SNPbox: a modular software package for large-scale primer design

Stefan Weckx, Peter De Rijk, Christine Van Broeckhoven and Jurgen Del-Favero*

Department of Molecular Genetics (VIB8), Bioinformatics Unit, Flanders Interuniversity Institute for Biotechnology, University of Antwerp, Universiteitsplein 1, B-2610 Antwerpen, Belgium

Received on February 10, 2004; revised on July 5, 2004; accepted on August 26, 2004
Advance Access publication September 3, 2004

ABSTRACT
Summary: We developed a modular software package SNPbox that automates and standardizes the generation of PCR primers and is used in the strategy for constructing single nucleotide polymorphisms (SNPs) maps. In this strategy, the focus of primer design can be either on the validation of annotated public SNPs or on the SNP discovery in exon regions or extended genomic regions, both by resequencing. SNPbox relies on Primer3 for the primer design and combines this program with other publicly available software tools such as BLAST, Spidey and RepeatMasker, and newly developed algorithms. Primer conditions were chosen such that PCR amplifications are uniform for each PCR amplicon facilitating the use of high-throughput genetic platforms. SNPbox can also be used for the design of primer sets for mutation analysis, STR marker genotyping and microarray oligo design. Of the 2500 primer sets designed by SNPbox, 95% successfully amplified genomic DNA under uniform PCR conditions.

Availability: The software is available from the authors upon request.
Contact: jurgen.delfavero@ua.ac.be
Supplementary information: SNPbox_supplement.

INTRODUCTION
Single nucleotide polymorphisms (SNPs) are the most frequent DNA sequence variations in the human genome with an average spacing of 1–2 kb (Cooper et al., 1985; Holden, 2002; Sachidanandam et al., 2001), and are therefore the markers of choice in genetic studies aiming at identifying susceptibility genes for complex diseases (Rafalski, 2002). SNPs can be retrieved from public databases like dbSNP (Sherry et al., 1999, 2001), HGVbase (Fredman, 2002) and JSNP (Hirakawa et al., 2002). However, the majority of SNPs in these databases have not yet been validated as true polymorphisms and/or their polymorphic content still needs to be determined in the population under investigation (Vieux et al., 2002; Marth et al., 2001). As a result, the map density when considering only validated public SNPs is often too small for detailed genetic studies. To efficiently construct high-density SNP maps, a combined strategy of SNP validation and discovery is required. For both steps, high-quality PCR and sequencing primers need to be generated, preferably in a fast, automated process, while carefully taking repeat sequences into account. Furthermore, the use of these primers in high-throughput laboratory environments requires that they amplify DNA under consistent and well-defined criteria.

Nowadays, a number of primer design tools are available as web applications or as stand-alone programs (Chen et al., 2003; Haas et al., 1998, 2003; Li et al., 1997; Proutski and Holmes, 1996; Raddatz et al., 2001; Rozen and Skaletsky, 2003). Although the efficiency of these programs is beyond dispute, most of these programs can only design one primer set at a time and therefore are less useful in large-scale primer design projects.

We present the program SNPbox offering a modular strategy for highly automated and standardized primer design in the construction of high-density SNP maps and mutation analysis based on resequencing of target sequences.

STRATEGY
SNPbox automates the primer design for a number of well-defined genomic sequences, further called ‘objects’. These objects are the starting points to define targets for which Primer3 will design primers within a frame of 70 bp 5′ and 3′ of the target. The default length of a target is 450 bp but can be changed if required. In case an object is less than the optimal target length, it is first symmetrically extended 5′ and 3′ till the optimal length is reached (Fig. 1A). When the object is >450 bp, multiple overlapping targets are defined. Small and neighboring objects will be joined into one or more targets. When SNPbox encounters repeat sequences while selecting suitable target sequences, they can be included in the target

*To whom correspondence should be addressed.
Fig. 1. SNPbox: frames, objects and targets. (A) Symmetric extension of an object till the optimal target length is reached. ‘e’ represents the extension which is the same at both sides. ‘f’ is the primer design frame of 70 bp in which Primer3 is allowed to select primers. (B) Example of the saturation module. Repeats R1 and R2 fulfill the inclusion criteria. Repeat R3 is less than the inclusion cutoff value of 300 bp, but cannot be included because of the polymorphic character of the repeat. R4 represents a repeat belonging to the interspersed repeat class, but is excluded because the repeat size is longer than the threshold size of 300 bp. The repeats R5 and R6 are interspersed repeats, but since the inter-repeat distance is <140 bp (two times the primer selection frame), they are considered as one large repeat that exceeds the threshold size of 300 bp.

SNPs are selected only if they fulfill the following criteria: a sequence similarity of ≥95% over a minimum length of 40 bp with a maximum E-value of 1E−10. The positions of the SNPs in the genomic sequence are determined and used to define the object as the region 30 bp upstream and 30 bp downstream of each SNP. Objects found within a region of 300 bp are joined into one object. In the exon module, coding sequences are identified within a genomic sequence by aligning cDNA and/or expressed sequence tag (EST) sequences using the Spidey program (Wheelan et al., 2001). In the object definition, exons are symmetrically extended by 50 bp on both sides to include the branch point and the splicing sites. In case an excluded repeat sequence is near an exon, the exon can be extended by 25 bp or not at all. Objects within a region of 250 bp are joined into one object.

In the saturation module, the objects are the parts of the genomic sequence between the excluded repeats and can consist of a regulatory region, introns of a specified gene, a complete gene or an extended chromosomal region. Targets are defined with a default overlap of 35 bp. Taking the frame of 70 bp into account in which primers are selected, the real overlap between the amplicons will be a maximum of 175 bp, including the primers. Since the length of an object is not necessarily a multifold of the optimal target length, SNPbox aims to design targets approaching the optimal target length as close as possible (Fig. 1B). The output of SNPbox consists of an HTML page with a graphical representation of the annotated genomic sequence and hyperlinks to a variety of files, allowing easy inspection of data. A tab-delimited file contains the
primer sequences, genomic position and PCR amplification conditions. Also GC-content is calculated per 50 bp of amplicon that translates into a value for average GC-content, and a minimal and a maximal GC-content.

SNPbox was successfully used in our laboratory to design primers for about 2500 targets, all on human DNA. In >95%, the PCR amplifications resulted in one specific amplion of expected size using the built-in PCR conditions without the need for optimization. SNPbox was also used to design primer sets for all exons of the human genome, based on Ensembl data. For the 208.202 exons, 227.187 objects were designated and for 98.62% of these, a target could be defined. SNPbox designed primer pairs for 98.53% of the targets, resulting in a global success rate of 97.17% (Weckx et al., 2004).

In conclusion, given that the standardization of the ‘in silico’ primer design for defined targets produced high success rates for both primer selection and subsequent PCR amplification, the software package SNPbox is a valuable asset for laboratory-involved in resequencing projects particularly when aimed at generating long distance high-density SNP maps.

ACKNOWLEDGEMENTS

We thank Dirk Van den Bossche for technical support, and Dominique Audenaert, Godelieve Claes and Rosa Rademakers for their valuable feedback during the development, fine-tuning and use of SNPbox. This work was in part funded by the Special Research Fund of the University of Antwerp, the Fund for Scientific Research Flanders and the Inter-University Attraction Poles program P5/19 of the Belgian Federal Science Policy Office.

REFERENCES


