Software

ELXR: a resource for rapid exon-directed sequence analysis

Jeoffrey J Schageman[*†‡¶], Christopher J Horton[*], Sijing Niu[*], Harold R Garner[*‡§¶] and Alexander Pertsemlidis[*†¶]

Addresses: [*]Eugene McDermott Center for Human Growth and Development, University of Texas Southwestern Medical Center, Harry Hines Boulevard, Dallas, TX 75390, USA. [†]Frank M. Ryburn Jr. Cardiology Center, University of Texas Southwestern Medical Center, Harry Hines Boulevard, Dallas, TX 75390, USA. [‡]Center for Biomedical Inventions, University of Texas Southwestern Medical Center, Harry Hines Boulevard, Dallas, TX 75390, USA. [§]Department of Biochemistry, University of Texas Southwestern Medical Center, Harry Hines Boulevard, Dallas, TX 75390, USA. [¶]Department of Internal Medicine, University of Texas Southwestern Medical Center, Harry Hines Boulevard, Dallas, TX 75390, USA.

Correspondence: Jeoffrey J Schageman. E-mail: jeff.schageman@utsouthwestern.edu

Published: 28 April 2004

Genome ***Biology*** 2004, **5:**R36

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2004/5/5/R36

Received: 8 January 2004
Revised: 13 February 2004
Accepted: 14 April 2004

Abstract

ELXR (Exon Locator and Extractor for Resequencing) streamlines the process of determining exon/intron boundaries and designing PCR and sequencing primers for high-throughput resequencing of exons. We have pre-computed ELXR primer sets for all exons identified from the human, mouse, and rat mRNA reference sequence (RefSeq) public databases curated by the National Center for Biotechnology Information. The resulting exon-flanking PCR primer pairs have been compiled into a system called ELXRdb, which may be searched by keyword, gene name or RefSeq accession number.

Rationale

With the vast amount of human genome sequence now publicly available [1], many researchers are mining these data to detect genetic variation with the hope of better understanding human disease. Most genetic variations are in the form of single-nucleotide polymorphisms (SNPs) and insertions/deletions. Of these, nonsynonymous SNPs are believed to be most frequently associated with disease phenotypes [2] as they may contribute to pathological amino-acid substitutions or nonsense mutations in the protein product. Gene resequencing at the exon level has become the standard method of detecting coding SNPs in human populations [3,4].

The process of resequencing individual genes is usually performed at the exon level in the following manner. First, a messenger RNA (mRNA) sequence from a gene of interest is obtained from a sequence database, such as those available from the National Center for Biotechnology Information (NCBI) [5]. Next, the corresponding genomic sequence must be identified and retrieved. Once both genomic and mRNA sequences are obtained, exon/intron structure is determined via sequence alignment of the two and/or tools for splice-site prediction. Polymerase chain reaction (PCR) and sequencing primer pairs are then designed such that they flank each exon. Following PCR, the resulting amplicons containing individual exons are sequenced and compared to the corresponding sequences from other individuals in a population to detect sequence variation. While often taken for granted, these initial design steps can be a significant informatics hurdle, and if done improperly, can result in a waste of laboratory resources.

To address these issues, we have developed an integrated informatics tool, called ELXR, to accomplish the same goals in a fraction of the time. ELXR is a web-based computer program (CGI) that incorporates publicly available bioinformatics tools into one sequence-analysis resource to completely automate PCR/sequencing primer-pair design for

resequencing exons and their flanking regions. Results reported to the user include annotated genomic sequence containing the query mRNA, start and stop codon locations, and a per-exon display of primer pairs with their respective locations and properties. Also located at the ELXR website is a queryable database, ELXRdb, consisting of pre-computed ELXR PCR/sequencing primer pairs for all human (15,365 as of June 2002), mouse (8,583 as of June 2002), and rat (4,552 as of July 2003) entries from the NCBI-curated RefSeq project [6]. ELXR and ELXRdb, along with documentation, are freely available web services located at [7] and [8], respectively.

Computational and testing resources

ELXR is being used to design PCR and sequencing primers for the resequencing of 750 candidate genes implicated in cardiovascular disease as part of the high-throughput SNP-sequencing pipeline in the NHLBI Program for Genomic Applications at University of Texas Southwestern University (UTSW-PGA) [9]. ELXR was tested and validated using a randomly chosen 14-gene subset, which collectively consisted of 154 putative exons determined by complementary DNA (cDNA) to genomic sequence alignments. PCR was carried out using the Advantage-GC 2 PCR Kit (Clontech). DNA sequencing was carried out using the ABI PRISM BigDye Terminators v3.1 Cycle Sequencing Kit, and sequence data was collected on a 3730 DNA Analyzer, both of which are supplied by Applied Biosystems.

The source code for ELXR was written using the Perl scripting language and utilizes a general CGI module as well as various BioPerl modules [10], including Seq and SeqIO for sequence input and output processing, as well as the Tools::Sim4 and Tools::Blast modules for Sim4 and BLAST output parsing. Perl is available for all major operating systems and documentation and download information for BioPerl is available from [11]. Graphical representation of aligned exons was developed using Java.

Sequence processing and algorithm

Input for ELXR may be a RefSeq (mRNA) accession number or a FASTA-formatted nucleotide sequence. Users may specify parameters related to primer picking options, species and output format. The automatic design of exon-flanking primers is accomplished in several steps (Figure 1), beginning with input processing. If the user input is a RefSeq accession number, the genomic contig identifier may be extracted from the NCBI LocusLink [6] annotation. If a FASTA-formatted sequence is used as input or the cognate LocusLink entry does not exist, a BLAST [12,13] search is performed to align the input sequence to an NCBI-curated genomic contig. These genomic sequence resources are available for download via FTP at [14].

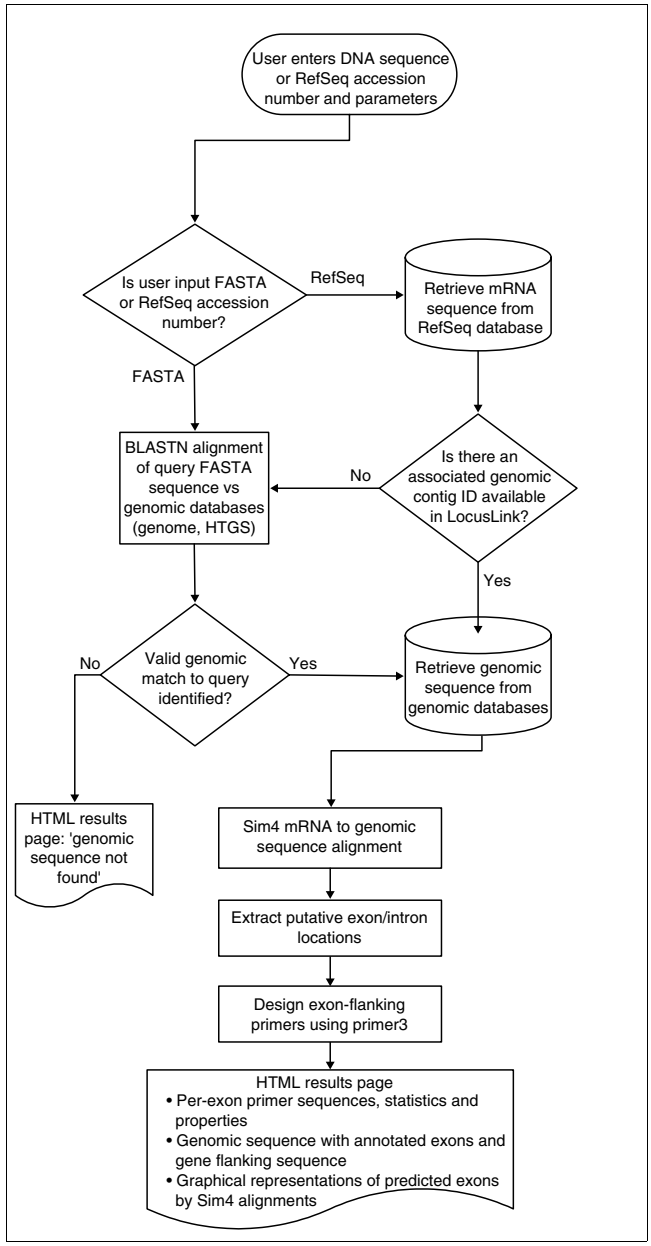

Figure 1
ELXR sequence processing flow for each mRNA/EST sequence query. HTGS, high-throughput genomic sequence.

One issue that had to be resolved involved BLAST alignment specificity when determining the correct parent genomic sequence. For some mRNA queries, if only the top-scoring BLAST result is chosen, erroneous, high-scoring matches can result from alignments to pseudogenes or genomic duplications. For this reason, BLAST 'hits' are filtered by local alignment score as well as by the fraction of identical nucleotide matches. As we expect near perfect alignments (to high-quality genomic sequence), the default fraction identical threshold is set to 0.96. This and other BLAST filtering

parameters may be tuned to user specification in the ELXR web form.

Occasionally, because of incompleteness of the curated genome databases, a genomic contig cannot be identified. In these cases, a secondary BLAST search using the NCBI high-throughput genomic sequence contigs [15] is used to ensure that a comprehensive search of all NCBI genomic sequence resources has been performed.

With mRNA and genomic sequences retrieved, putative exon locations and splice sites are identified using Sim4 [16]. Sim4 rapidly aligns cDNA sequence to genomic sequence and reports exon/intron boundaries by sequence position. Users may add higher sensitivity to small external exons as well as the removal of input sequence poly(A) tails using checkboxes on the ELXR web form. These options correspond to the Sim4 'N' and 'P' options, respectively.

Primer3 [17] is used to design PCR primers from sequences that flank, and are in close proximity to, putative exonic sequences determined by Sim4 alignments. The ELXR user interface allows the user to change many of the parameters used by Primer3, such as primer-annealing temperature, length, GC content and maximum self-complementarity. In addition, each designed primer is screened against a repetitive element database to reduce nonspecific priming in PCR reactions where whole genomic DNA serves as a template.

In many cases, exons are too large to be PCR amplified and sequenced as a single product, mostly owing to sequence quality read-length limitations imposed by current high-throughput fluorescent sequencing technologies. To address this issue, aligned exons larger than a user-specified optimum product size are automatically subdivided into segments of that optimal size where the adjacent segments overlap by 50 base pairs. Primers are designed for PCR amplification of each overlapping segment. Sequencing of these amplicons forms an efficient tiling path across a large exon.

To avoid the low-quality base calls that are typically found near the beginning and end of each sequence, we include a buffer region between the primer-annealing location and the point at which high-quality sequence is essential for clearly detecting sequence variation. The size of this buffer region is under user control and effectively increases the 50 bp product overlap that applies to exons with multiple products, and also adds to the user-defined exon flanking sequence for non-overlapping PCR products.

Output format

Results from ELXR include a set of hyperlinks consisting of a Primer3 primer summary for each aligned exon, Sim4 genomic alignments, a primer-pair summary for each aligned exon, an mRNA coverage assessment, and a FASTA-formatted nucleotide sequence which encompasses the query mRNA sequence. The coverage assessment conveys how much of the query mRNA sequence was found by an alignment to the parent genomic sequence. If there are more than 10 unaligned nucleotides at the 5' or 3' ends of the query mRNA, this unaligned sequence is also reported to allow the user to run ELXR a second time in the hope of aligning it to another genomic contig. Aligned exons as well as introns in this segment are indicated using annotation similar to that of BLAT-derived [18] output included in the University of California Santa Cruz (UCSC) Genome Browser [19] in which aligned exon sequences are presented in upper-case letters and remaining sequence segments are in lower-case letters. In cases where a RefSeq accession number is used as input, ELXR highlights start and stop codons to indicate the location of a protein-coding region in the genomic sequence and reports the associated exon numbers that contain these codons. This information is useful in situations where users want to select coding exons exclusively for resequencing. A graphical representation of Sim4 aligned gene structure is also provided at the top of each results page (Figure 2). Each segment representing an exon is hyperlinked to the Primer3 results page for that exon. Lastly, all resulting primer designs are compiled into a single text file that is hyperlinked to allow for easy evaluation and custom primer ordering.

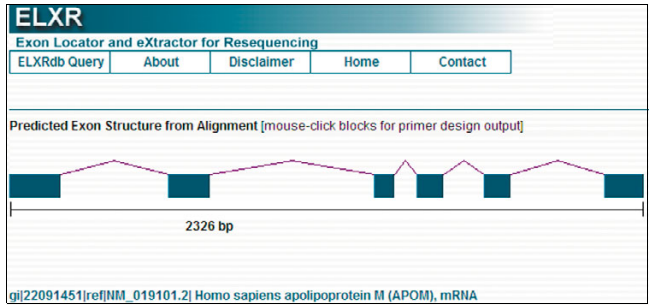

Figure 2
Sample output from ELXR. Graphical depiction of the human apolipoprotein M gene structure derived from ELXR's Sim4 component.

Validating the method

To provide ample validation of ELXR as an automated method, a comparison with manual exon processing and primer design was carried out. Manual exon processing was performed by lab technicians using online tools that include the UCSC Genome Browser, NCBI BLAST and Primer3 in conjunction with numerous cut-and-paste operations.

The 154 exons from the test set of 14 UTSW-PGA genes were chosen for resequencing in a cohort of 24 individuals and 164 ELXR primer pairs were generated and ordered. The discrepancy between the number of exons and the number of primer pairs ordered reflects the fact that some larger (mostly terminal, 3' UTR-containing) exons were covered by multiple

overlapping PCR products. Successfully PCR amplified and bidirectionally sequenced exons were tallied and compared to analogous results from a previously analyzed set of 864 manually processed exons (891 PCR products) also from the UTSW-PGA. The Primer3 parameter for PCR product size range in Primer3 was set to 350 to 450 bp with 400 being the optimum size, as most exons can be entirely amplified in this range. In these comparisons, a successful test was defined as a resultant single exon-containing product that aligned appropriately to control sequences for a given sequence alignment. The basis for determining success or failure is the combination of quality measures taken at both PCR and sequencing steps of the exon resequencing process. All synthesized primers as well as PCR products are verified for specificity and size by agarose gel electrophoresis. We consider successful those reads for which PCR products have been verified and where the resulting sequences are properly assembled into a sequence alignment using the Phred/Phrap/Consed software package [20-22]. Phrap uses a window-based quality method for aligning high-quality sequences. Parameters for this method were set to program defaults. All initial primer designs for both methods were performed using default ELXRdb parameters. Occasionally, these parameters were modified when primer designs would fail because of sequence-specific issues such as very high GC content or low-complexity regions. This subsequent parameter 'tweaking' usually corrected all primer design failures. All post-primer design procedures such as PCR amplification and optimization, sequencing, and sequence alignment evaluation were carried out in identical fashion for both methods.

Evaluations of these datasets based on comparisons of processing time and success-to-failure percentage revealed that comparable results were obtained more than eight times faster using ELXR (Table 1). PCR or sequencing-failure frequency does not appear to be related to whether or not ELXR was used. For example, PCR failures due to nonspecific priming or no product at all are approximately the same, varying by only 1-2%. The above comparison should not be interpreted as a test of primer design, as the manual and automated methods rely on the same primer design algorithm (Primer3). There is a difference between the two approaches, however, in that the manual method does not typically rely on a standard set of parameters for primer design, whereas the automatic method imposes such a constraint. The fact that the two methods yielded comparable results indicates that there is little or no penalty in trading some flexibility in parameter selection for an increase in speed. In the light of these observations, the UTSW-PGA group has subsequently converted from manual processing to the automated ELXR method.

Table 1

Time and performance comparison

Method	N	PT	PS	APT/gene
Manual	864	891	84.5%	1.50 h
ELXR	154	164	89.3%	0.18 h

N indicates the number of exons and PT the number of associated primer pairs chosen and tested for PCR and sequencing for each primer-picking method. PS indicates the percentage of PCR products that resulted in high-quality sequence products and subsequent SNP detection. APT/gene indicates the average processing time per gene using each method. This table is not intended to describe the performance of Primer3 (which is used in both methods), but only to illustrate that whereas success was comparable with both methods, exon identification and primer-pair design was more than eight times faster using ELXR compared to nonautomated processing methods.

Database generation and statistics

The accompanying database, ELXRdb, consists of pre-computed ELXR runs for all human, mouse, and rat mRNA sequences in the NCBI-curated RefSeq project. Creation of this database required that mRNA entries be processed in a single batch, and ELXR and Primer3 parameters had to be standardized. These parameters are available from the ELXR web site in the 'About' section.

In addition to the experimental validation described above, we attempted to obtain a more global validation by comparing some of the statistics generated from processing all available curated RefSeq mRNA entries to those reported in the literature. Generation of the ELXRdb enabled us to exploit these aggregate statistics to survey the genome on the basis of individual ELXR results (Table 2). RefSeq entries processed were those that have genomic contig accession numbers associated with them via LocusLink annotation, greater than 95% mRNA sequence coverage by alignment to genomic sequence, and result in little or no erroneous, small exon alignments from Sim4. Occasionally, Sim4 has trouble aligning small initial and terminal exons, leading to distorted measures of intron size and genomic extent. Therefore, Sim4 was run using the N and P flags. Sim4 also has a basic 'exon core' determination parameter (maximal segment pair threshold), K, which is normally set to 16 for aligning to genomic sequences that are a few kilobase pairs (kb) in length. Typically, NCBI genomic reference contigs identified with ELXR are megabase pairs in length, thus increasing the probability that the Sim4 alignment could be erroneous, especially for smaller initial and terminal exons. It is recommended in Sim4 documentation that K be increased as genomic length increases beyond a few kb. In an effort to increase exon specificity, ELXR dynamically increases K linearly with the genomic contig length. In addition, we set the minimum exon size to 8 bp. This number is somewhat arbitrary, but reduced erroneously large intron sizes in 28 out of 30 gene tests.

Examination of the resulting statistics revealed that with the exception of genomic extent (length in base pairs from the beginning of the first exon to the end of the last exon) and

Table 2

Statistical assessment of ELXRdb for mouse and human compared with analogous statistics initially reported by the public human genome sequencing project

	Mouse	Human	HGC [1]
Dataset			
Total RefSeq mRNAs processed	5,110	12,260	1,804
Number of primer pairs successfully designed	59,874	179,065	
Number of exons processed	44,715	123,757	
Single-exon genes	348	853	
Averages			
Intron size	3,644 bp	4,573 bp	3,365 bp
Internal intron size	3,057 bp	3,876 bp	
Exon size	218 bp	244 bp	
Exon number	8.7	10	8.8
Initial exon size	204 bp	240 bp	
Terminal exon size	633 bp	874 bp	
Internal exon size	143 bp	151 bp	145 bp
Coding-sequence nucleotides	1,315 bp	1,558 bp	1,340 bp
Coding sequence amino acids	438	520	447
Genomic extent	38 kb	62 kb	27 kb

Exon statistics were compiled from Sim4. Coding sequences are as defined in GenBank annotation. Empty fields in the HGC column indicate that there were no values for these measures provided in [1].

mean intron size, statistical measures were comparable to those originally reported in the literature. This discrepancy was not completely unexpected, as some Sim4 alignments resulted from comparing mRNA to genomic contigs that consist of both finished and draft sequence. The alignments to draft sequence may yield artificially large intron sizes (and thus genomic extents) due to the inclusion of sequence gaps.

The ELXR program (along with other methods) has an obvious limitation when genomic sequence is not available for a given mRNA. Nevertheless, with the NCBI human finished sequence nearly complete, we found that 93% of all human RefSeq cDNA entries aligned to genomic sequence with coverage of 95% or higher.

The interface to ELXRdb is designed such that a user can not only retrieve pre-computed ELXR runs corresponding to RefSeq mRNAs, but also retrieve particular sequence segments resulting from individual ELXR analyses (Figure 3). These include 5' or 3' untranslated sequences plus flanking regions, and exonic or intronic sequences separated into FASTA-formatted sequences. This functionality is convenient for use in other types of analyses such as scanning multiple 5' upstream regions for conserved DNA motifs in potential promoters.

Alternative uses and future enhancements

ELXR and ELXRdb, as described above, can greatly increase productivity for any high-throughput SNP sequencing project or individual investigation by automating most of the steps before PCR amplification and sequencing of exon-containing amplicons.

ELXR is also suited for other applications that involve primer design, determination of exon/intron boundaries and SNP discovery. These applications include the design of real-time PCR (RT-PCR) primers for mRNA quantification [20], the examination of potential exon/intron boundaries to assist in the evaluation of gene splice variants, the design of PCR primers to amplify CpG island sequences for methylation studies [21], and the resequencing of promoters and evolutionarily conserved noncoding regions of the genome in the search for SNPs associated with disease [22].

To serve these ends, we have made some of the ELXR parameters changeable to extend functionality beyond primer pair design for exon resequencing specifically. One such parameter controls the size of the parent genomic segment that surrounds each aligned mRNA. This segment can be increased to a maximum of 10,000 bases flanking the 3'- and

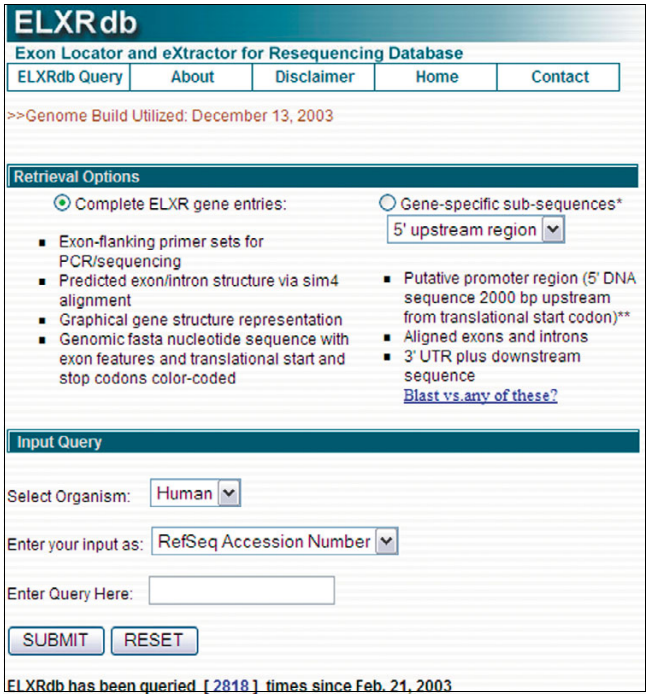

Figure 3
The ELXRdb entry retrieval interface.

5'-most exons. This is useful in studies that involve transcriptional binding site analysis or searching for conserved DNA motifs in promoter regions of orthologous genes. In addition, as Sim4-predicted gene structure is a component of the ELXR output, several individual FASTA-formatted sequence queries may be run using ELXR for detecting splice variants in mRNA sequences or expressed sequence tags (ESTs). Aligned exons may be compared at the sequence level by viewing the ELXR-annotated genomic segment. In addition, other FASTA-formatted sequences features such as introns or promoter regions can also be used as input. In these cases, Sim4 aligns the feature to a genomic contig, and then Primer3 primers are used to design overlapping PCR products tiling across the entire input sequence.

As publicly available genome and mRNA resources become more complete, other organisms will be added to the ELXR system and ELXRdb will be updated accordingly. One area where further work is warranted is the instance when high GC content or low-complexity sequence prohibits optimal primer design for a given sequence. Future versions of ELXR will include a method to automatically reanalyze sequences that fail primer design by relaxing the primer-picking parameters in Primer3. This can be accomplished now in ELXR, but not for individual exons. We also hope to include a mechanism by which users can annotate primer designs, providing feedback on success or failure and increasing the value of the resource as a whole.

Acknowledgements

This work was supported with funding from a Program for Genomic Applications (grant number 5U01HL6688002) from the National Heart Lung and Blood Institute and the National Cancer Institute (grant number R33CA81656). The authors wish to thank H. Hobbs, W. Crider, J.W. Fondon and B. Munjuluri for valuable comments and contributions and the McDermott Sequencing Core Facility for assistance with validation.

References

1. The International Human Genome Consortium: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409:**860-921.
2. Sunyaev S, Hanke J, Aydin A, Wirkner U, Zastrow I, Reich J, Bork P: **Prediction of nonsynonymous single nucleotide polymorphisms in human disease-associated genes.** *J Mol Med* 1999, **77:**754-760.
3. Ma X, Jin Q, Forsti A, Hemminki K, Kumar R: **Single nucleotide polymorphism analyses of the human proliferating cell nuclear antigen (pCNA) and flap endonuclease (FEN1) genes.** *Int J Cancer* 2000, **88:**938-942.
4. Ohnishi Y, Tanaka T, Yamada R, Suematsu K, Minami M, Fujii K, Hoki N, Kodama K, Nagata S, Hayashi T *et al.*: **Identification of 187 single nucleotide polymorphisms (SNPs) among 41 candidate genes for ischemic heart disease in the Japanese population.** *Hum Genet* 2000, **106:**288-292.
5. **NCBI** [http://www.ncbi.nih.gov]
6. Pruitt KD, Maglott DR: **RefSeq and LocusLink: NCBI gene-centered resources.** *Nucleic Acids Res* 2001, **29:**137-140.
7. **ELXR** [http://elxr.swmed.edu]
8. **ELXRdb** [http://elxr.swmed.edu/elxrdb_query.html]
9. **UT Southwestern Program for Genomic Applications** [http://pga.swmed.edu]
10. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JG, Korf I, Lapp H *et al.*: **The bioperl toolkit: perl modules for the life sciences.** *Genome Res* 2002, **12:**1611-1618.
11. **BioPerl** [http://bioperl.org]
12. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403-410.
13. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.
14. **NCBI Genomes FTP site** [ftp://ftp.ncbi.nih.gov/genomes]
15. **NCBI HTGS Sequence FTP** [ftp://ftp.ncbi.nih.gov/blast/db/FASTA/htgs.gz]
16. Florea L, Hartzell G, Zhang Z, Rubin GM, Miller W: **A computer program for aligning a cDNA sequence with a genomic DNA sequence.** *Genome Res* 1998, **8:**967-974.
17. Rozen S, Skaletsky H: **Primer3 on the WWW for general users and for biologist programmers.** *Methods Mol Biol* 2000, **132:**365-386.
18. Kent WJ: **BLAT - the BLAST-like alignment tool.** *Genome Res* 2002, **12:**656-664.
19. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D: **The human genome browser at UCSC.** *Genome Res* 2002, **12:**996-1006.
20. Zheng H, Yan W, Toppari J, Harkonen P: **Improved nonradioactive RT-PCR method for relative quantification of mRNA.** *Biotechniques* 2000, **28:**832-834.
21. Herman JG, Graff JR, Myohanen S, Nelkin BD, Baylin SB: **Methylation-specific PCR: a novel PCR assay for methylation status of CpG islands.** *Proc Natl Acad Sci USA* 1996, **93:**9821-9826.
22. Nobrega M, Pennacchio LA: **Comparative genomic analysis as a tool for biological discovery.** *J Physiol* 2004, **554:**31-39.