W548–W552 *Nucleic Acids Research, 2005, Vol. 33, Web Server issue*
doi:10.1093/nar/gki483

SOP^{3}v2: web-based selection of oligonucleotide primer trios for genotyping of human and mouse polymorphisms

Steven Ringquist, Christopher Pecoraro, Crystal M. S. Gilchrist, Alexis Styche, William A. Rudert, Panagiotis V. Benos[1] and Massimo Trucco*

Division of Immunogenetics, Department of Pediatrics, Rangos Research Center, Children's Hospital of Pittsburgh, University of Pittsburgh School of Medicine, 3460 Fifth Avenue, Pittsburgh, PA 15213, USA and [1]Department of Computational Biology and Bioinformatics, University of Pittsburgh School of Medicine, Pittsburgh, PA 15261, USA

Received February 14, 2005; Revised March 29, 2005; Accepted April 18, 2005

ABSTRACT

SOP^{3}v2 is a database-driven graphical web-based application for facilitating genotyping assay design. SOP^{3}v2 accepts data input in numerous forms, including gene names, reference sequence numbers and physical location. For each entry, the application presents a set of recommended forward and reverse PCR primers, along with a sequencing primer, which is optimized for sequence-based genotyping assays. SOP^{3}v2-generated oligonucleotide primer trios enable analysis of single nucleotide polymorphisms (SNPs) as well as insertion/deletion polymorphisms found in genomic DNA. The application's database was generated by warehousing information from the National Center for Biotechnology Information (NCBI) dbSNP database, genomic DNA sequences from human and mouse, and LocusLink gene attribute information. Query results can be sorted by their biological relevance, such as nonsynonymous coding changes or physical location. Human polymorphism queries may specify ethnicity, haplotype and validation status. Primers are developed using SOP^{3}v2's core algorithm for evaluating primer candidates through stability tests and are suitable for use with sequence-based genotyping methods requiring locus-specific amplification. The method has undergone laboratory validation. Of the SOP^{3}v2-designed primer trios that were tested, a majority (>80%) have successfully produced genotyping data. The application may be accessed via the web at http://imgen.ccbb.pitt.edu/sop3v2.

INTRODUCTION

The design of locus-specific primer sets for use during genetic analysis often requires combining information from multiple sources, and as a result it can be time consuming when validating assays for large numbers of polymorphisms. Data warehousing of publicly available information (e.g. genome sequence, the location of genomic polymorphisms and locus-related information), coupled with software applications for optimizing the generation of locus-specific primers, can increase the efficiency of assay development. The SOP^3 version 2 (SOP^{3}v2) software created for the development of genotyping assays was built using warehoused data from existing genome projects, utilizing the identity and location of relevant polymorphisms [i.e. single nucleotide polymorphisms (SNPs) as well as insertion/deletion events] for analysis by the application's primer design algorithm. The method automates processes such as collection of the genomic sequence and locus-associated functional information, and identification of polymorphic residues. It also takes into consideration restrictions that should be applied for sequencing applications, such as optimum length of the PCR amplification product, minimum distance between sequencing primer and polymorphic residue, and the presence of nearby polymorphisms. The present version of the application, SOP^{3}v2, allows querying of human and mouse genomic polymorphisms. The program accepts as input gene locus symbols, SNP reference sequence numbers or chromosomal physical location. For human polymorphisms, SOP^{3}v2 incorporates haplotype, ethnicity and validation attributes. The output is a list of oligonucleotide primers recommended for use in sequencing-based genotyping to evaluate the inheritance patterns of SNP markers, allowing the collection of genetic data for analysis of their association with inherited phenotypes. In the event that the user specifies multiple SNPs for genetic

*To whom correspondence should be addressed. Tel: +1 412 692 6570; Fax: +1 412 692 5809; Email: mnt@pitt.edu analysis, the application aids in the primer design process by being able to sort SNPs according to various properties. These include such properties as heterozygosity, validation status, ethnicity, the relationship between an SNP and a particular genetic structural element and whether the SNP codes for an amino acid change in the resulting protein. All of these additional refinements can, in turn, add to the usefulness of particular primer trios during genetic mapping studies (1).

Other software applications have been developed for PCR primer design, but they may not fully exploit the principal benefits when optimized sequencing assays are used for genotyping, i.e. the rapid accumulation of genotyping data when a sequencing primer is annealed proximal to the polymorphic site (2). Nor have they taken advantage of the benefits associated with data warehousing of readily available genomic datasets to provide a primer design system independent of a user-provided FASTA sequence. The SOP^{3}v2 software for developing primer trios for locus-specific PCR and rapid sequencing enable improved laboratory workflow during the typing of genomic polymorphisms associated with inherited phenotypes.

USING THE WEBSITE

System configuration

Warehoused genomic sequence data from the human and mouse genome projects were downloaded from the University of California Santa Cruz (UCSC) Genome Browser, consisting of the National Center for Biotechnology Information (NCBI) human genome release Build 35 of the finished human genome assembly, mouse genome release Build 33 and dbSNP release Build 123. Updates to the SOP^{3}v2 database are performed twice a year. The SOP^{3}v2 application and warehoused databases were designed to be accessible as an Internet-available web site (http://imgen.ccbb.pitt.edu/sop3v2). The application is written in preprocessor hypertext protocol (PHP) version 5. 0.3 and queries an associated MySQL 4.1 database developed on a Linux SuSE Enterprise Server 8 for the AMD64 operating system with Apache version 2.0.48 on a customized computation server (@Xi Computer, San Clemente, CA, USA) consisting of dual AMD Opteron 246 64-bit processors with 1024 kB Cache, 8192 MB random access memory and three 250 GB drives.

Input

The SOP^{3}v2 application enables user-defined queries to be examined against SNP and genome sequence databases. The SOP^{3}v2 application allows primer trio design for genetic analysis of human as well as mouse polymorphisms. The application accepts as input class gene names, SNP reference sequence numbers or chromosomal physical locations; thus, multiple queries of the same class (e.g. gene names or reference sequence numbers) can be examined simultaneously (Figure 1). Input can be entered into a text entry box provided by the application interface or by uploading a text file containing a list of loci or reference sequence numbers. Presentation of a set of queries provides the application with directions to extract the genomic sequence for up to 2000 nucleotides flanking each SNP. For example, the application accepts user input from which the algorithm can be directed to return loci-associated attributes such as SNP location, the sequence identity of the polymorphic residue, heterozygosity value for various ethnic groups and validation status. User input is also allowed via a pair of dropdown menus in order to direct the sorting of the data output, providing a tool for customization of the returned data to aid in choosing high-priority polymorphisms for laboratory evaluation.

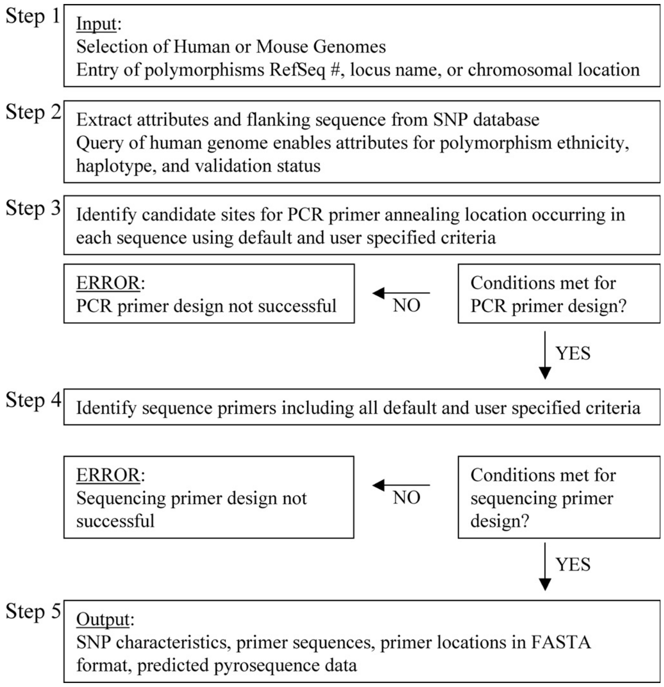

Step 1
Input:
Selection of Human or Mouse Genomes
Entry of polymorphisms RefSeq #, locus name, or chromosomal location
Step 2
Extract attributes and flanking sequence from SNP database
Query of human genome enables attributes for polymorphism ethnicity, haplotype, and validation status
Step 3
Identify candidate sites for PCR primer annealing location occurring in each sequence using default and user specified criteria
ERROR:
PCR primer design not successful
NO
Conditions met for PCR primer design?
YES
Step 4
Identify sequence primers including all default and user specified criteria
ERROR:
Sequencing primer design not successful
NO
Conditions met for sequencing primer design?
YES
Step 5
Output:
SNP characteristics, primer sequences, primer locations in FASTA format, predicted pyrosequence data

Figure 1. Schematic of the SOP^{3}v2 application used for the selection of primer trios during the design of pyrosequencing assays.

Design of locus-specific PCR primers

The graphical interface for the SOP^{3}v2 application provides a cluster of user-defined settings to allow customization of primer trio design. Genomic sequences from queried polymorphisms and their flanking sequences are returned to SOP^{3}v2's core algorithm to provide for selection of suitable primer trios. The choice of PCR primers is based on multiple criteria, including the distance between PCR primer annealing sites and the SNP, in order to provide for an amplification product that is large enough to be evaluated by agarose gel electrophoresis but not so large as to reduce product yield, as well as allowing for the selection of the optimal genomic flanking sequence for annealing of the sequencing primer. A variety of additional criteria for optimizing PCR amplification yield are included. For example, the search identifies suitable primers at multiple annealing sequences on either side of the SNP. The list of candidate PCR primers is evaluated for their ability to form competing secondary structures that may interfere with primer annealing, monitoring the predicted stability of primer annealing by calculating the melting temperature (T_m) and measurement of sequence complexity by maintaining representative frequencies of each nucleotide residue in the final primer sequence (3). Identification of nearby polymorphic sites is performed in order to improve the design of primer trios for sequencing as well as enabling possible multiplexing of assays for genotyping these residues.

Lists of candidate forward and reverse PCR primers that pass the individual tests are generated. These oligonucleotide sequences are then compared in order to obtain compatible primer pairs for use during PCR amplification. This is accomplished by analyzing primers for regions of reverse complementary sequence that may initiate primer dimer amplification. Final optimization of PCR primers for use during genotyping is accomplished by modifying PCR forward primer by the addition of an extra-genomic nucleotide motif to the 5′ end. This results in the suppression of competing sequencing reactions during genotyping analysis, which may occur whenever the template strand 3′ end is capable of annealing to an internal sequence motif, thus repressing the initiation of intra-molecular sequencing (4). The resulting PCR primers are returned to the user as oligonucleotide sequences recommended for laboratory testing.

Sequencing primer selection algorithm

The design of the primer for initiating sequencing is accomplished after the selection of the PCR primer pair. Sequencing primer selection is optimized for providing high signal-to-noise sequencing data by placing the 3′ end of the primer as near to the polymorphic residue as possible while maintaining annealing specificity and stability. Specificity is determined by selection of the 3′ end of the primer annealing site as proximal to the polymorphic residue as possible that is free of competing secondary structure and without repeating sequence motifs elsewhere within the amplified region. Primer annealing stability is maintained by requiring a user-defined T_m for annealing to the template strand. Sequencing primers are chosen on either side of the polymorphism and the one best matching the above criteria is recommended for laboratory testing.

Narrowing the query results by functional attribute

User-defined options are provided to aid in the exploitation of attributes contained within the collection of warehoused databases. The query can be focused on each SNP's associated attributes. When querying for polymorphisms based on locus symbol or chromosomal region, the application provides options for choosing reference sequence contained SNPs linked to specific genetic structural elements. For instance, the user can select to have the application design primers for polymorphisms associated with exons, introns or motifs associated with mRNA splice sites. Human genomic data are associated with heterozygosity value, ethnicity, validation status and whether the polymorphism has been included in the HapMap project. Each of these attributes can be selected for inclusion during primer trio design and polymorphisms meeting any of the criteria individually will be returned.

Output

The recommended primer trios, the identity of the SNP, genomic flanking sequence, SNP attributes and the expected pyrogram are among the elements displayed by the application output (Figure 2). Results can be sorted by attribute (e.g. locus symbol, reference sequence number, heterozygosity value or validation status). The data can be presented as a small view, consisting of the recommended primer trio, or a full view with a graphical image of the FASTA-formatted flanking sequence as well as nearby SNPs and simulated pyrosequence data. The data may also be viewed as FASTA-formatted text by selecting the text button. The returned nucleotide sequences flanking the SNP indicate nearby polymorphisms by color-coding these residues. A legend is provided explaining the color-coding scheme. The output may also be exported as an Excel-formatted file that can be saved in a directory provided on the server and downloaded onto the user's local computer.

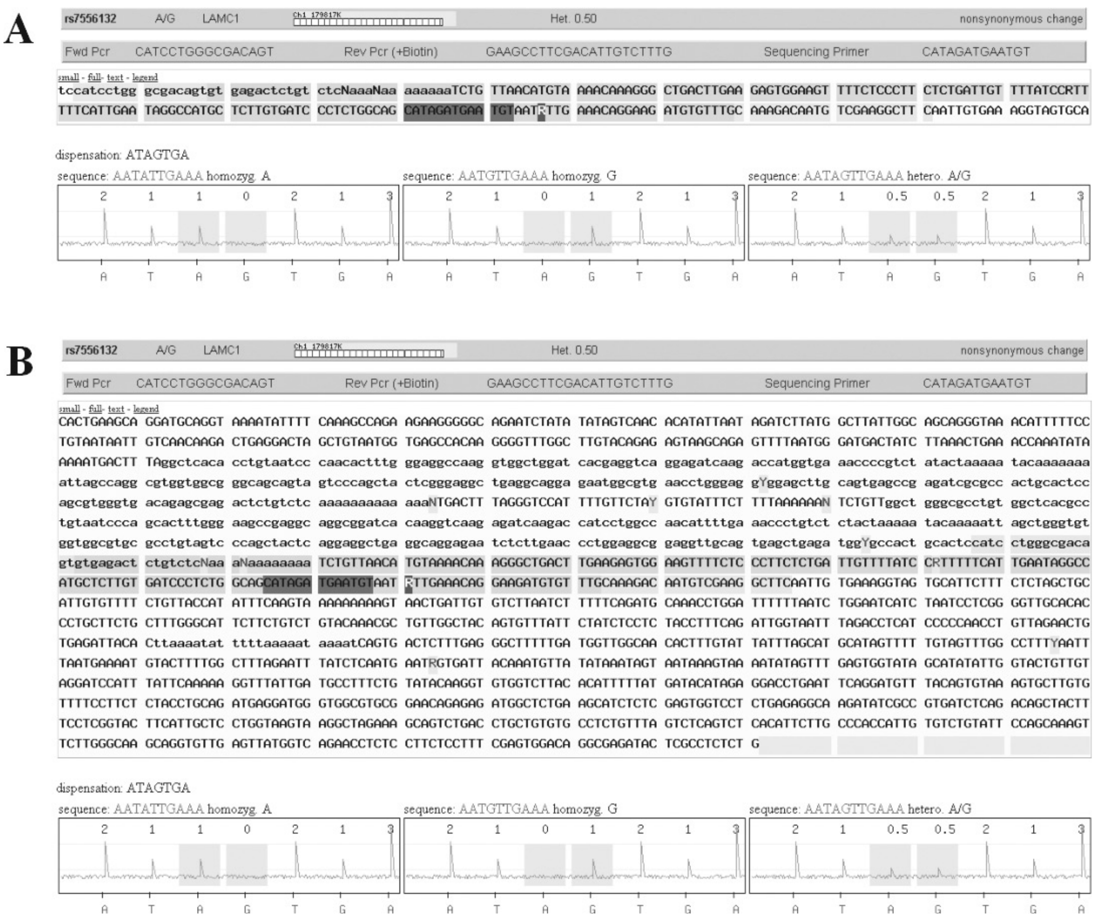
Validated genotyping assays

The SOP^{3}v2 website is linked to a list of validated primer trios for genotyping. Available through a link located on the SOP^{3}v2 homepage, the application presents primer trios that have successfully generated genotyping data in the laboratory (Table 1). Contributions to the dataset are encouraged and are made through a web-accessible form, providing users with a means to communicate their results using SOP^{3}v2-designed primers. These data are freely available and will be used to increase the application's impact on studies involving genetic analysis.

DISCUSSION

The objective of the SOP^{3}v2 application has been to provide a method to effectively search genomic databases for SNPs and insertion/deletion polymorphisms in order to design genotyping assays. An earlier version of the application, focusing on human polymorphisms, has been used to design primers that have been validated for a variety of SNPs within loci suspected of correlating with risk of developing disease (5). To the best of our knowledge, SOP^{3}v2 is the first software application that can generate PCR and sequencing primer trios based on querying genomic regions using the information available in multiple genomic databases. In SOP^{3}v2, the application has been expanded to incorporate information deposited as part of the HapMap project and the validation status of human polymorphisms, as well as to provide a tool for evaluating mouse genetics. Localizing the NCBI dbSNP database and the data from the UCSC GoldenPath human and mouse genomes allowed the creation of SQL queries to present genomic variation data as well as the annotation of functional genomic elements in a customized format. Using the SOP^{3}v2 application for a multi-loci genetic marker association study is aided by the use of the integrated LocusLink annotation, which provides the start and end markers of a gene, as well as the start and end markers of exons and introns and the location of polymorphisms associated with splice site regions. Updates of the SOP^{3}v2 database are scheduled to occur twice annually, with the next update to include locus information as well as functional and population attributes from the Entrez Gene database. SOP^{3}v2-designed primer trios continue to be evaluated in the laboratory as part of ongoing research into the genetics of complex diseases using human cohorts as well as mouse model systems (6–8).

The program aids genotyping methods by enabling locus-specific PCR amplification via improved rates of validation of primer trios and decreased time required to develop assays for



```
rs7556132    A/G    LAMC1    Het. 0.50    nonsynonymous change
Fwd Pcr  CATCCTGGGCGACAGT    Rev Pcr (+Biotin)  GAAGCCTTCGACATTGTCTTTG    Sequencing Primer  CATAGATGAATGT

tccatcctgg gcgacagtgt gagactctgt ctcNaaaNaa aaaaaaTCTG TTAACATGTA AAACAAAGGG CTGACTTGAA GAGTGGAAGT TTTCTCCCTT CTCTGATTGT TTTATCCRTT
TTTCATTGAA TAGGCCATGC TCTTGTGATC CCTCTGGCAG CATAGATGAA TGTAATRTTG AAACAGGAAG ATGTGTTTGC AAAGACAATG TCGAAGGCTT CAATTGTGAA AGGTAGTGCA

dispensation: ATAGTGA
sequence: AATATTGAAA homozyg. A    sequence: AATGTTGAAA homozyg. G    sequence: AATAGTTGAAA hetero. A/G
```

```
rs7556132    A/G    LAMC1    Het. 0.50    nonsynonymous change
Fwd Pcr  CATCCTGGGCGACAGT    Rev Pcr (+Biotin)  GAAGCCTTCGACATTGTCTTTG    Sequencing Primer  CATAGATGAATGT

CACTGAAGCA GGATGCAGGT AAAATATTTT CAAAGCCAGA AGAAGGGGGC AGAATCTATA TATAGTCAAC ACATATTAAT AGATCTTATG GCTTATTGGC AGCAGGGTAA ACATTTTTCC
TGTAATAATT GTCAACAAGA CTGAGGACTA GCTGTAATGG TGAGCCACAA GGGGTTTGGC TTGTACAGAG AGTAAGCAGA GTTTTAATGG GATGACTATC TTAAACTGAA ACCAAATATA
AAAATGACTT TAggctcaca cctgtaatcc caacactttg ggaggccaag gtggctggat cacgaggtca ggagatcaag accatggtga aaccccgtct atactaaaaa tacaaaaaaa
attagccagg cgtggtggcg ggcagcagta gtcccagcta ctcgggaggc tgaggcagga gaatggcgtg aacctgggag gYggagcttg cagtgagccg agatcgcgcc actgcactcc
agcgtgggtg acagagcgag actctgtctc aaaaaaaaaa aaaNTGACTT TAGGGTCCAT TTTGTTCTAY GTGTATTTCT TTTAAAAAAN TCTGTTggct gggcgcctgt ggctcacgcc
tgtaatccca gcactttggg aagccgaggc aggcggatca caaggtcaag agatcaagac catcctggcc aacattttga aaccctgtct ctactaaaaa tacaaaaatt agctgggtgt
ggtggcgtgc gcctgtagtc ccagctactc aggaggctga ggcaggagaa tctcttgaac cctggaggcg gaggttgcag tgagctgaga tggYgccact gcactccatc ctgggcgaca
gtgtgagact ctgtctcNaa aNaaaaaaaa TCTGTTAACA TGTAAAACAA AGGGCTGACT TGAAGAGTGG AAGTTTTCTC CCTTCTCTGA TTGTTTTATC CRTTTTTCAT TGAATAGGCC
ATGCTCTTGT GATCCCTCTG GCAGCATAGA TGAATGTAAT RTTGAAACAG GAAGATGTGT TTGCAAAGAC AATGTCGAAG GCTTCAATTG TGAAAGGTAG TGCATTCTTT CTCTAGCTGC
ATTGTGTTTT CTGTTACCAT ATTTCAAGTA AAAAAAAAGT AACTGATTGT GTCTTAATCT TTTTCAGATG CAAACCTGGA TTTTTTAATC TGGAATCATC TAATCCTCGG GGTTGCACAC
CCTGCTTCTG CTTTGGGCAT TCTTCTGTCT GTACAAACGC TGTTGGCTAC AGTGTTTATT CTATCTCCTC TACCTTTCAG ATTGGTAATT TAGACCTCAT CCCCCAACCT GTTAGAACTG
TGAGATTACA Cttaaaatat ttttaaaaat aaaatCAGTG ACTCTTTGAG GGCTTTTTGA TGGTTGGCAA CACTTTGTAT TATTTAGCAT GCATAGTTTT TGTAGTTTGG CCTTTYAATT
TAATGAAAAT GTACTTTTGG CTTTAGAATT TATCTCAATG AATRGTGATT ACAAATGTTA TATAAATAGT AATAAAGTAA AATATAGTTT GAGTGGTATA GCATATATTG GTACTGTTGT
AGGATCCATT TATTCAAAAA GGTTTATTGA TGCCTTTCTG TATACAAGGT GTGGTCTTAC ACATTTTTAT GATACATAGA GGACCTGAAT TCAGGATGTT TACAGTGTAA AGTGCTTGTG
TTTTCCTTCT CTACCTGCAG ATGAGGATGG GTGGCGTGCG GAACAGAGAG ATGGCTCTGA AGCATCTCTC GAGTGGTCCT CTGAGAGGCA AGATATCGCC GTGATCTCAG ACAGCTACTT
TCCTCGGTAC TTCATTGCTC CTGGTAAGTA AGGCTAGAAA GCAGTCTGAC CTGCTGTGTG CCTCTGTTTA GTCTCAGTCT CACATTCTTG CCCACCATTG TGTCTGTATT CCAGCAAAGT
TCTTGGGCAA GCAGGTGTTG AGTTATGGTC AGAACCTCTC CTTCTCCTTT CGAGTGGACA GGCGAGATAC TCGCCTCTCT G

dispensation: ATAGTGA
sequence: AATATTGAAA homozyg. A    sequence: AATGTTGAAA homozyg. G    sequence: AATAGTTGAAA hetero. A/G
```

Figure 2. Illustration of the SOP^{3}v2 application's output. (**A**) The output overview. (**B**) The detailed output setting. On screen, the location of primers and nearby polymorphisms are color-coded as PCR primers (yellow), sequencing primer (red), nearby polymorphisms (gray), amplified region (gold) and reference polymorphism (blue).

Table 1. Associated dataset of validated primer trios for genetic analysis

Reference sequence	Locus	Forward PCR primer	Reverse PCR primer	Sequencing primer	Product length
rs1015408	PRKCB1	acc cattctagcctgtgtcagg	gcctttcctactgtccagat	ttatctgcacttt	145
rs10797819	LAMC1	tgccgagctaaggttcaatca	aaaaaggggatagggcatgag	tatttttgaagaaaggga	201
rs1316757	ITGB1	accaatgttttgtttcaatgggatata	gacttatgtattagctgtcagg	gggaataaaatactta	139
rs1927349	COL4A2	cgggccaaggatacatatacacaca	ttctccgtcacctattcttca	cacacgcacagat	85
rs2056402	ITGBA2	gtcctasaataatacagagaaatgagt	cattgtgggcaagcatcagc	tgagtaaaagtctga	88
rs2062011	BCL2	aaattgggaggtacaagcctc	ggtgatctctggagttttcac	gccacgttgt	116
rs2256455	ITGB1	ccggtacattagatgtgatttgatgt	gaagtaggcattccttcctgt	aaggtttgaccatta	143
rs26679	ITGA2	tcgtaagtccttcccaccttgag	gcacattgctgaaaccagaat	tgatttatgaaaacag	206
rs27890	ITGA2	acggaagttgaacaagacgaaaagg	agaaatgtccagtgtccctg	ctctaagtatccaag	199
rs326	LPL	tcacactcaggccacatca	ggtgaaagcccagtaacataaga	ggtaatttgagagcctaac	133
rs3747408	COL4A5	aatgcacttggtccaaaaggtga	tccagggagcccatctaag	tttcccaggacct	81
rs4308887	COL4A5	acgacatatacaatttaatttttccat	acaaagggctgaatcaagtga	ccttgtttgattcct	92
rs4771667	COL4A2	attcctgtcctcacctgaaacca	gctccccggcctagagc	gttggctgggctt	145
rs4773161	COL4A5	acgatgccatcacttttcttcaaga	caatcctaatacaaataatagca	actgttttatttactg	106
rs5929099	COL4A5	ccgtgttgtgggagggagataac	gaaaaactaagagatatgcagaa	ggcatcccattag	173
rs5929126	COL4A5	acgtgcctgtgagacagacttaat	gttccgtagctgctatcatca	gattttcttttacaaat	173
rs6792117	TGFBR2	cgtatagcactgtttgacattctgga	ccatctgggagagagataaga	gtattttttcctca	193
rs7410919	LAMC1	actccgtcacggggttaca	gcaacacaggtctaaagaatctc	tcatggtccc	156

the evaluation of a sample's genotype. The availability of a dataset of validated primer trios furnishes a unique resource for proven assay conditions. The SOP^{3}v2 application provides distinct advantages for designing genotyping studies in that assays can be developed *de novo* or from the list of proven primer trios, thus increasing the rate of successful sample analysis. An additional benefit of the method is that the application accepts user-defined input for searching chromosome regions as well as genetic loci and reference sequence defined polymorphisms, providing multiple choices for selecting the most suitable genetic markers for evaluation during the screening of large genomic regions.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

This work was supported by grants MCB0316255(PVB) from the NSF, 1R01LM007994-01(PVB) from the NIH-NLM, U19-AI056374-01 Autoimmunity Centers of Excellence (S.R. and M.T.), RO1DK24021 (M.T.) from the National Institutes of Health and ERHS #00035010 (S.R. and M.T.) from the Department of Defense. Funding to pay the Open Access publication charges for this article was provided by ERHS #00035010.

Conflict of interest statement. None declared.

REFERENCES

1. Botstein,D. and Risch,N. (2003) Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nature Genet*., **33** (Suppl.), 228–237.
2. Ronaghi,M., Karamohamed,S., Pettersson,B., Uhlen,M. and Nyren,P. (1996) Real-time DNA sequencing using detection of pyrophosphate release. *Anal. Biochem*., **242**, 84–89.
3. Schildkraut,C. and Lifson,S. (1965) Dependence of the melting temperature of DNA on salt concentration. *Biopolymers*, **3**, 195–208.
4. Ronaghi,M., Pettersson,B., Uhlen,M. and Nyren,P. (1998) PCR-introduced loop structure as primer in DNA sequencing. *Biotechniques*, **25**, 876–878.
5. Alexander,A.M., Pecoraro,C., Styche,A., Rudert,W.A., Benos,P.V., Ringquist,S. and Trucco,M. (2005) SOP3: a web-based tool for selection of oligonucleotide primers for single nucleotide polymorphism analysis by pyrosequencing. *Biotechniques*, **38**, 87–94.
6. Mathews,C.E., Leiter,E.H., Spirina,O., Bykhovskaya,Y., Gusdon,A.M., Ringquist,S. and Fishel-Ghodsian,N. (2005) *mt*-ND2 allele of the ALR/Lt mouse confers resistance against both chemically-induced and autoimmune diabetes. *Diabetologia*, **48**, 261–267.
7. Ringquist,S., Alexander,A.M., Rudert,W.A., Styche,A. and Trucco,M. (2002) Pyrosequence based typing of alleles of the HLA-DQB1 gene. *Biotechniques*, **33**, 166–175.
8. Ringquist,S., Alexander,A.M., Styche,A., Pecoraro,C., Rudert,W.A. and Trucco,M. (2004) HLA class II DRB high resolution genotyping by pyrosequencing: comparison of group specific PCR and pyrosequencing primers. *Hum. Immunol*., **65**, 163–174.