Nucleic Acids Research, 2006, Vol. 34, Web Server issue ***W651–W655***
doi:10.1093/nar/gkl125

SNPmasker: automatic masking of SNPs and repeats across eukaryotic genomes

Reidar Andreson[1,2], Tarmo Puurand[1] and Maido Remm[1,2,*]

[1]Department of Bioinformatics, University of Tartu, Estonia and [2]Estonian Biocentre, Tartu, Estonia

Received February 14, 2006; Revised February 23, 2006; Accepted March 13, 2006

ABSTRACT

SNPmasker is a comprehensive web interface for masking large eukaryotic genomes. The program is designed to mask SNPs from recent dbSNP database and to mask the repeats with two alternative programs. In addition to the SNP masking, we also offer population-specific substitution of SNP alleles in genomic sequence according to SNP frequencies in HapMap Phase II data. The input to SNPmasker can be defined in chromosomal coordinates or inserted as a sequence. The sequences masked by our web server are most useful as a preliminary step for different primer and probe design tasks. The service is available at http://bioinfo.ebc.ee/snpmasker/ and is free for all users.

INTRODUCTION

Human genome contains millions of single nucleotide polymorphisms (SNPs). There are many different technologies for determining the alleles of SNP markers in human DNA samples (1). Most of these technologies use PCR and/or primer extension for analysis of SNPs. Unfortunately, primer-based technologies are sensitive to repeats and to variations in the genome. Repeats around SNPs may cause failure of assay or give mixed signals from different genomic regions. Variations may cause biased signal due to allele-specific binding of primers. The SNPs in human genome are not distributed uniformly. In the current dbSNP database (release 125) ~2 million SNPs out of total number of ten million are located within 25 bp or less from another SNP. A previous study has demonstrated that closely located SNPs may have affected the performance of assays in the Human HapMap Project (2), because of the interference with primers/probes used in assays. Thus, it is important to avoid both repeats and SNPs within primers. The most efficient way to avoid unwanted regions within primers is masking the template sequence before designing primers. DNA masking is typically done by replacing nucleotide regions with certain properties with 'N' characters, or by converting the nucleotides within the region to lower-case letters. Repeats are most frequently masked by the program called RepeatMasker (Smit,A.F.A., Hubley,R. and Green,P. http://www.repeatmasker.org/). Low-complexity regions are often masked by DUST program, a built-in part of the BLAST software package (3).

Several programs exist that offer masking of SNP locations in user-defined genomic regions. For example, the Genome Browser at UCSC website (http://genome.ucsc.edu/) allows retrieving masked DNA for user-specified regions. Both SNPs and the repeats can be masked with different options. Similar service with fewer options is offered by the SNP Research Facility at Washington University (http://snp.wustl.edu/bio-informatics/human-snp-annotation.html) and by the Institute of Human Genetics, GSF, Germany (http://ihg.gsf.de/cgi-bin/snps/seq1.pl).

All the previously mentioned websites offer retrieval of masked DNA by entering chromosomal coordinates. This is necessary because SNP locations are typically defined by chromosomal coordinates. However, often users do not know the exact coordinates of their DNA region of interest in a given assembly. In this case the location of the query sequence within the genome should be determined before masking of SNPs. One way to do that is by sequence homology search. For example, SNP BLAST web interface at NCBI (http://www.ncbi.nlm.nih.gov/SNP/snpblastByChr.html) allows searching with user's query sequence against the sequence database of SNP flanking regions. After finding homologies, SNP BLAST highlights the differences between the query and the target sequence, including SNP positions. The server can also mask human repeats. Searching SNPs in other genomes represented in dbSNP (4) is possible, but the retrieval of longer genomic regions by coordinates is not possible with this program. Thus, the existing web pages allow the retrieval of masked sequences by chromosomal coordinates or by homology search, but none of them allows both (Table 1).

We have put these functionalities for SNP and repeat-masking together into one web service. In addition, we have added a possibility to mask the repeats by a different and more specialized manner using a custom-made

*To whom correspondence should be addressed. Tel: +372 7375001; Fax: +372 7420286; Email: maido.remm@ut.ee

Table 1. Comparison of different web pages for masking SNPs and repeats

	SNP masking types	Repeat-masking programs	Region defined by coordinates	Region found by sequence homology search
Genome browser UCSC	lower-case, by color, bold/italic	RepeatMasker	Yes	No
SNP Research Facility Washington University in St. Louis	IUPAC	RepeatMasker	Yes	No
GSF Munich, Germany	'N'	RepeatMasker	Yes	No
SNP BLAST NCBI	IUPAC	RepeatMasker	No	Yes
SNPmasker University of Tartu	any character, IUPAC, lower-case, by HapMap frequency	RepeatMasker, GenomeMasker	Yes	Yes

program called GenomeMasker (5) and a possibility to change SNP alleles in a sequence.

IMPLEMENTATION

Input

SNPmasker is currently able to mask sequences from two genomes: human and mouse. However, the program can easily be configured to accept additional genomes from the ENSEMBL database (6). The sequence of interest can be defined in two principally different ways: by chromosomal coordinates and by sequence (Figure 1). The sequence can be inserted by pasting it into the text box or uploading file in FASTA format.

Databases

SNPmasker uses sequence databases from ENSEMBL (ftp://ftp.ensembl.org/pub/release-35/). For each genome, all tables from database *homo_sapiens_core_35_35h* and from *mus_musculus_35_34c* are installed. SNP locations are retrieved from dbSNP database (ftp://ftp.ncbi.nih.gov/snp/). Additional database is required for storing HapMap (7) allele frequency data, which we create locally. The HapMap database is created by counting and storing allele frequencies of each SNP in each population. The counting is done by using tables downloaded from the Phase 2 database (http://www.hapmap.org/genotypes/2006-01/non-redundant/).

Localization of input sequences

SNP masking can only be done if the location of a sequence in the genome is known. If the location is not defined by the user, then it must be found by a homology search. The homology search is performed by MEGABLAST (8) program against chromosome sequences. The location is considered unique if MEGABLAST finds a single match with length of 90% of query sequence that must have >90% identity with the target genome. The query sequence must contain at least one 100 bp long exact match (or 16 bp, if the user-given sequence is shorter than 100 bp). If a unique location cannot be determined by MEGABLAST, the masking is cancelled. In this case MEGABLAST alignments are presented to the user for further analysis and corrections of the input sequences/coordinates.

Masking of SNPs

The program has two major functionalities that can be used either together or separately—the masking of SNP positions and the masking of repeats. SNP masking is implemented as follows. Once the location is determined by the user or by a homology search, the program compares the coordinates of sequence with the coordinates of known SNPs and verifies whether the given sequence region contains any SNPs. If a SNP is found within the sequence, it is masked by replacing the existing character with a lower-case character, any user-defined symbol or an IUPAC symbol. A unique option, not offered by other similar services, is changing the sequence according to SNP major allele nucleotide—the nucleotide that is most frequent in certain human populations. This option is available for the human genome only and is based on the HapMap Phase2 data (public release #20), offering separate masking for CEPH, Japanese, Chinese and African (Yoruban) populations. When comparing the human genomic sequence with the HapMap genotype data, we discovered that in the current Golden Path sequence about 25% of SNP positions (~900 000 nt over the whole genome) are representing a minor allele—the less frequent variant of a nucleotide. These figures are similar for all four populations represented in the HapMap. Changing sequence at these positions to represent most frequent allele in a given population might be useful for research projects, which are concentrating on individuals (or cell lines) from a given population only. Deletions or insertions are currently not masked.

Masking of repeats

Masking of repeats is optional. If the repeats are masked, the masking can be performed with either RepeatMasker or GenomeMasker. GenomeMasker is a novel masking program that was developed specifically for the PCR primer design and therefore has several differences compared to the traditional masking programs like DUST or RepeatMasker. GenomeMasker exhaustively counts all 16 nt motifs in a given genome and masks the abundant (>10 occurrences by default) motifs. Because PCR primers are single-stranded, this masking method is also strand-specific—the 'upper' and 'lower' strands can be masked separately, if necessary. This is useful for the PCR primer design around the markers or for the other target regions in the genome. For the PCR primer design only the upper strand should be masked on the left side of the target region and only the lower strand should be masked on the right side of the target region. Primer design from the lower-case-masked sequences is facilitated by the program called GM_Primer3 (executable available at http://bioinfo.ebc.ee/download/, an on-line version can be found at http://bioinfo.ebc.ee/cgi-bin/primer3_www.cgi). The GM_Primer3 is essentially a modified version of Primer3 with additional

Figure 1. Web interface for SNPmasker input.

ability to reject the primer candidates ending with a lower-case letter.

EXAMPLES

SNPmasker is primarily designed for masking the sequences before primer and probe design. The exact masking style may vary depending on the purpose and on the technology requirements. Here are some examples of using SNPmasker for different purposes. For illustration we have used a region from human the chromosome 2, nucleotides 19 341 544–519 342 344. This region contains several SNPs and repeated regions.

If one is interested in the PCR primer design for a genomic PCR, then the best masking style is 'N' masking for SNPs and strand-specific lower-case masking for the repeats. When used together with GM_Primer3 program, this masking style

will avoid any SNPs within the primer and any repeats overlapping with the 3′-part of the primer. An example of such PCR-specific masking is shown in Figure 2B. This example demonstrates a case where the investigator is interested in amplification of the region around third SNP (rs851320). SNP together with the 25 bp of the flanking region is defined as a target for the amplification (shown in *italic*). GenomeMasker is used in a strand-specific mode which means that only the upper strand is masked on the left side of the target and only the lower strand is masked on the right side of target region. The strand-specific masking allows finding more potential PCR primer candidates in this repeat-rich region.

For hybridization probe design or for other purposes it may be useful to mask SNPs together with RepeatMasker (Figure 2C).

A. human_35 chr=2:19341544-19342344 (original sequence)

```
TTTAAAATTAGCTCAGGAAATTAACAGTTTATTATGGTCCTGTTAATGTTATTACTATTATTCACAACA
TCCTCCTCTCTCTATTTTTCTCAGGTTTATCAATTGGTCTACTGTAACTTCATTTTTCCAGATGGATAG
TGGGTACAATAAATAGCTTTAGCAAAAGCTTTAAATGACAAGTTTAAACATAATATTCCAAATGAAAAA
AAATAAGGAAATGGCACACATGCGCACATGCACATACATTCACAGGTGGCTTTTTAACATGGATGTAGA
AAGGTGTGTATGAAGTGGGGAGGGAGATTCATTGACACCAACAGCATTAAGAAAGCCACTGAGGTGGTT
TCTTTGGCTCCAGCAAAATTTGGGATCAATTTTTCTTGGCAAGAATTACGACTCCGGGGATGCTAAAAG
GACATTAAAGTAGCCACAGATAGAATTAACTCTATCTATAACAGATAGAATTAACGATTGCTCTGGAAG
CCTTCCAGAGTGGTGGGCTAATATAGCGTATCACAGAATTGGGAGCACAGGGGAACCACAGAGGTGAGA
AACCAGTCCTGGTTTCTCAATCACCACCAATCAATGTGGAGAGAGTCATTGTGTTAAGTTACTAGTGTA
AGTGAGGCCCTGTGATGGCCACCTGTAGCAGGGACTTTCGTTGTTCAGAGACTCTCTCTGTGTGTGTGT
GTGTGTGTGTGTGTGTGTCTGTGTGTTCGTGTGTGTTGGTCTCAGAAGTCACTTATTTAGAAAAAGGAA
GCCAACTTAATAAGAAAAGCAATAGGAAAAATTTGGTCACAG
```

B. GenomeMasker masking; target 375 425; SNPs masked with 'N'

```
tttaaaattagctcaggaaattaacagtttattatGGtcctgttaatgttattactattattcacaaca
tcctcctctctctatttttctcaggtttatcaaTTGGTCTactgtaacttcatttttccagatggatag
tgggtaNaataaatagctttagcaaaagctttaaatgacaagtttaaacataatattccaaatgaaaaa
aaataaggaaatggcacacatgcgcacatgcacatacattcacaggtggctttttaaCATGGATGTAGA
AAGGTGTGTatgaNgtggggagggagattcattGACACCAACAGCATTAAGAAAGCCACTGAGGTggtt
tctttggctccaGCAAAATTTGGGATCAATTTTTCTTGGCAAGAATTACGACTCCNGGGATGCTAAAAG
GACATTAAAGTAGCCACAGATAGAATTAACTCTATCTATAACAGANAGAATTAACGATTGctctggaag
ccttccagagtggtgggCTAATATAGCGTATCACAGAATTGGGAGCACAGGGGAAccacagaggtgaga
aaCCAGTCCTGGTTTCTCAATCACCACCAATCAATGTGGAGAGagtcattgtgttaagtTACTAGTGTA
AGTGAGGCCCTgtgatggccacctgtaGCAGGGACTTTCGTTGTTcagagactctctNtgtgtgtgtgt
gtgtgtgtgtgtgtgtgtctgtgtgttcgtgtgtgttggtctcagaagtcacttatttagaaaaaggaa
gccaacttaNtNagaaaagcaataggaaaaatttggTCACAG
```

C. RepeatMasker masking; SNPs masked with 'N'

```
TTTAAAATTAGCTCAGGAAATTAACAGTTTATTATGGTCCTGTTAATGTTATTACTATTATTCACAACA
TCCTCCTCTCTCTATTTTTCTCAGGTTTATCAATTGGTCTACTGTAACTTCATTTTTCCAGATGGATAG
TGGGTANAATAAATAGCTTTAGCAAAAGCTTTAAATGACAAGTTTAAACATAATATTCCAAATGAAAAA
AAATAAGGAAATGGCACACATGCGCACATGCACATACATTCACAGGTGGCTTTTTAACATGGATGTAGA
AAGGTGTGTATGANGTGGGGAGGGAGATTCATTGACACCAACAGCATTAAGAAAGCCACTGAGGTGGTT
TCTTTGGCTCCAGCAAAATTTGGGATCAATTTTTCTTGGCAAGAATTACGACTCCNGGGATGCTAAAAG
GACATTAAAGTAGCCACAGATAGAATTAACTCTATCTATAACAGANAGAATTAACGATTGCTCTGGAAG
CCTTCCAGAGTGGTGGGCTAATATAGCGTATCACAGAATTGGGAGCACAGGGGAACCACAGAGGTGAGA
AACCAGTCCTGGTTTCTCAATCACCACCAATCAATGTGGAGAGAGTCATTGTGTTAAGTTACTAGTGTA
AGTGAGGCCCTGTGATGGCCACCTGTAGCAGGGACTTTCGTTGTTCAGAGACTCTCTNtgtgtgtgtgt
gtgtgtgtgtgtgtgtgtctgtgtgttcgtgtgtgtTGGTCTCAGAAGTCACTTATTTAGAAAAAGGAA
GCCAACTTANTNAGAAAAGCAATAGGAAAAATTTGGTCACAG
```

D. SNPs masked with lower-case [HapMap-JPT], frequency > 80%

```
TTTAAAATTAGCTCAGGAAATTAACAGTTTATTATGGTCCTGTTAATGTTATTACTATTATTCACAACA
TCCTCCTCTCTCTATTTTTCTCAGGTTTATCAATTGGTCTACTGTAACTTCATTTTTCCAGATGGATAG
TGGGTAtAATAAATAGCTTTAGCAAAAGCTTTAAATGACAAGTTTAAACATAATATTCCAAATGAAAAA
AAATAAGGAAATGGCACACATGCGCACATGCACATACATTCACAGGTGGCTTTTTAACATGGATGTAGA
AAGGTGTGTATGAAGTGGGGAGGGAGATTCATTGACACCAACAGCATTAAGAAAGCCACTGAGGTGGTT
TCTTTGGCTCCAGCAAAATTTGGGATCAATTTTTCTTGGCAAGAATTACGACTCCaGGGATGCTAAAAG
GACATTAAAGTAGCCACAGATAGAATTAACTCTATCTATAACAGATAGAATTAACGATTGCTCTGGAAG
CCTTCCAGAGTGGTGGGCTAATATAGCGTATCACAGAATTGGGAGCACAGGGGAACCACAGAGGTGAGA
AACCAGTCCTGGTTTCTCAATCACCACCAATCAATGTGGAGAGAGTCATTGTGTTAAGTTACTAGTGTA
AGTGAGGCCCTGTGATGGCCACCTGTAGCAGGGACTTTCGTTGTTCAGAGACTCTCTCTGTGTGTGTGT
GTGTGTGTGTGTGTGTGTCTGTGTGTTCGTGTGTGTTGGTCTCAGAAGTCACTTATTTAGAAAAAGGAA
GCCAACTTAATAAGAAAAGCAATAGGAAAAATTTGGTCACAG
```

Figure 2. Examples of different masking styles. Masked repeats are shown in boldface and SNPs are highlighted in red for the visualization in this Figure. (**A**) Original sequence from the human genome sequence, assembly NCBI35.1. (**B**) The same sequence masked for PCR primer design with the GenomeMasker using parameter 'target'. Asymmetrical masking is used—on the left side of target the upper strand is masked, on the right side of target the lower strand is masked. The middle part around the third SNP (shown in *italic*) is the target region which is chosen to be amplified. (**C**) The sequence masked with the RepeatMasker. (**D**) Population-specific masking of SNPs. The original nucleotides in the genome sequence have been substituted with a population-specific (lower-case) nucleotides using HapMap frequency information.

A novel way of masking or changing a sequence by population-specific allele frequency is shown in Figure 2D. Here you can see that compared to the original sequence, 2 nt have been replaced ('C' to 't' and 'G' to 'a') because these are the major alleles in Japanese dataset. The major allele frequencies of the marked SNPs in Japanese dataset are 1.00 and 0.93, respectively. Both alleles that have been replaced are the major alleles in all other HapMap populations as well. The other SNPs within this sequence remain unchanged because their major allele is already present in the Golden Path sequence or because HapMap database does not contain information about them.

PERFORMANCE

Major steps in the algorithm are the localization of the sequence by a homology search, the localization of SNPs on the sequence and finding repeats by RepeatMasker or GenomeMasker. RepeatMasker is generally too slow for real-time masking. Therefore, large genomes are typically pre-masked with RepeatMasker. If the user requests RepeatMasker-masked sequence then the corresponding region is retrieved from the database. We have taken a similar approach by downloading RepeatMasker-masked genome sequences from ENSEMBL database and installed them in a local database. Fortunately, GenomeMasker program is several orders of magnitude faster than RepeatMasker and can thus be executed each time user submits a new masking job. This allows executing it with slightly different options each time, giving the user more flexibility in masking.

Queries of dbSNP and HapMap databases and masking with GenomeMasker take only seconds to complete even with large input sizes. The main limiting factor is MEGABLAST execution time for sequences uploaded by the user. Therefore the sequence region masked by our web server is currently limited to a maximum length of 100 000 bp and single sequence per file, which should take no more than 2.5 min (Table 2). Please contact the authors for masking larger sequences or multiple sequence regions.

Table 2. The performance of SNPmasker for different tasks

Job	1 kb	10 kb	100 kb
Sequence from FASTA file			
No repeat-masking, SNPs masked with 'N'	32 s	35 s	142 s
No repeat-masking, SNPs masked using HapMap allele frequency	32 s	35 s	146 s
GenomeMasker, SNPs masked with 'N'	38 s	40 s	148 s
RepeatMasker, SNPs masked with 'N'	32 s	35 s	142 s
Sequence defined by chromosomal coordinates			
No repeat-masking, SNPs masked with 'N'	14 s	14 s	14 s
No repeat-masking, SNPs masked using HapMap allele frequency	14 s	14 s	17 s
GenomeMasker, SNPs masked with 'N'	15 s	15 s	18 s
RepeatMasker, SNPs masked with 'N'	14 s	14 s	14 s

ACKNOWLEDGEMENTS

This work was supported by the Estonian Ministry of Education and Research grant 0182649s04, grant 6041 from Estonian Science Foundation and grant EU19730 from Enterprise Estonia. The authors thank Tõnis Org, Elin Lõhmussaar and Neeme Tõnisson for a critical reading of the manuscript, and Katre Palm and Signe Sumerik for a valuable help with English grammar. Funding to pay the Open Access publication charges for this article was provided by the Estonian Ministry of Education and Research grant 0182649s04.

Conflict of interest statement. None declared.

REFERENCES

1. Syvanen,A.C. (2005) Toward genome-wide SNP genotyping. *Nature Genet*., **37**, S5–S10.
2. Koboldt,D.C., Miller,R.D. and Kwok,P.Y. (2006) Distribution of human SNPs and its effect on high-throughput genotyping. *Hum. Mutat*., **27**, 249–254.
3. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.
4. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*., **29**, 308–311.
5. Andreson,R., Reppo,E., Kaplinski,L. and Remm,M. (2006) GenomeMasker package for designing unique genomic PCR primers. *BMC Bioinformatics*, **7**, 172.
6. Birney,E., Andrews,D., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cox,T., Cunningham,F., Curwen,V., Cutts,T. *et al*. (2006) Ensembl 2006. *Nucleic Acids Res*., **34**, D556–D561.
7. Altshuler,D., Brooks,L.D., Chakravarti,A., Collins,F.S., Daly,M.J. and Donnelly,P. (2005) A haplotype map of the human genome. *Nature*, **437**, 1299–1320.
8. Zhang,Z., Schwartz,S., Wagner,L. and Miller,W. (2000) A greedy algorithm for aligning DNA sequences. *J. Comput. Biol*., **7**, 203–214.