`-s`	slow search; 0% to 5% more sensitive; 2.5× slower than default

   These flags make significant differences when the input sequences are long. If only a quick check is desired, the `-qq` *flag may be used for fast results. On the other hand, if the quality of the result is more critical, the default (with none of the above options selected), or even* `-s`*, should be used.*

   It is possible to recruit more processors for RepeatMasker by using the `-pa(rallel)` *flag, which only works when there are many input files or if the query files are big (>50 kb).*

   WU-BLAST can be used to replace cross_match if the flag `-w(ublast)` *(see Alternate Protocol) is used.*

d. Output options support the following frequently used formats (for other available options refer to `repeatmasker.help`):

`-a`	shows the alignments in a `.align` output file;
`-small`	returns complete `.masked` sequence in lower case
`-xsmall`	returns repetitive regions in lowercase (rest capitals) rather than masked
`-x`	returns repetitive regions masked with Xs rather than Ns
`-gff`	creates an additional General Feature Finding format output

Note that the `-cut` *option is not supported in the current release of RepeatMasker; however, the function may be obtained by contacting Robert Hubley (rhubley@systemsbiology.org).*

ALTERNATE PROTOCOL

RUNNING REPEATMASKER WITH WU-BLAST

Running RepeatMasker for larger sequences (e.g., whole genome for *Homo sapiens*) will take a significant amount of time. The processing time can be reduced roughly 30-fold by using WU-BLAST as the engine for RepeatMasker, to replace cross_match (Bedell et al., 2000). Although RepeatMasker with WU-BLAST has better processing time, the combination also has some limitations: (1) low-complexity repeats are not as efficiently masked as when RepeatMasker is used with cross_match; (2) some output formats are not supported; and (3) the accuracy of the results returned by the combination of RepeatMasker with WU-BLAST has not been assessed.

NOTE: Investigators unfamiliar with the Unix environment should read *APPENDIX 1C* and *APPENDIX 1D*.

Necessary Resources

Hardware

Unix or Linux workstation

Software

RepeatMasker (see Basic Protocol 2)
WU-BLAST 2.0: contact *licensing@blast.wustl.edu*
Repbase Update repeat libraries (see Basic Protocol 2)

Files

A FASTA file or a collection of FASTA files (*APPENDIX 1B*). Note that there is no size limit for running RepeatMasker with WU-BLAST on command line. The example file used in this protocol is the fully sequenced whole *C. elegans* genome, 102,287,094 bp in length, downloaded from the WormBase (*http://www.wormbase.org*) FTP site (*ftp://ftp.wormbase.org/pub/wormbase/genomes/elegans/sequences/dna/*).

1. Download and install programs—RepeatMasker, WU-BLAST, and Repeat library files. Note that until June 2004, MaskerAid (Bedell et al., 2000) was necessary for the WU-BLAST to be used with the RepeatMasker. That functionality is now implemented and does not need to be integrated separately. For this example, make a directory called `repeat` and then copy the `RepeatMasker/` directory into this directory. To do this, first change to the home directory and then make a new directory named `repeat` using `mkdir`. Use `cd` to change directory to `repeat`, as follows:

```
[mta57@grouse ~]mkdir repeat
[mta57@grouse ~]cd repeat
```

Copy `RepeatMasker/` *into this directory. Copy WU-BLAST package into this directory as well and unpack it:*

```
[mta57@grouse repeat]$ gunzip -WU_BLAST |tar xvf -
```

`wu_blast/` *directory will be seen after unpacking*

Programs within the `wu_blast/` *directory, like* `blastp`*, and* `blastx`*, are executable after unpacking.*

2. Change the permission of the programs and the directories.

   For this example:

   ```
   [mta57@grouse repeat]$ chmod u+x RepeatMasker
   [mta57@grouse repeat]$ chmod u+x wu-blast
   ```

3. Set the correct paths by running the Configure Script, as described in Basic Protocol 2.

   To add a WU-BLAST search engine, enter:

   ```
   Enter path: /home/mta57/repeat/wu-blast
   ```

4. Create a new directory for input and output files.

   RepeatMasker output files will be written to the same directory as the input file resides.

   For this example, type the following:

   ```
   [mta57@grouse repeat]$ mkdir RepeatMasker_file
   [mta57@grouse repeat]$ cd RepeatMasker_file/
   [mta57@grouse RepeatMasker_file]$
   ```

   Next, download or copy the FASTA file (`current.dna.fa.gz`*) for C. elegans genome to the directory and unpack it:*

   ```
   [mta57@grouse RepeatMasker_file]$ gunzip current.dna.
   fa.gz
   ```

5. Run program on command line using the flag `-w(ublast).`

   For this example, run:

   ```
   [mta57@grouse RepeatMasker_file]$ ../RepeatMasker/
   RepeatMasker -w -species elegans current.dna.fa
   ```

   Here the flag `-w` *is used to indicate that WU-BLAST is used as the matching engine; the* `-species elegans` *is used to indicate that the C. elegans Repbase repetitive element library file is used, since the sequence is from C. elegans. Note that species names that contain multiple words need to be bracketed by quotation marks (e.g.,* `"caenorhabditis elegans"`*).*

   Other than the `-w` *option, which indicates that WU-BLAST is used, the command-line parameters and options are similar to those in Basic Protocol 2.*

GUIDELINES FOR UNDERSTANDING RESULTS

The output of RepeatMasker is written into five different files in the same directory where the query sequence or sequences reside. Only three files, those with `.out`, `.masked`, and `.tbl` extensions, contain results; others store processing information and are therefore not detailed here. If RepeatMasker is run via the Web server interface, the contents of these three files are written into one page (file), shown in Figures 4.10.2, 4.10.3, and 4.10.4, respectively.

Table 4.10.1 Columns of the `.out` File from Left to Right (also see Fig. 4.10.2)

Column	Content
SW score	Smith-Waterman score of the match
Perc div.	Percent substitutions in matching region compared to the consensus
Perc del.	Percent of bases opposite a gap in the query sequence (deleted bp)
Perc ins.	Percent of bases opposite a gap in the repeat consensus (inserted bp)
Query sequence	Name of query sequence
Position in query	
Begin	Starting position of match in query sequence
End	End position of match in query sequence
(Left)	Number of bases in query sequence past the end position of the current match
Matching repeat	
Repeat	Name of repeat
Class/family	The class of the repeat
Position in repeat[a]	
Begin	Starting position of match in repeat consensus sequence
End	End position of match in repeat consensus sequence
(Left)	Number of bases in repeat consensus sequence past the end of the current match
ID	Repeat identification number

[a] Note that if the repeat consensus matches the positive strand, the three subcolumns are begin, end, and (left); otherwise, the three subcolumns are (left), end, and begin.

The `.out` file (Fig. 4.10.2 in the Web example) is the annotation file that contains the cross_match summary lines. The file is basically self-explanatory. The columns of the `.out` file are described briefly in Table 4.10.1.

The matches (domains) are masked in the `.masked` file. This file can be parsed with the help of the BioPerl module (Bio::Tools::RepeatMasker, *http://www.bioperl.org*). The `.masked` file (Fig. 4.10.3) is the same as the query sequence, except that the repetitive elements are masked using Ns, Xs, or lowercase letters (if one has a `-x` or `-xsmall` flag on command line or checked the box "Mask with Xs or lower case to distinguish masked regions from Ns already in query" on the RepeatMasker Web site).

The `.tbl` file (Fig. 4.10.4) summarizes the annotation results shown in the `.out` file. Notably, the `.tbl` file states the percentage repetitive elements coverage.

COMMENTARY

Background Information

How RepeatMasker works

RepeatMasker finds and masks repetitive elements by aligning each of the query sequence(s) with each of the repeat consensus sequences in the repeat library file. Usually, cross_match is the engine that does the alignment, while RepeatMasker manages the whole process and parses the alignments. The program cross_match implements the Smith-Waterman (SW) alignment algorithm (Smith and Waterman, 1981). The program cross_match is one of the best applications for sequence alignment. The drawback of cross_match is that it is slow. To make RepeatMasker process faster, WU-BLAST can be used to replace cross_match (see Alternate Protocol). The alignment program WU-BLAST is a heuristic alignment algorithm. However, the sensitivity is reduced when running RepeatMasker with WU-BLAST.

Critical Parameters and Troubleshooting

Limitations and known bugs

For files with multiple long sequences (e.g., a file containing whole-chromosome sequences), RepeatMasker does not work well. All of the output entries are mislabeled as the first sequence (chromosome). There is a default maximal sequence length of 4 Mb. There are two ways to work around this limitation. One way is to change the default maximal sequence length value in the RepeatMasker script. Find the following line in the script:

```
$maxsize = 4000000;
```

and modify the value. Note that the memory requirements of the program go up as this value is increased. Another way is to break down each long (>4 Mb) sequence into shorter ones.

RepeatMasker does not fail explicitly even if one's hard disk is full; it actually gives apparently normal results. Therefore, when it is noticed that the results are far from those expected, there might be a disk space problem.

Using `-q` or `-qq` (see Basic Protocol 2, step 8) can speed things up, but the sensitivity is reduced. When WU-BLAST is used, the `-s` (slow) option is preferred, since the speed with WU-BLAST is reasonably fast and the masking results are better.

Analysis on smaller sequences (<2 kb) could be less accurate.

Note that previous version(s) of RepeatMasker had a problem with overwriting the files with the same names when multiple analyses were performed on the same input files. This is no longer a problem, since RepeatMasker creates output directories for each analysis.

Literature Cited

Bao, Z. and Eddy, S.R. 2002. Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res.* 12:1269-1276.

Bedell, J.A., Korf, I., and Gish, W. 2000. MaskerAid: A performance enhancement to RepeatMasker. *Bioinformatics* 16:1040-1041.

Jurka, J. 2001. Repbase update, a database and an electronic journal of repetitive elements. *Trends Genet.* 16:418-420.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. 2005. Repbase update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* 110:462-467.

Price, A.L., Jones, N.C., and Pevzner, P.A. 2005. De novo identification of repeat families in large genomes. *Bioinformatics* 21:Suppl 1:i351-358.

Smith, T.F. and Waterman, M.S. 1981. Identification of common molecular subsequences. *J. Mol. Biol.* 147:195-197.

Stein, L.D., Bao, Z., Blasiar, D., Blumenthal, T., Brent, M., Chen, N., Chinwalla, A., Clarke, L., Clee, C., Coghlan, A., Coulson, A., D'Eustachio, P., Fitch, D.H.A., Fulton, L., Fulton, R., Griffiths-Jones, S., Harris, T.W., Hillier, L.W., Kamath, R., Kuwabara, P.E., Marra, M., Mardis, E., Miner, T., Minx, P., Mullikin, J.C., Plumb, R.W., Rogers, J., Schein, J., Sohrmann, M., Spieth, J., Stajich, J.E., Wei, C., Willey, D., Wilson, R., Durbin, R., and Waterston, R. 2003. The genome sequence of *Caenorhabditis briggsae*: A platform for comparative genomics. *PLoS Biol.* 1:E45.

Internet Resources

http://www.repeatmasker.org/
RepeatMasker Web server

http://www.girinst.org/
Repbase Update

http://selab.janelia.org/recon.html
RECON Web site

http://bix.ucsd.edu/repeatscout/
RepeatScout Web site

http://www.phrap.org/consed/consed.html#howToGet
cross_match Web site

http://blast.wustl.edu/
WU-BLAST Web sites

http://genome.ucsc.edu/cgi-bin/hgGateway
UCSC Genome Browser

ftp://ftp.wormbase.org/pub/wormbase/genomes/elegans/sequences/dna/
WormBaseFTP site

http://www.repeatmasker.org/RepeatModeler.html
RECON site, the newest version of RECON is available from the RepeatMasker

http://www.bioperl.org
BioPerl Web site