W660–W664 *Nucleic Acids Research, 2006, Vol. 34, Web Server issue*
doi:10.1093/nar/gkl168

MutScreener: primer design tool for PCR-direct sequencing

Fengxia Yao, Ruifang Zhang, Zanhua Zhu, Kun Xia and Chunyu Liu[1],*

National Laboratory of Medical Genetics of China, Central South University, Changsha, Hunan,
P.R. China and [1]Department of Psychiatry, University of Chicago, Chicago, IL, USA

Received February 14, 2006; Revised and Accepted March 20, 2006

ABSTRACT

In searching for susceptibility genes, both positional cloning and candidate gene strategies have been helpful. Mutation screening is one of the many technologies that have been implemented in order to identify mutations or polymorphisms in candidate genes or genomic regions. Since human genome sequence is available, PCR-direct sequencing is one of the major methods for mutation screening or resequencing. Unfortunately, assay design can be laborious if multiple genes or large regions need to be investigated. To solve this conundrum a web-based application, MutScreener, has been developed. MutScreener assists in the analysis of human gene structure and design of PCR/sequencing primer. This application supports batch assay design based on either existing public gene annotation or custom gene annotation. The optional universal tagged primers can support high throughput resequencing processes. MutScreener is available for public use at http://bioinfo.bsd.uchicago.edu/MutScreener.html.

INTRODUCTION

The near-completed human genome sequence facilitates in the identification of disease genes. Searching for disease gene can start with either candidate gene approach or positional cloning approach. Both approaches rely on mutation screening to identify causal mutations or polymorphic markers. There are many different methods to screen for mutations. Methods include single strand conformational polymorphism (SSCP) (1), denaturing gradient gel electrophoresis (DGGE) (2), denaturing high performance liquid chromatography (DHPLC) (3), chemical cleavage (4) and chip-based hybridization sequencing (5). Although these 'scanning' methods are useful, they each either lack the ability to detect all variants in a target region, or can be expensive when setting up for each individual fragment. The chip-based methods require expensive setup and have variable success rates at the detection of heterozygosity. They also are not able to handle repeat sequence regions. Currently, many investigators consider the Sanger-based method best when performing genome sequencing although new cost efficient technologies might evolve soon. Therefore, PCR-direct sequencing is still one of the most widely used methods. Several steps are involved in the PCR-direct sequencing assay design: (i) analyzing the target gene's exon/intron structure in genomic DNA (gDNA), (ii) extracting target DNA sequences and (iii) designing PCR primers and sequencing primers for each PCR fragment. Appropriate assay design is essential. It will be time-consuming if many exons or large genomic regions are to be studied.

There are several web-based applications or services that perform PCR-sequencing assay design, including Variant-SEQr (http://www.appliedbiosystems.com), ExonPrimer (http://ihg.gsf.de/ihg/ExonPrimer.html), Genomic Primers (http://www2.eur.nl/fgg/kgen/primer/Genomic_Primers.html), ELXR (http://mutation.swmed.edu/ex-lax/) (6) and ampExon (http://bioinformatics.well.ox.ac.uk/mst-bin/ampexon.pl). However, the existing tools are not comprehensive for numerous reasons. Most of these web-based applications rely on public gene annotation. They do not support custom input and output data, therefore, they cannot be used to design an assay to screen one novel gene identified by the investigators. Many of the applications only handle one gene at a time, and do not support batch design for multiple genes. They are also unable to design assay to screen promoters, intronic regions or continuous genomic regions. Most of these applications do not design overlap amplicons to cover large exons, nor are they able to design sequencing primers including universal tagged primers.

MutScreener, however is a novel web-based application, that is able to execute all of the above needs as well as automate the assay design procedures. MutScreener annotates genes' structure and extracts sufficient sequences for each exon/promoter, then designs PCR and sequencing primers.

*To whom correspondence should be addressed. Tel: 773 834 3604; Fax: 773 834 2970; Email: cliu@yoda.bsd.uchicago.edu

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

SYSTEM

MutScreener provides a number of Perl coded CGI-driven web interfaces. Several Bioperl (7) modules are used to retrieve sequence data from the NCBI GenBank (http://www.ncbi.nlm.nih.gov/). Primer3 (8) is used to perform PCR primer design. In order to accommodate different types of input data of cDNA and gDNA, MutScreener has four different tracks (Tracks 1 to 4) that annotate gene structure and prepare resequencing targets. These tracks use BLAT (9), DNannotator (10) and UCSC (University of California, Santa Cruz) Genome Browser data (11) to analyze the exons, introns, promoters and to extract DNA sequence data. MutScreener then designs the appropriate PCR and sequencing primers. The workflow of MutScreener is illustrated in Figure 1.

GENE STRUCTURE ANALYSIS AND SEQUENCE PREPARATION

The four tracks of MutScreener use different methods to analyze the gene structure according to the input data as described below and prepare target sequence. Subsequently, all tracks use the same algorithm to design PCR and sequencing primers.

Track 1: inputs un-annotated genes and public genome sequences. The input data for Track 1 is a list of gene cDNA IDs (NCBI accession numbers or GI numbers) or cDNA sequences in FASTA format. When using the 'ID list' as the input, MutScreener will retrieve cDNA sequences from GenBank using Bioperl modules. Users should also specify on which chromosome the genes are located. MutScreener uses the BLAT program to align the cDNA sequences against the chromosomal gDNA sequence (UCSC Genome Browser's sequence data). With the BLAT results, MutScreener obtains the exon locations in the gDNA sequences and extracts the exon 5′ and 3′ flanking sequences in FASTA format (using adjustable parameters for length of flanking intronic sequence there is a default value of 200 bp). We compared a number of programs including Sim4 (12), est_genome (13), Spidey (14) and BLAT, and determined that BLAT usually performs better on gene intron/exon annotation (data not shown). As a result, we selected BLAT as our annotation engine. BLAT-based gene annotation parameters, which set for filtering the homologous sequences, can be set in the web interface.

Track 2: inputs genes annotated by DNannotator. The input data for Track 2 is the gene structure feature data file created by DNannotator (10), which performs batch gene annotation in human DNA. The gene structure (feature) data file is a tab-delimited text including gDNA sequence identifiers, feature names, feature types (exon or promoter), the orientation of the features, physical start and end positions and others. Users may upload this file with the corresponding gDNA sequence data in FASTA format. MutScreener then extracts exons and promoters with flanking sequences.

Track 3: inputs public annotated sequence data from UCSC Genome Browser. The input data for Track 3 is the data available from the UCSC Genome Browser's function of 'gene structure analysis and sequence extraction'. A detailed description of this input file can be found at http://bioinfo.bsd.uchicago.edu/UCSC_sample.html. The Genome Browser function supplies exon and putative promoter sequences. The limitation of this function is that only 'known genes' in the UCSC Genome Browser can be studied.

Track 4: inputs continuous gDNA sequences. The input data for Track 4 is gDNA sequences. This track is designed specifically for resequencing needs in a continuous gDNA fragment for the identification of all polymorphisms in the region regardless of the gene structure.

PRIMER DESIGN

Amplicon

With the sequence data prepared as described above, MutScreener then identifies amplicons that are PCR products for each sequencing target. In exon-centered gene resequencing each amplicon includes four elements: (i) Exon or promoter region. (ii) Splicing site region (default 30 bp). Variations in splicing sites are increasingly recognized as causes of the diseases (15). Statistics from the September 2005 Human Gene Mutation Database (http://www.hgmd.cf.ac.uk/hgmd0.html) have collected 9.4% in all the mutation entries as the splicing mutation type. (iii) The low quality base call region (default 30 bp). Because the beginning of a sequencing spectrum usually gives low quality data, this buffer region is used to ensure high quality sequencing data. (iv) PCR primer design region (default 70 bp). This region is defined in order to confine the PCR primers to the smallest possible region around the target sequence so that only crucial regions are studied. All the above parameters can be adjusted in the web interface.

When one target region is too large to be covered by a single PCR amplicon, MutScreener will use multiple amplicons to make a tiling path to cover the oversized region. Each amplicon fragment will have a minimum 120 bp overlap with the next fragment to ensure that the overlapped sequences are high quality sequencing data. The size of the PCR product and the target region decide the number of amplicons. Figure 2 illustrates the four elements of an amplicon in the PCR primer design.

Primer design parameters

After the amplicon range is determined, MutScreener will design the PCR and sequencing primers. The parameters for PCR primer design are identical to those of the Primer3 program (http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi). MutScreener has three repeat libraries (human, rodent and simple repeat) to filter the primers. Users can either select one of the libraries, or select 'none' to disable the filter. If MutScreener fails to pick up PCR primers using the users' parameters, it will re-design primers according to Primer3's default parameters, assuming that the users' condition is more stringent.

Because of the sequencing read length limits the parameter of 'quality sequencing size' (default 500 bp) needs to be set. If the amplicon size exceeds this parameter, MutScreener will design multiple sequencing primers so multiple sequencing fragments can make up a tiling path to cover the whole amplicon.

Universal tags

Universal tags are able to substitute the sequencing primers. Universal tagged primers have the benefit of easy management of a high throughput resequencing project. MutScreener

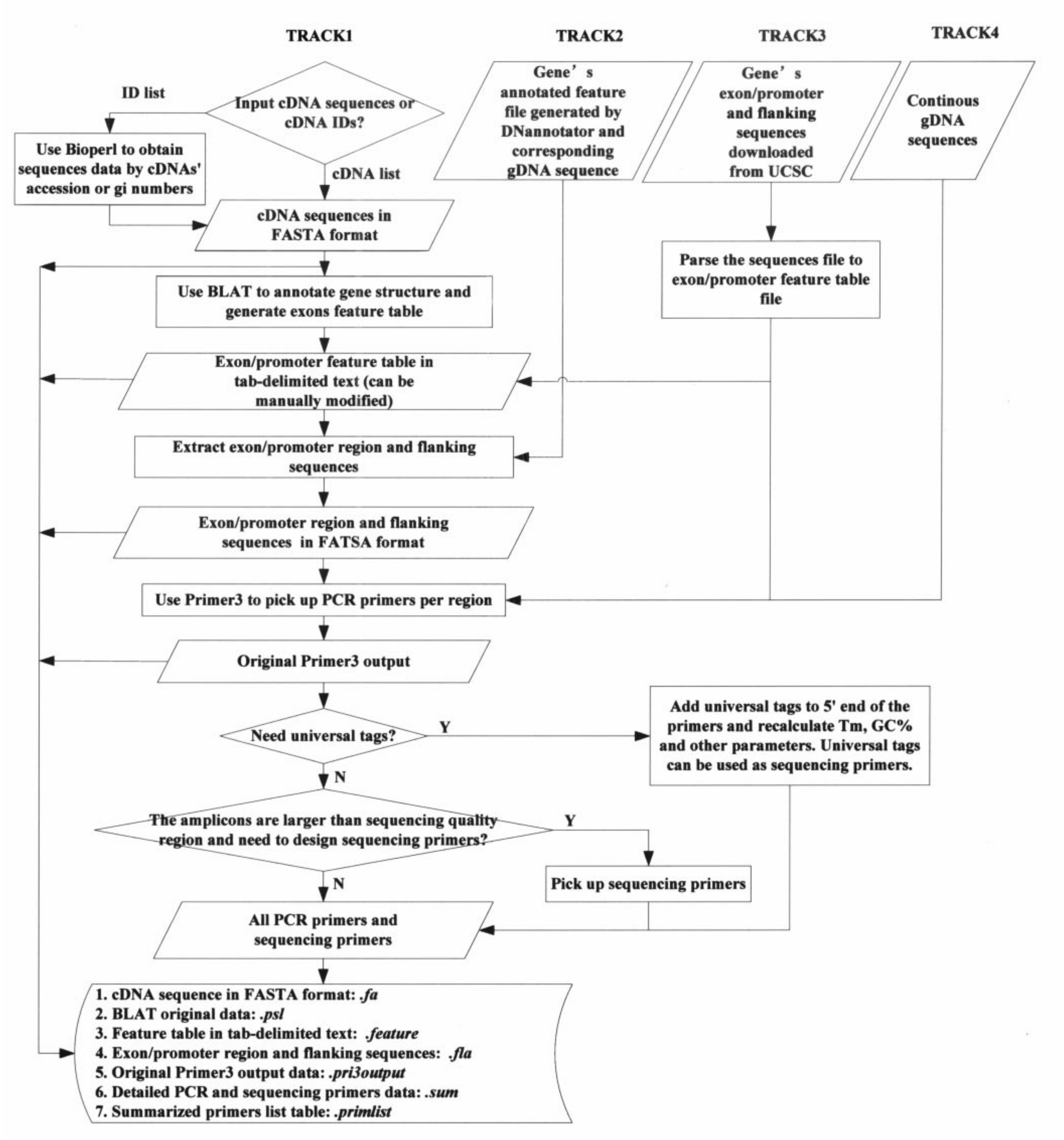
TRACK1
TRACK2
TRACK3
TRACK4
ID list
Input cDNA sequences or cDNA IDs?
cDNA list
Use Bioperl to obtain sequences data by cDNAs' accession or gi numbers
cDNA sequences in FASTA format
Gene' s annotated feature file generated by DNannotator and corresponding gDNA sequence
Gene' s exon/promoter and flanking sequences downloaded from UCSC
Continous gDNA sequences
Parse the sequences file to exon/promoter feature table file
Use BLAT to annotate gene structure and generate exons feature table
Exon/promoter feature table in tab-delimited text (can be manually modified)
Extract exon/promoter region and flanking sequences
Exon/promoter region and flanking sequences in FATSA format
Use Primer3 to pick up PCR primers per region
Original Primer3 output
Need universal tags?
Y
Add universal tags to 5' end of the primers and recalculate Tm, GC% and other parameters. Universal tags can be used as sequencing primers.
N
The amplicons are larger than sequencing quality region and need to design sequencing primers?
Y
Pick up sequencing primers
N
All PCR primers and sequencing primers
1. cDNA sequence in FASTA format: .fa
2. BLAT original data: .psl
3. Feature table in tab-delimited text: .feature
4. Exon/promoter region and flanking sequences: .fla
5. Original Primer3 output data: .pri3output
6. Detailed PCR and sequencing primers data: .sum
7. Summarized primers list table: .primlist

Figure 1. The workflow of MutScreener. Rectangles and rhombi show the processes that the program performs. Parallelograms show the data, either supplied by users or generated by MutScreener.

supplies the option of adding a universal tag to the 5′ end of either forward, reverse or both primers. MutScreener provides the choices of M13 forward and reverse primers as tags. M13 primers have no strong homologs in the human genome. Users can also use their own preferred sequences as tags.

OUTPUTS

The four tracks of MutScreener produce different outputs. Detailed outputs are summarized in Table 1. The results include cDNA sequences in FASTA format, chromosomal gDNA sequence (only for Track 1), original BLAT and

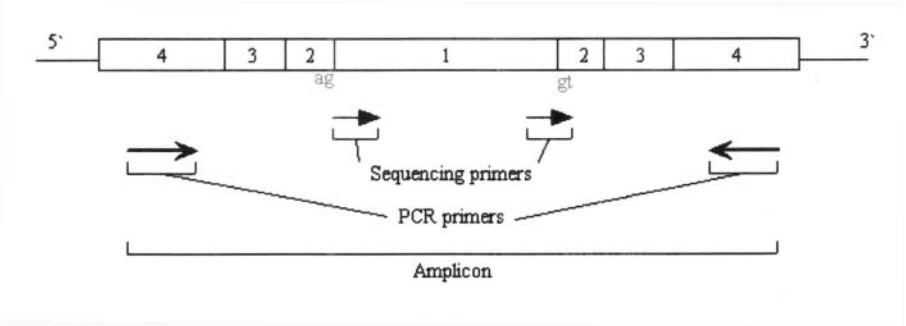

Figure 2. The elements of an amplicon in the PCR primer design. The box with a number 1 illustrates the exon or promoter region. The boxes with a number 2 inside of them show the splicing site region. The boxes with a number 3 explain the low quality base call region. The boxes with a number 4 demonstrate the PCR primer design region.

Table 1. The outputs of MutScreener

	Track 1	Track 2	Track 3	Track 4
cDNA sequences	m1.X.pa	—	—	—
Chromosomal gDNA sequence	chr.gb.gz	—	—	—
BLAT original output	m1.X.psl	—	—	—
Tab-delimited text feature table	m1.X.feature	—	M3.X.feature	—
Exon/promoter and flanking sequences in FASTA format	m1.X.fla	m2.X.fla	—	—
Primer3 original output	m1.X.pri3output	m2.X.pri3output	m3.X.pri3output	m4.X.pri3output
Summarized primers list in tab-delimited text	m1.X.primlist	m2.X.primlist	m3.X.primlist	m4.X.primlist
Detailed primers output	m1.X.sum	m2.X.sum	m3.X.sum	m4.X.sum

Primer3 program outputs, the tab-delimited gene feature data, exon/promoter and flanking sequences in FASTA format, summarized primer list in tab-delimited format, and detailed primer data similar to Primer3 Web output format. Two tab-delimited files make the results compatible with a database or a spreadsheet.

DISCUSSION AND CONCLUSION

We have compared MutScreener with other similar tools and web-based applications, such as: VariantSEQr, ExonPrimer, Genomic Primers, ELXR and ampExon (see Supplementary Data). MutScreener provides more comprehensive and efficient functions for PCR-sequencing assay design. MutScreener offers multiple choices of inputs including cDNA IDs, cDNA sequences, DNannotator annotated feature data, UCSC annotated sequences and continuous gDNA sequences. Users can select different tracks according to their data. MutScreener examines exons, promoter regions and splicing sites. Sequencing low quality regions are also considered at the amplicon design step.

Batch screening of genes is often needed for positional cloning or systematic resequencing projects. MutScreener can now design assays for resequencing genes from one chromosome. In the future, it will be improved to analyze genes from different chromosomes.

MutScreener can also design multiple sequencing primers for large amplicons. This feature is not provided by other programs. The option of using universal tagged primers enables high throughput resequencing.

MutScreener provides gene annotation and designs PCR and DNA sequencing primers. It assists researchers in mutation screening in multiple genes or genomic regions and supports universal tagged PCR primers for high throughput resequencing.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

ACKNOWLEDGEMENTS

This work was supported by the Young Investigator Grant from NARSAD (National Alliance for Research in Schizophrenia and Affective Disorders), Brain Research Foundation (to C.L.). Chinese State '863' program, '973' program and Ministry of Education grant (to K.X.). Support from the Geraldi Norton Memorial Corporation, the Eklund Family and Anita Kaskel Roe are gratefully acknowledged. The authors also thank Elliot S. Gershon, Jiajun Shi, Heather L. Fritz and Aquila J. Beach for helpful discussion and comments. Funding to pay the Open Access publication charges for this article was provided by '973' program.

Conflict of interest statement. None declared.

REFERENCES

1. Orita,M., Iwahana,H., Kanazawa,H., Hayashi,K. and Sekiya,T. (1989) Detection of polymorphisms of human DNA by gel electrophoresis as

single-strand conformation polymorphisms. *Proc. Natl Acad. Sci. USA*., **86**, 2766–2770.

2. Myers,R.M., Fischer,S.G., Lerman,L.S. and Maniatis,T. (1985) Nearly all single base substitutions in DNA fragments joined to a GC-clamp can be detected by denaturing gradient gel electrophoresis. *Nucleic Acids Res*., **13**, 3131–3145.
3. Xiao,W. and Oefner,P.J. (2001) Denaturing high-performance liquid chromatography: a review. *Hum. Mutat*., **17**, 439–474.
4. Ellis,T.P., Humphrey,K.E., Smith,M.J. and Cotton,R.G. (1998) Chemical cleavage of mismatch: a new look at an established method. *Hum. Mutat*., **11**, 345–353.
5. Drmanac,R., Drmanac,S., Chui,G., Diaz,R., Hou,A., Jin,H., Jin,P., Kwon,S., Lacy,S., Moeur,B. *et al*. (2002) Sequencing by hybridization (SBH): advantages, achievements, and opportunities. *Adv. Biochem. Eng Biotechnol*., **77**, 75–101.
6. Schageman,J.J., Horton,C.J., Niu,S., Garner,H.R. and Pertsemlidis,A. (2004) ELXR: a resource for rapid exon-directed sequence analysis. *Genome Biol*., **5**, R36.
7. Stajich,J.E., Block,D., Boulez,K., Brenner,S.E., Chervitz,S.A., Dagdigian,C., Fuellen,G., Gilbert,J.G., Korf,I., Lapp,H. *et al*. (2002) The Bioperl toolkit: perl modules for the life sciences. *Genome Res*., **12**, 1611–1618.
8. Rozen,S. and Skaletsky,H. (2000) Primer3 on the WWW for general users and for biologist programmers. *Meth. Mol. Biol*., **132**, 365–386.
9. Kent,W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res*., **12**, 656–664.
10. Liu,C., Bonner,T.I., Nguyen,T., Lyons,J.L., Christian,S.L. and Gershon,E.S. (2003) DNannotator: annotation software tool kit for regional genomic sequences. *Nucleic Acids Res*., **31**, 3729–3735.
11. Karolchik,D., Baertsch,R., Diekhans,M., Furey,T.S., Hinrichs,A., Lu,Y.T., Roskin,K.M., Schwartz,M., Sugnet,C.W., Thomas,D.J. *et al*. (2003) The UCSC Genome Browser Database. *Nucleic Acids Res*., **31**, 51–54.
12. Florea,L., Hartzell,G., Zhang,Z., Rubin,G.M. and Miller,W. (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res*., **8**, 967–974.
13. Mott,R. (1997) EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. *Comput. Appl. Biosci*., **13**, 477–478.
14. Wheelan,S.J., Church,D.M. and Ostell,J.M. (2001) Spidey: a tool for mRNA-to-genomic alignments. *Genome Res*., **11**, 1952–1957.
15. Cooper,T.A. and Mattox,W. (1997) The regulation of splice-site selection, and its role in human disease. *Am. J. Hum. Genet*., **61**, 259–266.