Nucleic Acids Research, 2005, Vol. 33, No. 8 ***e72***
doi:10.1093/nar/gni071

PieceMaker: selection of DNA fragments for selector-guided multiplex amplification

Johan Stenberg, Fredrik Dahl, Ulf Landegren and Mats Nilsson*

Department of Genetics and Pathology, Rudbeck Laboratory, Se-751 85, Uppsala, Sweden

Received February 11, 2005; Revised and Accepted April 1, 2005

ABSTRACT

We describe PieceMaker, a software tool for the design of applications of selector probes–oligonucleotide probes that direct circularization of target nucleic acid molecules. Such probes can be combined in parallel to circularize a selection of fragments from restriction digested total genomic DNA. These fragments can then be amplified in a single PCR using a common primer pair, yielding substrates for subsequent analyses, such as parallel genotyping or sequencing. However, designing multiplex selector assays is a laborious task. The PieceMaker program alleviates this problem by selecting restriction enzymes to generate suitable fragments for selection, and generating the output data required to design the selector probes.

INTRODUCTION

Selectors are oligonucleotide constructs that enable circularization of selected genomic fragments with the inclusion of a standard sequence, and subsequent amplification in a multiplex format (1). A selector has target-specific single-stranded 5′ and 3′ ends, joined by a general, double-stranded segment. A DNA sample is specifically fragmented by restriction digestion, and fragments containing sequences of interest are circularized by hybridization to the target-specific selector ends and ligation to the general segment. The general sequence is thus incorporated into the DNA circles, which can then be amplified in multiplex by PCR using a standard primer pair. By use of a structure-specific endonucleolytic cleavage reaction prior to ligation, the 5′ ends of restriction fragments can be removed, allowing circularization of truncated fragments of the desired lengths (2,3).

The selector method allows multiplexed amplification of selected genomic sequences. This is promising for a number of different DNA analytic applications, such as multiplexed single nucleotide polymorphism (SNP) genotyping (4), measurements of gene copy number (5) and resequencing (6). Currently, oligonucleotide synthesis costs are high for the long oligonucleotides required by the selector method. We are developing methods for parallel synthesis of large sets of oligonucleotide probes to decrease this cost. Furthermore, designing a selector application requires the selection of a combination of restriction enzymes that will generate fragments that contain the sequences of interest, and that are suitable for circularization and amplification. The requirements on restriction fragments include limits on the length of the removed sub-fragment and the minimum and maximum length of the selected fragments to allow circularization and to achieve an even amplification of different fragments. It is also necessary that all restriction fragments for which structure-specific cleavage is used have the same nucleotide at the position where this cleavage structure is formed, to allow the subsequent ligation of the fragments to the common part of the selector.

For a given set of target sequences, an optimal design is one that minimizes the number of parallel restriction reactions required to yield suitable fragments for all targets or, alternatively, one that maximizes the number of targets for which there are suitable fragments with a given number of restriction reactions. Finding an optimal design requires the evaluation of a very large number of target/enzyme combinations.

In the present work, a computer program, PieceMaker, has been developed, which performs *in silico* restriction digestion of target sequences, finds structure-specific endonuclease cleavage positions and selects combinations of restriction enzymes. The program is applied to example target sets using different parameter settings in order to evaluate the impact of parameter choice on design success rate.

METHODS

Implementation

PieceMaker runs through a graphical user interface, integrating the sequential steps of (i) *in silico* digestion, (ii) selection of structure-specific cleavage position, (iii) fragment evaluation, (iv) reaction combination selection and (v) fragment

*To whom correspondence should be addressed. Tel: +46 018 471 4816; Email: Mats.Nilsson@genpat.uu.se selection, all into one single program. Each of these five modules will be described below. Sequence and restriction enzyme data are provided by the user as input files, while application-specific parameters are set through the user interface.

In silico digestion

In the *in silico* digestion step, input sequences are cleaved by restriction enzymes to generate sets of fragments. Each input sequence is a $5'\rightarrow3'$ sequence of nucleotide symbols (including the degeneracy symbols), having a region of interest that is defined by two position values denoting the region's beginning and end. This region represents the sequence of interest, while the input sequence also contains flanking sequences, required for the design. A reaction represents a combination of one or more restriction enzymes to be used in a single digestion reaction. Each reaction exhibits one or more cleavage patterns, each made up of a recognition sequence, a plus strand cleavage position and a minus strand cleavage position. Table 1 describes example reactions.

For every combination of input sequence and reaction, restriction sites in the sequence are found by comparing the nucleotide sequence at every position with the recognition sequence of each cleavage pattern of the reaction. If a match is found, the cleavage positions for the plus and minus strands are determined by adding the plus and minus cleavage position values to the position of the match. As degeneracy symbols are allowed in the input sequence, cleavage positions are divided into two classes; certain and possible cleavage positions. The former are positions where the match is independent of the possible variants in any degenerate positions involved in the match, while the latter are positions where a match is found only for some variants. Restriction fragments are created for each neighboring pair of cut positions in the set. Each fragment is defined by its beginning and end within the input sequence, its polarity relative to the input sequence, the beginning, r_b, and end, r_e, of the region of interest within the fragment, and the position, p_c, of the $3'$-most of the possible cut positions within the fragment, if any. All fragments generated by cleaving an input sequence with a reaction are collected in a fragment set for that sequence-reaction couple.

Selection of structure-specific cleavage position

In this step, a position for the structure-specific cleavage is selected for all fragments. Only the sequence downstream of this position will be selected, i.e. included in the circular product. The position selection procedure is performed based on four parameters: the minimum and maximum selection lengths, s_{min} and s_{max}, define the allowed length span of the selected part of the fragment; the maximum flap length, f_{max}, defines the maximum allowed length of the sub-fragment that is removed; and the cleavage base, b, defines what nucleotide the cleavage should occur immediately $3'$ of. The length of the fragment is denoted by l. The choice of the cleavage position, p, determines the content, $C(p)$, of sequence of interest in the selected sub-fragment. The nucleotide at a position q is denoted $B(q)$. The above definitions are illustrated in Figure 1. The cleavage position must be located downstream of any possible cut positions and should be no further from the $3'$ end than the maximum selection length. Similarly, the position should be no closer to the $3'$ end than the minimum selection length and no further from the $5'$ end than the maximum flap length. If the fragment length is longer than the maximum flap length plus the maximum selection length, there will be no valid position, as either the flap or the selected fragment will exceed its limit for every possible position. The interval P of possible positions is defined as $P = [p_{min}, p_{max}]$, where:

$$p_{min} = \max(l - s_{max}, p_c + 1)$$
$$p_{max} = \min(l - s_{min}, f_{max})$$

For a position, p, to be a valid cleavage position, it must belong to the interval P, and the nucleotide at p must be equal to b, thus satisfying the following criteria:

$$p \in P$$
$$B(p) = b$$

An optimal cleavage position also maximizes the content, thus satisfying:

$$\forall q \in P \mid B(q) = b, C(p) \geqslant C(q)$$

Table 1. Cleavage patterns for an example set of reactions of one or two restriction enzymes

Reaction	Specificity	Sequence	Cleavage patterns	
			Plus cleavage position	Minus cleavage position
CviA II	C/ATG	CATG	1	3
Hph I	GGTGA (8/7)	GGTGA	13	12
		TCACC	−7	−8
Hpy188 I	TCN/GA	TCNGA	3	2
CviA II + Hpy188 I	C/ATG + TCN/GA	CATG	1	3
		TCNGA	3	2

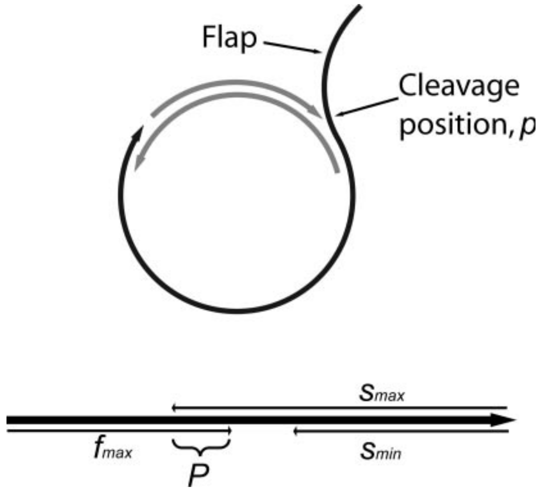


Figure 1. Illustration of definitions of design parameters. This Figure shows the definition of maximum flap length, f_{max}, and minimum and maximum selection length, s_{min} and s_{max}. The flap is the sub-fragment that is cleaved of by the structure-specific cleavage. The interval P of allowed positions for this cleavage is defined by f_{max}, s_{min} and s_{max} as well as by the presence of any possible restriction enzyme cleavage positions, none shown in this Figure.

The region of interest is a contiguous subsequence of the fragment, defined by the two positions r_b and r_e, where $r_\mathrm{b} \leqslant r_\mathrm{e}$. From this follows that:

$$C(p) = \begin{cases} r_\mathrm{e} - r_\mathrm{b} + 1 & p < r_\mathrm{b} \\ r_\mathrm{e} - p & r_\mathrm{b} \leqslant p < r_\mathrm{e} \\ 0 & p \geqslant r_\mathrm{e} \end{cases}$$

Four cases can be identified based on the values of r_b and r_e:

(i) $r_\mathrm{b} \leqslant p_\mathrm{min} + 1, r_\mathrm{e} \leqslant p_\mathrm{min}$
(ii) $r_\mathrm{b} \leqslant p_\mathrm{min} + 1, r_e > p_\mathrm{min}$
(iii) $p_\mathrm{min} + 1 < r_b < p_\mathrm{max}$
(iv) $r_\mathrm{b} \geqslant p_\mathrm{max}$

In case (i), $C(p)$ equals zero for all p within P. In case (ii), $C(p)$ is strictly decreasing until $p = r_\mathrm{e} + 1$, where $C(p)$ becomes zero. The optimal position is thus the first position in P that satisfies $B(p) = b$. In case (iii), any position in the sub-interval $P1 = [p_\mathrm{min}, r_\mathrm{b} - 1]$ that satisfies $B(p) = b$ is optimal, and if no such position exists, the solution is the same as in case (ii). In case (iv), finally, any position that satisfies the base criterion is optimal. In the actual implementation of cases (i) and (iv), the highest valid value of p is selected. This is also true for those instances of case (iii) where a valid p exists within $P1$. Thus, the algorithm selects the highest, optimal value of p, if any valid p exists.

Fragment evaluation

Each fragment set is evaluated and fragments that satisfy the evaluation criteria are accepted. Parameters for this evaluation are minimum content, c_min, minimum and maximum selection length, s_min and s_max, and maximum flap length f_max. Fragments are accepted if they satisfy the following criteria:

(i) $l \geqslant s_\mathrm{min}$
(ii) $l - s_\mathrm{max} \leqslant p \leqslant f_\mathrm{max}$
(iii) $C(p) \geqslant c_\mathrm{min}$

Note that criterion (ii) will be satisfied for all fragments where a valid position p was found in the previous step. If no such position was found, the fragment will not be accepted.

Reaction combination selection

After fragment evaluation, a fragment set with zero or more accepted fragments exists for each target–reaction couple. By selecting a combination of one or more reactions, a combined set of fragments is created for each target. The task in the reaction combination selection step is to select, for a given number n, a combination of at most n reactions that maximizes the number of targets for which the combined fragment set satisfies some application-specific condition of success. For every $k = 1, 2, \ldots, n$, each combination of k reactions is evaluated for each target until a combination is found that satisfies the success condition for all targets. If no such combination is found, the first combination tested among those yielding the highest number of successes is selected instead. Let r be the number of reactions and t the number of targets. The number of reaction combinations to be tested is then:

$$\sum_{k=1}^{n} \binom{r}{k}$$

The total number of success-condition evaluations thus grows roughly as $r^n \times t$, making this the most demanding step of the design process, as r and/or n increases.

Fragment selection

The selected combination of reactions may yield multiple suitable fragments for some of the targets, and thus a subset of those fragments can be selected for use. The scheme used for this selection is dependent on the application, e.g. the fragments closest in length to a certain value, or with the largest content of sequence of interest, may be selected.

Design examples

An SNP target set was created by random selection of 100 SNPs. A total of 1000 bases of flanking sequence on each side of the SNPs were downloaded with SNPper (7). The SNP identification numbers are shown in Supplementary Table S1. The region of interest was defined to be 20 nt on each side of the SNP, i.e. positions 981 through 1021. The exon target set consists of 101 targets, one for each coding exon of the genes ATM, RB1 and P53. The coding exon sequences were set as regions of interest and 1000 nt of flanking sequence on each side were downloaded from http://ncbi.nlm.nih.gov. The enzyme set used in both examples consists of the 15 enzymes displayed in Table 2.

RESULTS

The PieceMaker program finds an optimal solution to a defined selector assay design problem. The user defines this problem by specifying sequences with defined regions of interest, restriction enzyme reactions, parameters for cleavage position selection and fragment evaluation, maximum number of reactions in a combination and success criteria for reaction combinations. The input sequences are *in silico* digested in a set of restriction reactions, each containing one or more restriction enzymes. For restriction fragments that are longer than a specified length, an optimal position for the structure-specific cleavage is selected, followed by the evaluation of the fragments. Fragments that do not allow selection of a valid cleavage position or that do not satisfy the evaluation criteria are discarded. The success criteria are applied to every

Table 2. The 15 restriction enzymes used in the design examples

Enzyme	Sequence
Alu I	AG/CT
Bbv I	GCAGC (8/12)
Bcc I	CCATC (4/5)
Bsp1286 I	GDGCH/C
CviA II	C/ATG
Dde I	C/TNAG
FspB I	C/TAG
Hph I	GGTGA (8/7)
Hpy188 I	TCN/GA
HpyCH4 V	TG/CA
Mbo II	GAAGA (8/7)
Mly I	GACTC (5/5)
Mnl I	CCTC (7/6)
Mse I	T/TAA
Sty I	C/CWWGG

combination of reactions and one is selected that yields the maximum number of successes with the minimum number of reactions. If preferred, a non-redundant subset can be selected among the restriction fragments generated by the selected combination of reactions. The sequences of the selected fragments can then be used for designing the selector sequences either manually or, preferably, using oligonucleotide probe design software, such as ProbeMaker (J. Stenberg, M. Nilsson and U. Landegren, manuscript in preparation). The *in silico* digestion and structure-specific cleavage position selection steps of the design process are illustrated by a simplified design, shown in Figure 3.

Two sets of sequences were *in silico* digested with a set of reactions containing 15 restriction enzymes and all possible pairs of those enzymes. Cleavage position selection and fragment evaluation were performed using a number of different parameter choices, and for each set of parameters the best combinations of one, two and three reactions were selected. For the SNP target set, a success was called for a target if there was at least one accepted fragment containing the complete region of interest. For the exon target set, a success was called for a target if the complete region of interest was covered by the contents of a combination of one or more accepted fragments. The number of successes achieved using different parameter sets are shown in Figure 4A and B. The trends are the expected ones; success rate improves when reactions are used in parallel, when allowing a wider range of fragment lengths, and a longer flap. Resulting sets of fragments from the best designs are shown in Supplementary Tables S2 and S3. In the SNP set, design was unsuccessful for one target regardless of parameter choice. The SNP of this target (rs5746536) is located in a region of repeated sequence with ∼90% A/T-content. None of the reactions used produced a suitable restriction fragment for this target.

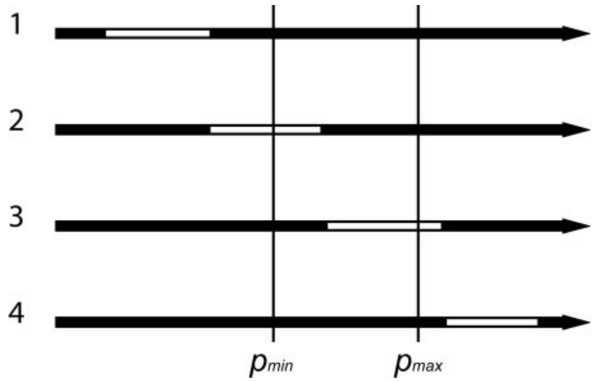
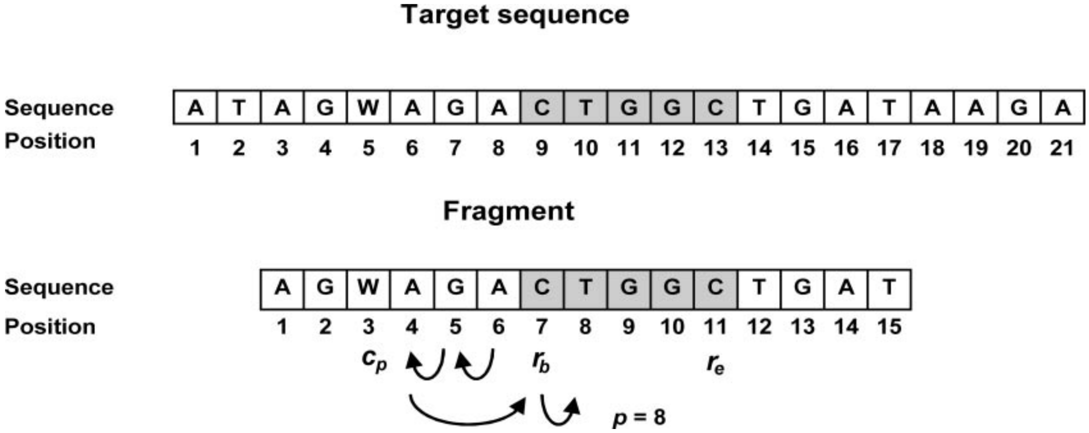


Figure 2. The different cases of structure-specific cleavage position selection. The position of the region of interest (white) within the fragment (black) in relation to $p_{\min}$ and $p_{\max}$ determines the algorithm by which the cleavage position is selected.

DISCUSSION

An early version of PieceMaker was used to design the set of 96 selectors used in the demonstration of the selector method (1). This work also investigated how an increasing number of targets affected the design success rate and found that there was no adverse effect. The full version of the program has now been completed and is described in the present work. To further demonstrate the utility of the program and to examine the impact of parameter choices on the results, several designs were carried out on two sets of target sequences; a set of 100 targets containing SNPs and a set of 101 targets containing the coding regions of the exons of the ATM, RB1 and P53 genes.

As seen in the results of the design examples, design success rate is greatly improved by allowing longer flaps and a wider range of selection lengths. The choice of maximum flap length should depend on the mechanism of the structure-specific cleavage, which may be hindered by long flap lengths and the associated risk of secondary structure. Using Platinum *Taq* DNA polymerase (Invitrogen) as the structure-specific


Target sequence

Sequence	A	T	A	G	W	A	G	A	C	T	G	G	C	T	G	A	T	A	A	G	A
Position	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21

Fragment

Sequence	A	G	W	A	G	A	C	T	G	G	C	T	G	A	T
Position	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

Figure 3. *In silico* digestion. A target sequence is digested *in silico* in a reaction consisting of a single enzyme cleaving in the middle of the recognition sequence 'TA'. This reaction is certain to cleave the target sequence at two positions; between positions 2 and 3, and between positions 17 and 18. Cleavage between positions 5 and 6 is possible, but not certain, since it depends on the nucleotide actually present at position 5. This digestion thus generates one fragment, having a length, *l*, of 15, with the region of interest beginning at position $r_b = 7$ and ending at position $r_e = 11$. A possible restriction enzyme cleavage position c_p exists at position 3. Selection of structure-specific cleavage position. Finding an optimal cleavage position, *p*, using the parameters *b* = 'T', $s_{\max} = 15$, $s_{\min} = 5$ and $f_{\max} = 10$ next proceeds as follows: the interval *P* is [4, 10] according to the definition of *P*. Since $p_{\min} + 1 < r_b < p_{\max}$, we have an instance of case (iii). Thus, we start searching for a 'T' in the sub-interval [4, 6]. Positions are interrogated in the order 6, 5, 4. As no 'T' was found, we continue the search in the sub-interval [7, 10], starting from position 7. We find a 'T' at position 8, and thus set *p* = 8. This gives $C(p) = C(8) = 3$.

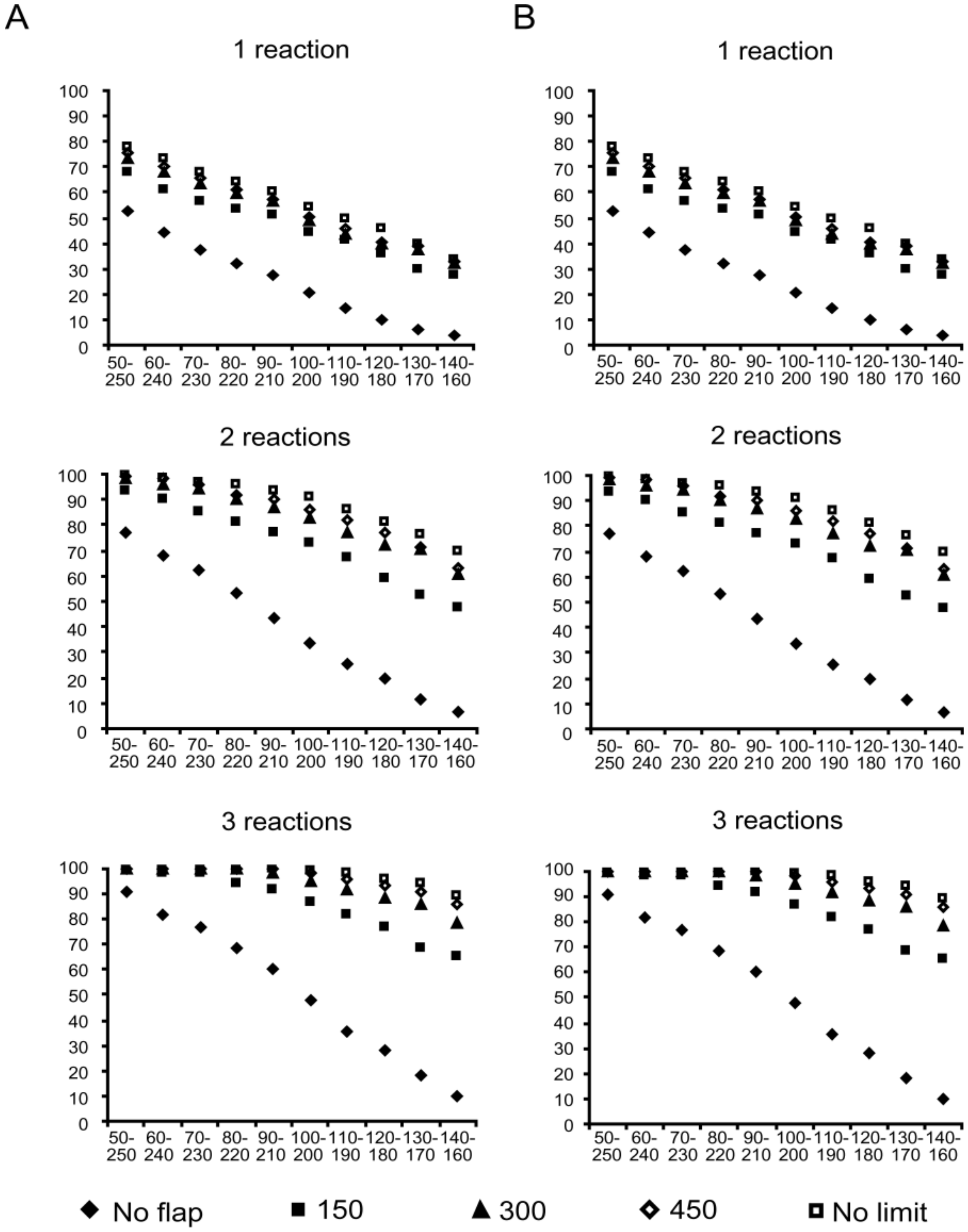


Figure 4. Impact of parameter choice on design success rate. The *x*-axes show length limits for selected fragments; *y*-axes show proportion of targets for which the design was successful, in per cent. (**A**) The SNP target set. (**B**) The exon target set.

endonuclease, we have previously used flap lengths of up to 600 with no adverse effects on assay performance.

In most cases, the limits on selection length will be determined by the requirements of the application, rather than those of the reaction mechanism. Specifically, longer selected fragments may result in less PCR product than with shorter ones, while a wider distribution of lengths may yield a worse quantitative representation of the different amplicons than a narrower one would. Very short fragments will be difficult to circularize owing to the rigidity of the double-stranded segment of the selector, while for long fragments the distance between the selector binding sites will tend to reduce the efficiency of circle formation. We have successfully circularized fragments from 100 up to 1000 nt in length.

The average success rate of selector application design also depends on other application-specific properties, such as target region length, the amount of variability in the target sequences and the requirements on the resulting fragment set. The design-success performance for new types of applications involving selectors is difficult to predict and will have to be examined as these applications appear.

In this work, one obstacle for applying the selector method to large sets of targets has been overcome by the development of computer software to find optimal designs for given selector applications.

AVAILABILITY

The PieceMaker software is written in Java (Sun Microsystems) and should thus run on any system with a Java Runtime Environment (JRE) installed. The software was compiled and tested using Java version 1.4.2 and thus a JRE compliant with this version may be required. The compiled Java class files required to run PieceMaker are available free of charge for academic users by request to the authors.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

This work was supported by grants from the Wallenberg Consortium North, the EU-FP6 integrated project MolTools, and the Swedish Research Councils for natural sciences and technology, and for medicine. Funding to pay the Open Access publication charges for this article was provided by the MolTools project.

Conflict of interest statement. None declared.

REFERENCES

1. Dahl,F., Gullberg,M., Stenberg,J., Landegren,U. and Nilsson,M. (2005) Multiplex amplification enabled by selective circularization of large sets of genomic DNA fragments. *Nucleic Acids Res.*, vol 33, e71.
2. Lyamichev,V., Brow,M.A. and Dahlberg,J.E. (1993) Structure-specific endonucleolytic cleavage of nucleic acids by eubacterial DNA polymerases. *Science*, **260**, 778–783.
3. Holland,P.M., Abramson,R.D., Watson,R. and Gelfand,D.H. (1991) Detection of specific polymerase chain reaction product by utilizing the 5′–3′ exonuclease activity of *Thermus aquaticus* DNA polymerase. *Proc. Natl Acad. Sci. USA*, **88**, 7276–7280.
4. Pastinen,T., Kurg,A., Metspalu,A., Peltonen,L. and Syvanen,A.C. (1997) Minisequencing: a specific tool for DNA analysis and diagnostics on oligonucleotide arrays. *Genome Res.*, **7**, 606–614.
5. Schouten,J.P., McElgunn,C.J., Waaijer,R., Zwijnenburg,D., Diepvens,F. and Pals,G. (2002) Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. *Nucleic Acids Res.*, **30**, e57.
6. Shendure,J., Mitra,R.D., Varma,C. and Church,G.M. (2004) Advanced sequencing technologies: methods and goals. *Nature Rev. Genet.*, **5**, 335–344.
7. Riva,A. and Kohane,I.S. (2002) SNPper: retrieval and analysis of human SNPs. *Bioinformatics*, **18**, 1681–1685.