Published online 12 June 2007

Nucleic Acids Research, 2007, Vol. 35, No. 11 ***e83***
doi:10.1093/nar/gkm410

Selection of restriction endonucleases using artificial cells

Yu Zheng* and Richard J. Roberts

New England BioLabs, Inc.,240 County Road, Ipswich, MA 01938, USA

Received February 9, 2007; Revised April 27, 2007; Accepted May 6, 2007

ABSTRACT

We describe in this article an *in vitro* system for the selection of restriction endonucleases using artificial cells. The artificial cells are generated in the form of a water-in-oil emulsion by *in vitro* compartmentalization. Each aqueous compartment contains a reconstituted transcription/translation mix along with the dispersed DNA templates. In the compartments containing endonuclease genes, an endonuclease expressed *in vitro* cleaves its own DNA template adjacent to the gene, leaving a sticky end. The pooled DNA templates are then ligated to an adaptor with a compatible end. The endonuclease genes are then enriched by adaptor-specific PCR on the ligation mix. We demonstrate that the system can achieve at least 100-fold enrichment in a single round of selection. It is sensitive enough to enrich an active endonuclease gene from a 1:10^5 model library in 2–3 rounds of selection. Finally, we describe experiments where we selected endonuclease genes directly from a bacterial genomic DNA source in three rounds of selections: the known PstI gene from *Providencia stuartii* and the new TspMI gene from *Thermus* sp. *manalii*. This method provides a unique tool for cloning restriction endonuclease genes and has many other potential applications.

INTRODUCTION

Restriction endonucleases have been the workhorse of molecular biology for the past 30 years (1). They catalyze the breakage of phosphodiester bonds on DNA backbones at specific sites and, together with their companion methyltransferases, are part of bacterial defense systems against the invasion of bacteriophages. Expression of restriction endonucleases in *Escherichia coli* without the proper protection of the companion methyltransferases usually results in cell death. For this reason, restriction endonucleases have proven to be difficult candidates for direct cloning or for engineering efforts to change their properties using living hosts (2). For example, the traditional cloning approach (3) relies on the fact that the restriction endonuclease gene (RE gene hereafter) and its companion DNA methyltransferase gene often sit close on the chromosome allowing selection for the methyltransferase gene and its flanks to carry along the RE gene. A complete *in vitro* approach would diminish the effect of cell toxicity and may be better suited for many purposes. One such scheme has previously been applied to the selection of restriction enzyme genes (4), in which the selection is based on using a DNA polymerase to incorporate dUTP-biotin to the sticky ends generated by the restriction endonuclease in water-in-oil emulsion. DNA templates with dUTP-biotin extensions are then captured on streptavidin-coated beads and amplified. Using this method, a selection efficiency of ~10-fold enrichment was obtained in a single round. This relatively low efficiency limits the use of this method to certain specific applications. For instance, six rounds of selections were needed to select an active FokI gene from a randomized FokI library at three codon positions.

An ideal selection system is a simplified Darwinian process, in which only genes surviving the imposed selection criteria are allowed to propagate. Among many crucial requirements of this process are the separation of distinct genotypes and the linkage between genotype and phenotype. Living hosts such as *E. coli* cells fulfill these requirements by cell membrane encapsulations and by the viability of the selected clones. *In vitro* methods that have been developed based on these considerations include *in vitro* compartmentalization (IVC) (5), mRNA display (6) and ribosomal display (7) etc. While the various display methods are useful choices for the selection of binding, *in vitro* compartmentalization provides the necessary ingredients for carrying out activity-based selections in a cell-like environment. Since its introduction in 1998 (5), IVC has been applied to a wide range of biomolecular engineering applications (8).

The *in vitro* compartmentalization (IVC) (5) technique generates as many as 10^9–10^{10} individual aqueous droplets in oil. In our selection procedure, the aqueous phase in

*To whom correspondence should be addressed. Tel: (978)3807441; Fax: (978) 380-7406; Email: zhengy@neb.com

each droplet contains the reconstituted transcription/translation system (9) and is capable of protein translation from the linear DNA templates dispersed inside. Being stable over the process of selection, these droplets provide a simplified means to mimic *E. coli* cells as 'artificial cells'. The selection scheme utilizes the restriction endonuclease's ability to generate defined sticky ends on DNA templates, which, in cellular compartments, ensures the linkage between genotype and phenotype for selection. Briefly, active endonuclease is expressed *in vitro* and cleaves its encoding DNA templates in the same droplet, leaving a defined sticky end at the tail. The recovered DNA templates and an excess of double-stranded adaptors with compatible sticky ends are then ligated. Only those templates that have been cleaved by the encoded endonuclease and carry intact sticky ends can be ligated efficiently. They are then amplified using adaptor-specific PCR to enrich the RE genes.

Model selections were carried out using libraries containing an excess of a Green Fluorescent Protein (GFP) gene spiked with various amounts of the gene encoding the PstI restriction endonuclease (recognition sequence CTGCA↓G), which would generate a four-base 3′-overhang. We show that at least 100-fold enrichment is reached in a single round of selection. Multiple rounds of selection are carried out to achieve successive enrichment. Finally, as a 'real' test of the system's selection power, we challenge it by using libraries constructed from the genomic DNA of a single bacterial species. We show that by three rounds of iterative *in vitro* selections, the RE gene becomes the single dominating DNA species in the resulting library. Using this method, we have cloned the PstI gene from *Providencia stuartii* and the TspMI gene (10) from *Thermus sp. manalii*.

We believe that the *in vitro* approach offers a unique route for endonuclease selection and engineering. The general principle demonstrated here may be applicable to a broad range of other genes encoding selectable enzymatic activities.

MATERIALS AND METHODS

All PCRs were carried out using the high-fidelity Phusion polymerase (Finnzyme) according to the manufacturer's instructions. All oligos were synthesized at New England Biolabs (NEB, see Table S1 in Supplementary Materials for oligo details). DNA purifications, if not otherwise specified, used the spin-column procedure (Qiagen). Enzymes, if not otherwise specified, are all from NEB.

Model library construction

The PstI gene was first cloned into the pLT7K vector (11) and then amplified from the plasmid. The GFP gene was amplified from the pIVEX-GFP vector (Roche). The 5′-untranslated regions upstream of the T7 promoter are the same for both templates. Both reverse primers have two tandem repeats of the PstI recognition site (CTGCAG) (Figure 2a). PCR products were gel purified. The concentration of purified DNA was determined by A_{260} readings and by gel electrophoresis. Model libraries were constructed by mixing the PstI and GFP templates in variable molar ratios, 1:100, $1:10^3$, $1:10^4$ and $1:10^5$, with a final concentration of 10 ng/μl.

Calculation of theoretical enrichment

A Poisson distribution is assumed to describe the distribution of DNA templates in aqueous droplets. This implies that all droplets are of equal volumes. The probability that a droplet has $n = 0,1,2$ or more DNA templates can be calculated by:

$$f(n,\lambda) = \frac{e^{-\lambda}\lambda^n}{n!}$$

where λ is the ratio between the number of DNA templates and the number of droplets in the emulsion.

We assume that all the DNA templates in the same droplet containing at least one RE gene are selected, i.e. 100% selection efficiency. Thus the number of RE genes after selection is:

$$N_{re} = N\sum_{n=1} f(n,\lambda)\cdot\sum_{k=1}^{n}\binom{n}{k}\cdot p^k\cdot(1-p)^{n-k}\cdot k$$

where N is the total number of droplets in the emulsion and p is the percentage abundance of the RE gene in the starting library. The number of 'carryover' genes, which refer to those non-RE templates residing in the same droplets with an endonuclease gene, is:

$$N_{co} = N\sum_{n=1} f(n,\lambda)\cdot\sum_{k=1}^{n}\binom{n}{k}\cdot p^k\cdot(1-p)^{n-k}\cdot(n-k)$$

The final ratio between the selected RE gene and the 'carryover' gene, if we assume no bias in the PCR amplification, is:

$$r = \frac{N_{re}}{N_{co}}$$

The theoretical enrichment is approximately:

$$E = \frac{r}{p/(1-p)}$$

Genomic library construction

Bacterial strains were obtained from the NEB strain collection. Here, ~10 μg of purified genomic DNA (gDNA) was sheared using a nebulizer (Invitrogen, K7025-05) according to the manufacturer's instructions. Sheared gDNA was precipitated by isopropanol, re-suspended in water and size-selected from 1to3 kb by agarose gel electrophoresis. The ends of size-selected gDNA are heterogeneous and were blunted using Phusion polymerase ($3'\rightarrow5'\text{exo}^+$) with dNTP at 72°C for 2 h. Purified blunt-ended gDNA fragments were phosphorylated using T4 polynucleotide kinase at 37°C for 1 h.

Vector pYZ6 is derived from pIVEX2.4 (Roche), with the following modifications: (i) the NruI site (TCG↓CGA) and the MscI site (TGG↓CCA) have been added to the

multiple cloning region immediately after the ribosome-binding site to allow the insertion of blunt-ended DNA fragments; (ii) two PstI sites are added after the multiple cloning region (two TspMI sites in the TspMI genomic selection). Circular pYZ6 plasmid was linearized by NruI digestion and purified. Ligation between the gDNA fragments and pYZ6 was carried out using T4 ligase in the presence of NruI (1 U/10 μl) at room temperature overnight. DNA was purified from the ligation mixture. One microliter of purified DNA was transformed into chemical competent cells (NEB Turbo) to judge the library quality and estimate the extent of coverage of the genome.

Emulsion PCR (emPCR) was then performed to 'clonally' amplify linear DNA templates from the ligated gDNA using primers 561 and 825III (12). Briefly, 200 μl of the aqueous PCR mix was added to 400 μl stirring oil mix [4.5% v/v Span 80 (Fluka), 0.45% v/v Tween 80 (Sigma), 0.05% Triton-X100 (EM Science) in light mineral oil (Sigma)] at 1000 rpm in a dropwise manner over 1.5 min. After the addition was complete, the stirring was continued for 5 min. The emulsion was pipetted into 10 aliquots of 50 μl in PCR tubes and overlayed with mineral oil. Reactions were heated to 98°C for 60 s, then cycled 30 times (98°C 10 s, 55°C 20 s, 72°C 90 s), then 7 min at 72°C. The primers for emPCR anneal to the vector arms: the forward primer 561 is ~100 nt upstream of the T7 promoter and the reverse primer 825III is downstream of the PstI sites (Figure S1 in the Supplementary Materials). Amplified DNA from emulsion PCR was purified as described in (12) (Figure 4a) and used for *in vitro* selection.

Selection using *in vitro* compartmentalization

The reconstituted PURE system (Post Genome Institute, Japan) was used for *in vitro* transcription/translation reactions. Fifty microliters of chilled aqueous mix (25 μl solution A, 10 μl solution B, 14 μl H_2O, 1 μl library) was added to 450 μl stirring oil mix [0.5% v/v Triton X-100 (EM Science) and 4.5% v/v Span 80 (Fluka) in light mineral oil (Sigma)] at 1200 rpm (Variomag Telesystem HP15P) and stirred for an additional 5 min. The emulsion was incubated at 37°C for 2 h to allow *in vitro* transcription/translation. In PstI selections, the reactions in the emulsion were stopped by first heating to 80°C for 20 min and then adding 50 μl quenching buffer (10 mM Tris, 20 mM EDTA, pH = 8.0). The emulsion was then spun for 15 min at 14 000 rpm at 4°C. The upper oil phase was removed and the residual emulsion was broken by extracting with 1 ml of water-saturated ether. Residual ether was removed by spinning for 5 min in a Speedvac. The DNA library was recovered by the spin-column procedure and eluted in 50 μl buffer EB (Qiagen).

Purified DNA after each emulsion selection is ligated with an excess (>100 fold) of short double-stranded adaptors (100–200 nt). Adaptors are excised from purified DNA by restriction enzyme digestions (see Supplementary Materials for details). Two microliters out of 10 μl ligation mix was used for adaptor-specific PCR (initial 98°C 60 s, 30 cycles of 98°C 10 s, 55°C 20 s, 72°C 60 s, final extension 72°C 7 min). Forward primers used in successive rounds of selection are nested to increase the specificity of PCR (see Supplementary Materials for details). After PCR, DNA was spin-column purified and was used for the next round of selection.

Cloning DNA after selection

DNA bands after selection were excised and purified on an agarose gel. Selected DNA was then digested with the restriction enzyme (PstI in PstI selection and XmaI in TspMI selection) and ligated into pLT7K. pLT7K was designed to accommodate toxic genes (11). Ligated DNA was transformed into NEB Turbo and plated onto LB-Amp plates. Plates were grown at 37°C overnight. Individual clones were picked and grown in LB media with ampicillin. Plasmids were extracted by the mini-prep procedure and sequenced.

RESULTS

Specific enrichment of RE genes in model selections

The selection of RE genes relies on their ability to generate sticky ends which are later used for ligation and PCR amplification, as illustrated in Figure 1. DNA templates for *in vitro* selection are engineered so that at one end there are the necessary elements for efficient transcription and translation (T7 promoter, ribosome-binding site) and at the other, there are two tandemly repeated PstI recognition sites as the substrates (Figure 2a). DNA templates mixed with the *in vitro* transcription/translation system are dispersed into up to 10^{10} aqueous droplets as artificial cells (5). In droplets containing RE genes, active endonuclease is expressed *in vitro* and cleaves its own encoding DNA templates, leaving sticky ends at the tail. Active endonuclease molecules are confined to individual droplets to ensure the genotype–phenotype linkage. After the reaction in the emulsion is stopped, DNA templates are pooled and put into a ligation mixture with an excess of adaptors which have compatible sticky ends. Adaptor-specific PCR is then carried out to specifically amplify DNA templates to which the adaptor has ligated. This is achieved by using the reverse primer which only hybridizes to the adaptor while the forward primer is common to all DNA templates.

We constructed model libraries which consist of two DNA templates, one has the PstI open reading frame (ORF) and the other has the GFP ORF. The two templates were mixed in variable molar ratios with decreasing concentrations of the PstI template. The PstI template is ~1.3 kb in size and the GFP template is ~1.2 kb (Figure 2a). ~10^{10}(1 μl of model library at 10 ng/μl) template molecules were used when starting all model selections. The same amount was also used in the control experiments. As a positive control, the initial library was digested with pure PstI enzyme, followed by adaptor ligation and PCR amplification. In principle, all templates in the library should be amplified in the positive control experiment and the final molar ratio between the templates reflects the selection efficiency in the absence of a genotype–phenotype linkage. A negative control was

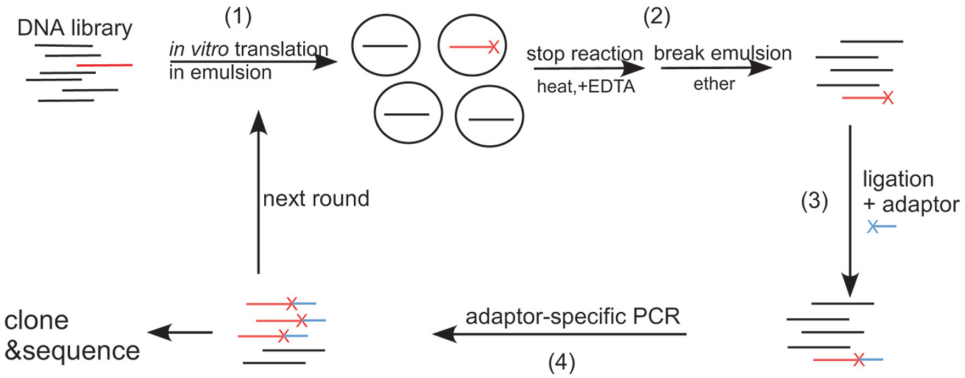

Figure 1. Diagram of the *in vitro* selection of endonucleases. (1) The DNA library (endonuclease gene in red) in the mix containing the PURE system is dispersed into aqueous droplets in a water-in-oil emulsion. Active endonuclease is expressed *in vitro* and cleaves the tail of its encoding gene at its recognition sequence (shown in red) in the droplets. (2) The reaction in the emulsion is quenched by heating and adding EDTA. The emulsion is broken by adding water-saturated ether. The DNA library is recovered from the aqueous phase. (3) Ligation is performed between the recovered DNA and an excess of a double-stranded adaptor with a compatible sticky end. (4) Adaptor-specific PCR is performed using the ligation mix. After PCR, the purified DNA either enters the next round of selection or proceeds to cloning and sequencing.

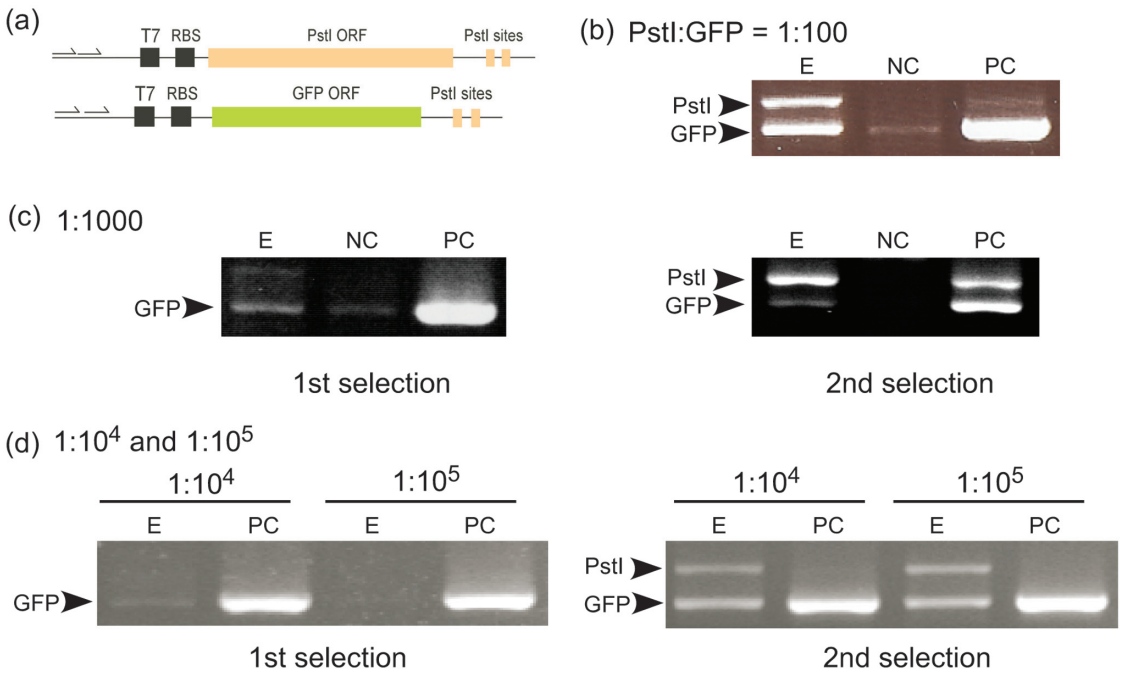

Figure 2. Model selections using a library with two DNA templates. **(a)** DNA templates used in the model libraries. **(b)** Selection after adaptor-specific PCR for the library PstI: GFP = 1:100. E: emulsion selection; NC: negative control, the library is put directly into the ligation reaction and subjected to PCR; PC: positive control, the library is first digested with the pure PstI enzyme, purified, then subjected to ligation and PCR. **(c)** Results of adaptor-specific PCR in two rounds of selection for the 1:1000 library. **(d)** Results of adaptor-specific PCR in two rounds of selection for the $1:10^4$ and $1:10^5$ libraries.

carried out by directly putting the initial library into the ligation reaction followed by PCR amplification. Since DNA templates are blunt-ended PCR products, they should not ligate to adaptors with sticky ends and thus not be amplified. A negative result in the negative control suggests there is no non-specific ligation and amplification.

Results from a single round of selection using a PstI:GFP = 1:100 library are shown in Figure 2b. In the positive control, GFP is preferentially amplified (Figure 2b, lane PC) since it is the dominant species in the starting library, and in the negative control, almost no DNA was amplified (Figure 2b, lane NC). In contrast, after emulsion selection, a bright band corresponding to the PstI template appears with comparable intensity on top of the GFP band (Figure 2b, lane E). Taken together, these experiments suggest a specific enrichment of the PstI template. Similar results were observed in the first round of selection using the 1:1000 library (Figure 2c, left panel). The final molar ratio between PstI and GFP after

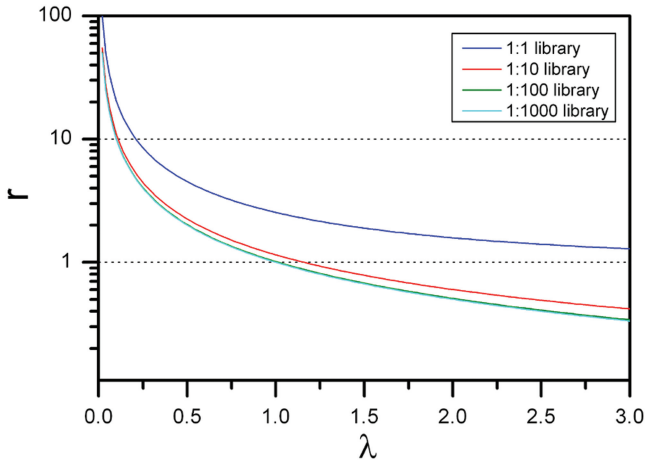

Figure 3. Theoretical analysis of selection efficiency. The Y-axis represents *r*, which is the ratio between the RE gene and the 'carryover' gene after a single round of selection; the X-axis represent λ, which is the number of DNA templates used in the selection relative to the number of droplets.

selection, as judged from the band intensities, is larger than 1:10, which implies at least a 100-fold enrichment.

During an IVC selection, all templates that reside in the same droplet with an RE gene would be amplified as 'carryover'. Thus a trade-off has to be made about the amount of DNA template present when starting a selection. On the one hand, it is ideal to use less so that there is no more than one gene in a droplet; on the other hand, less DNA often brings difficulty in sample recovery and lowers the reaction efficiency in the downstream processing. By assuming a Poisson process on the distribution of templates over the 10^{10} droplets, one can actually compute the molar ratio (*r*) between the selected endonuclease gene and the 'carryover' non-endonuclease gene after an 'ideal' selection (see derivation in Methods and Materials). The curves in Figure 3 show the predicted relationship between the ratio (*r*) and the number of DNA templates relative to the number of droplets (λ). Notice that for libraries of 1:10 or lower, the ratio after selection (*r*) is very close and largely dependent on the number of DNA templates entering the selection. The relative enrichment (*E*), however, is highest for the library 1:1000. If using a fixed amount of library 1:100 or 1:1000, for example 10^{10} ($\lambda = 1$, or ~10 ng for 1 kb DNA templates), the best *r* value after a single round of selection is ~1:1. In these cases, for the target gene to become the dominant species (as shown in Figure 2), one has to use less than 10^{10} templates or multiple rounds. Another interesting observation is that an enrichment from the original library is always predicted independent of the amount of DNA templates used. This is because the droplets only with the non-target genes are not selected. The theoretical analyses warrant a strategy of starting with a larger amount of the DNA library and gradually reducing the template amount in the later rounds.

To eliminate the 'carryover' templates, different adaptors have to be used between successive selections. Figure 2c shows the multiple rounds of selection using the 1:1000 library. After the first round, the PstI template is enriched by more than 100-fold (Figure 2c, left panel, lane E). The purified DNA after the first PCR was directly used in the next round of selection, after which the PstI template has become the dominant DNA species in the library (Figure 2c, right panel, lane E). Selections using the $1:10^4$ and $1:10^5$ libraries are shown in Figure 2d. After the first selection, very little DNA is amplified, which presumably represents 'carryover' GFP templates (Figure 2d, left panel, lane E). After the second selection, bands corresponding to both PstI and GFP appear on the gel with a ratio of ~1:1. These results suggest a consistent 100-fold enrichment in each round of selection.

Genomic selection of the PstI gene

Having established that RE genes are effectively enriched from the model libraries, we continued to challenge the system with more complex libraries. A 'real' test is to do an *in vitro* selection using libraries constructed from a bacterial genome where we know an active RE gene exists. The size of a typical bacterial genome differs from less than 1M bases to close to 10 M bases. The size of a typical Type II RE gene is ~1 kb. In principle, the expected library complexity constructed from a bacterial genomic DNA should be between 10^5 and 10^6. We first chose to select the known PstI gene from its native host *Providencia stuartii* and later a new thermostable endonuclease TspMI gene from a *Thermus sp*.

A schematic diagram for the genomic library construction is shown in Figure S1 in the Supplementary Materials. Briefly, pure genomic DNA (gDNA) was sheared to less than 5 kb fragments using a nebulizer. Fragmented gDNA was then size-selected (1k–3k), blunt-ended and phosphorylated (see Methods and Materials for details). The resulting gDNA fragments were ligated with the linearized vector which has the necessary elements for *in vitro* transcription/translation and selection. The ligated gDNA with the vector was then 'clonally' amplified by using emulsion PCR (12). The advantages of using emulsion PCR are to reduce amplification bias and increase the quality of the genomic library. Amplified linear gDNA templates were used directly in the *in vitro* selection.

During the selection process, we monitored the presence of two reference genes in the libraries before and after selection by PCR: one is the target PstI gene and the other is a fragment from the DNA methyltransferase gene M.PstII (13), which does not possess endonuclease activity and is not located in the vicinity of the PstI gene on the chromosome. Figure 4 shows the whole process of genomic selection. The starting genomic library is shown in Figure 4(a), in which both PstI and the M.PstII fragment are present. Notice that the band intensities do not necessarily reflect their proportional abundance in the genomic library since there may be differences in individual PCR efficiencies.

The gel in the upper Figure 4(b) shows the results of the first adaptor-specific PCR for the emulsified genomic library and the negative control. There is no apparent

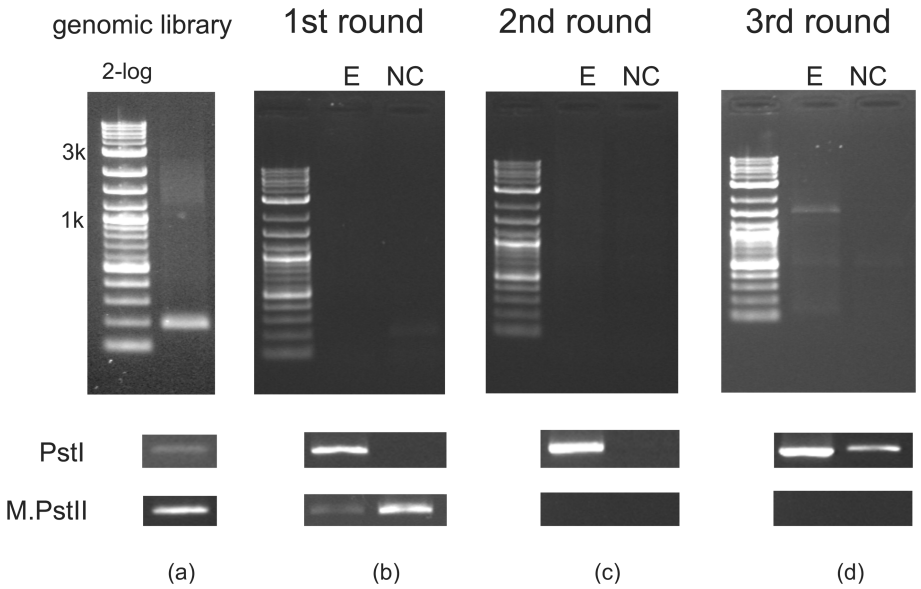

Figure 4. Genomic selection of the PstI gene from *Providencia stuartii*. **(a)** (upper) starting genomic library. The smear between 1 and 3 kb contains the amplified genomic templates after emulsion PCR. The lower band corresponds to the amplified short fragment from the empty plasmids. (lower) PCR amplification of the target PstI gene and the reference M.PstII gene in the starting library. **(b)** (upper) PCR after the first round of selection. E: emulsion selection; NC: negative control selection. (lower) PCR comparison of the PstI gene and the M.PstII gene in the samples E and NC above [same for (c) and (d)]. **(c)** PCR after the second round of selection. **(d)** PCR after the third round of selection. Notice a single band appears in lane E but is absent in lane NC. It corresponds to the selected PstI genomic fragment.

difference between the two PCRs. However, individual PCR on the two reference genes suggests that the PstI gene was enriched in the emulsified library but not in the negative control. The control gene, M.PstII, is clearly not amplified (Figure 4b, bottom). The fact that less M.PstII is present in the emulsified library than in the negative control may be due to greater DNA sample loss in the emulsion selection. Lane E in Figure 4(b) was purified and used for the next round of selection [Figure 4(c)], after which only the desired PstI gene is present in the emulsified sample and other contaminating genes, such as M.PstII, have been diluted away [Figure 4(c), bottom]. Although it seems that only PstI-bearing templates survive the second selection, it is not enough to stand out on the gel. Lane E in Figure 4(c) was purified and put into a third round of selection, after which a band of ~1.5 kb appears in the emulsified library but not in the negative control [Figure 4(d), upper]. Individual PCRs on the reference genes support a consistent enrichment in the third round. It was later confirmed that this 1.5 kb band harbors the complete PstI genomic fragment. These results strongly suggest that enrichment of the PstI gene follows the expected course.

The ~1.5 kb band from the third selection was gel-purified, digested with PstI enzyme and cloned into pLT7K for sequencing. The plasmid pLT7K is engineered to accommodate extremely toxic genes by combining controlled repression of the cloned gene and an anti-sense promoter to counter the lethal effects of basal expression (11). Ten clones were picked randomly and six of these contained DNA inserts. The sequenced inserts were compared with the fully sequenced PstI RM system (14) and the results confirm that there is one major product in the selected genomic DNA, which encompasses the full PstI open reading frame, with 3 nt upstream of the start codon and ~300 nt downstream of the stop codon. This result unambiguously shows that the selected DNA is indeed from the genomic DNA source and not from any possible contamination. Interestingly, the fact that all of the selected genomic fragments start 3 nt upstream of the PstI start codon indicates there may be selection pressure on the translation efficiency during genomic selections.

Genomic selection of the TspMI gene

We then applied the *in vitro* selection method to another thermostable endonuclease TspMI (recognition sequence C↓CCGGG) from *Thermus sp*. (10), which had not been cloned before. TspMI is optimally active at 75–80°C and retains ~20% activity at 37°C (14). Based on these facts, *in vitro* selection differs slightly from the selection of PstI gene in that: (i) in the library construction, the ligation steps between the genomic fragments and the vector are performed in the presence of NruI and MscI enzyme individually to minimize the chance that either enzyme cuts inside the TspMI gene and destroys the target gene to be selected; (ii) the emulsion reaction was first incubated at 37°C for *in vitro* transcription/translation and later moved to 65°C briefly for efficient DNA cleavage and (iii) since the TspMI enzyme cannot be deactivated by heat, only quenching buffer was used to stop the reaction and the process of DNA recovery was performed on ice. For comparison, traditional methylase selection (3) was also performed to map the genomic region harboring the TspMI RM system (YZ and RJR, unpublished data).

Figure 5 shows the adaptor-specific PCR after each round of selection using the library derived from the NruI

ligation. As a result, multiple bands are observed after the third selection. These bands were digested by XmaI (recognition sequence C↓CCGGG), an isoschizomer of TspMI, and cloned into the vector, pLT7K. Twenty clones were randomly picked and five clones contained DNA inserts. The five clones with inserts were sequenced and contain an open reading frame of ~1.1 kb. This open reading frame coincides with the endonuclease gene acquired by the traditional methylase selection approach and later was confirmed to encode the active TspMI endonuclease gene (data not shown). Analysis of the 5 sequenced clones with genomic inserts suggests that the selected genomic fragments all start 36 nucleotides downstream from the predicted start codon and end at variable sites after the stop codon, leading to the pattern of multiple bands on the agarose gel. Selection using the library derived from MscI ligation yields no bands (data not shown). It was later found that there are multiple MscI sites inside the TspMI open reading frame so that the endonuclease gene had been destroyed during the ligation step. The sequence of the TspMI RM system has been deposited in GenBank (accession number EF426476).

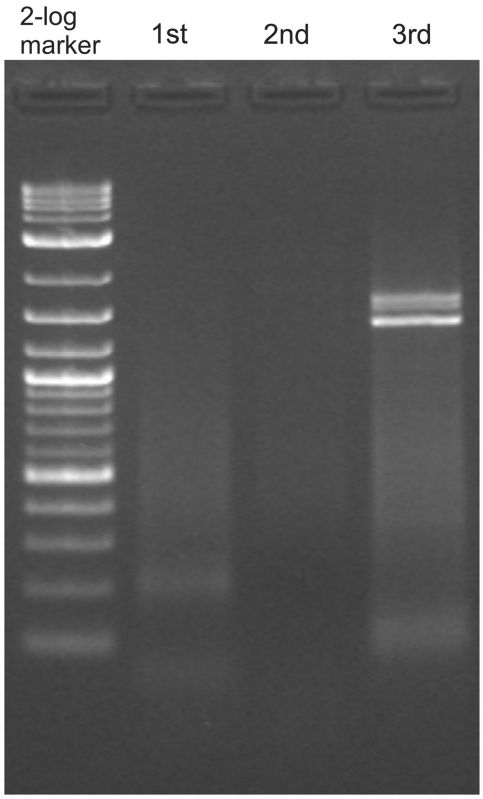

Figure 5. Genomic selection of the TspMI gene from *Thermus sp*. The lanes (1, 2, 3) show the adaptor-specific PCR amplification after each round of selection. Bands in the third lane were cloned and confirmed to encode the TspMI gene.

The TspMI restriction-modification system is interesting in several ways. It contains the usual R and M genes as well as a nicking endonuclease gene (V gene) that is often found with m5C DNA methyltransferases. These endonucleases recognize the G–T mismatches that are formed following cytosine deamination, a spontaneous event that would be mutagenic if left uncorrected. On the basis of sequence comparison, the TspMI gene appears to be a member of a new family of genes recognizing CCCGGG, since it is quite dissimilar to the known families of genes represented by SmaI and XmaI (Table 1). In REBASE there are six genes in the SmaI family and seven in the XmaI family all of which are accompanied by DNA methyltransferases that form N4-methylcytosine. In the three known cases, it is the second base in the recognition sequence that is modified and given the sequence similarity among this set, it is likely they all modify this same base. In contrast, M.TspMI is likely to be an m5C methyltransferase showing only limited similarity to M.NmeAI ($C^{m5}CGG$).

The protein sequence of TspMI is only remotely similar to BsoBI ($P > 0.1$), which is another thermostable

Table 1. Sequence families of restriction enzymes that recognize CCCGGG

Restriction enzyme	Source	Methyltransferase
SmaI (CCC↓GGG)	***Serratia marcescens* Sb**	**$C^{m4}CCGGG$**
Cli245ORF1935P	*Chlorobium limicola*	unknown
CphBORF2531P	*Chlorobium phaeobacteroides* BS1	unknown
CphORF2524P	*Chlorobium phaeobacteroides*	unknown
XcaVORF1110P	*Xanthomonas campestris* pv. *vesicatoria str*. 85-10	unknown
XveIIP	*Xanthomonas campestris* pv. *vesicatoria*	unknown
XmaI (C↓CCGGG)	***Xanthomonas malvacearum***	**unknown**
Cfr9I	***Citrobacter freundii***	**$C^{m4}CCGGG$**
Pac25I	*Pseudomonas alcaligenes*	unknown
XcyI	***Xanthomonas cyanopsidis* 13D5**	**$C^{m4}CCGGG$**
MhuORF2537P	*Methanospirillum hungatei* JF-1	unknown
StpORF334P	*Symbiobacterium thermophilum*	unknown
XaxGORFAP	*Xanthomonas axonopodis* pv. *glycines* plasmid AG1	unknown
TspMI (C↓CCGGG)	**Unidentified thermophile**	**m5C**

The methyltransferases for the top two classes of RM systems are all closely related by sequence suggesting they all modify the second cytosine residue in the recognition sequence to form N4-methylcytosine. In contrast, M.TspMI is an m5C methyltransferase showing limited similarity to M.NmeAI ($C^{m5}CGG$). Enzymes shown in bold have been characterized biochemically, the others are predicted on the basis of sequence similarity to XmaI or SmaI.

restriction enzyme and recognizes C↓YCGRG (Y = C/T, R = A/G). This would be consistent with the relaxed specificity of BsoBI, which recognizes the two sequences, CCCGAG and CTCGAG, as well as the specific sequence, CCCGGG, which is recognized by TspMI. Note that the relative position of cleavage within the recognition sequence is the same for both enzymes. Figure S2 in the Supplementary Materials shows a multiple alignment of TspMI and BsoBI together with two other related enzymes, AvaI and NspIII that also recognize C↓YCGRG. Sequence conservation between TspMI and the BsoBI family is localized in the catalytic motif EXK (shown in the red box of Figure S2) (15). One interesting observation is that the conserved histidine residue in this region of the BsoBI family, which was suggested to act as a base to deprotonate a water molecule as a nucleophile (15), is replaced by a serine residue in TspMI. This suggests a slightly different catalytic mechanism possibly through serine-mediated nucleophilic attack. Two residues in BsoBI, Asp246 and Lys81 (blue arrows in Figure S2), which were suggested to recognize degenerate base pairs and are conserved in the BsoBI family, have changed in TspMI, with aspartate conserved and lysine changed to phenylalanine. Again this suggests a slightly different base recognition mechanism, possibly in accordance with the tightened specificity of TspMI. Overall, the sequence alignment between TspMI and the BsoBI family supports the idea that TspMI should adopt a structure similar to BsoBI and is a remote homolog of the BsoBI family.

DISCUSSION

Reconstructing biological systems *in vitro* is one of the many interesting challenges in synthetic biology (16). When compared with the *in vivo* host, an *in vitro* system sometimes offers unique advantages, for instance, it is ideal for toxic genes such as restriction endonucleases for which approaches using living hosts have proven to be difficult (2). Minimal *in vitro* systems are in general more configurable and reaction intermediates are more accessible, making them more amenable to engineering. For the genetic selection of DNA or RNA, without the barrier of a cell membrane and the limitations of transformation, an *in vitro* system is capable of exploring even larger libraries and functionalities.

In this article, by using *in vitro* compartmentalization to generate myriad aqueous droplets in oil as artificial cells, we are able to selectively amplify RE genes from bacterial genomes. The so-called 'artificial cells' themselves do not undergo Darwinian selection, but simply provide a means to link genotype and phenotype. The selection itself requires a method to distinguish those genotypes that have been changed from those that have not. With over 100-fold enrichment in each round of selection, typically three rounds are needed to enrich a specific gene to 'homogeneity' from a bacterial genome. From the genomic selection of PstI, we observe that actually most of the contaminating genes have been removed from the library after the second round of selection. We used inverse PCR and DNA sequencing to determine the ends of the selected genomic fragments containing the PstI gene after the second round selection, and found that there is only one variant present in the library at that point, which was later amplified in the third round (data not shown). This contradicted our intuition that many different templates encompassing the PstI gene would be selected and the final result would be a DNA smear on the gel. It can be explained from several perspectives. First, it seems that there is a strong selection pressure on the translation efficiency of the DNA templates, which gives a selective advantage to those templates with ribosome-binding sites very close to the start codon of the target gene for efficient translation. In the genomic selection of the PstI gene, one end of the selected DNA fragment is just 3 nt upstream of the start codon. On the other hand, the 'substrate' ends of the templates provide little influence on the translation efficiency and is presumably less strictly selected. This is supported by the TspMI selection, in which the selected genomic fragments end at variable points after the stop codon as far as ~300 nt apart. However, we have no explanation for the lack of variability observed in the PstI genomic selection. Second, it may result from the non-randomness of the shearing process using the nebulizer, in which strand breakage is heavily influenced by the local AT content of the genomic DNA (YZ and Chudi Guan, to be published).

Doi, N. *et al*. have previously applied IVC in selecting RE genes (4). Their method uses a DNA polymerase to incorporate dUTP-biotin to the sticky ends generated by the restriction endonuclease, permitting affinity-based purification of the genes. Using this method, they were only able to obtain a selection efficiency of ~10-fold in a single round. This is likely due to the fact that any DNA fragments that might have resulted from non-specific cleavage in the compartments could have become labeled and hence, selected. Due to this relatively low efficiency, more iterations are required to recover the desired genotype, which severely limits the potential applications. For instance, six rounds of selections are needed to select active FokI gene from a randomized FokI library at three codon positions (expected library complexity of ~8000). In contrast, our method exploits the full potential of the sequence specificity available at the sticky end by requiring ligation of the adaptor. Non-specific cleavage products or damaged ends would not result in amplification. This results in much higher enrichment during each round of selection and experimentally we find that greater than 100-fold enrichment can be obtained. This has allowed us to select genes from genomic libraries and would also permit a much greater sampling of sequence space during the selection of mutants.

When compared with the conventional methylase selection, which often requires a purified endonuclease and the selection process targets the DNA methylase, the *in vitro* method directly targets the RE genes in the library and merely requires a specified recognition site and a set of DNA adaptors. Thus, it should provide a possible route for searching environmental DNA samples for RE genes with desired specificities, which in principle contains a much greater genetic diversity from many co-existing microbial species.

Table 2. Statistics of restriction endonuclease genes having their own recognition sites within their coding sequences*

Recognition site length	Total number	Number of genes w/sites (percentage)			
		0 site	1 site	2 sites	More than 2 sites
4 base	117	64(55%)	32 (27%)	14(12%)	7(6%)
5 base	80	52(65%)	19(23%)	6(8%)	3(4%)
6 base	157	138(88%)	14(9%)	4(3%)	1(1%)
7 base	15	11(73%)	4(23%)	0	0
8 base	10	10(100%)	0	0	0

*All sequence data were retrieved from REBASE (14) as of February 2007. Only experimentally verified RE genes are included in the analysis. Degenerate bases, 'RYMKSW' (e.g. R = A or G) are counted as 0.5 base; while 'BDHV' (e.g. B = C or G or T) are counted as 0.25 base.

There are possible limitations with the current approach. Because the ligation is the key step in selection, it may be less effective to select endonucleases which generate shorter overhangs or blunt cuts. This may be alleviated by placing the recognition site of a nicking enzyme close to the blunt cut site and use the nicking enzyme to convert the blunt end to a suitable sticky end (17). Frequent cutters, such as those recognizing 4-base sites, sometimes fall outside of the application range since they tend to destroy their own genes. These enzymes are completely fine in living bacteria since there is always a companion DNA methyltransferase to protect the host. Nevertheless, it appears that the selective disadvantage of having self-destructing sites has driven a significant proportion of frequent cutters to lose the recognition sites within their genes. Table 2 lists the statistics of those RE genes having their own recognition sites within their genes. For example, for a gene of 1kb in size, the probability that it does not have a particular 4-base site is $\sim$0.0004 [i.e. $(1\text{-}1/128)^{1000}$]. This sharply contrasts with the observation that over half of the 4-base-recognizing RE genes do not have their own sites in their coding sequences (Table 2).

The *in vitro* method described here has the potential to be extended to other applications, with the most relevant being the directed evolution of genes encoding enzymes (4). This requires both extreme specificity and sensitivity in the selection method to allow an efficient search in the vast sequence space. In fact, even with the ability to select from a 10^{10} library, one can only possibly vary 6–7 codons with saturation. Nonetheless, numerous directed evolution experiments suggest that sometimes only a few amino acid substitutions could bring considerable changes in biochemical properties. For highly diverse libraries, theoretical derivations under simplified assumptions in this article support a reasonable strategy of using relatively large amounts of template in the early rounds of selections, with sacrificed specificity but presumably a high sensitivity, and decreasing amounts of the library in later rounds with improved specificity.

The method we describe in this article depends on our ability to select the desired genotype based on a specific alteration caused by its translated phenotype. The reconstituted *in vitro* transcription/translation system plays a crucial role: it is free of many unwanted constraints that are often lethal when using living hosts, and it offers considerable modularity for potential engineering. A similar approach should allow the isolation of mutants with specifically desired properties, such as altered cleavage positions that might result in novel sticky ends, improved thermal stability by selecting at a desired temperature as well as numerous other properties that might increase the practical utility of these enzymes. In all of these cases, the selections can be done on bulk preparations thanks to the DNA-modifying nature of these enzymes. For a wide variety of other useful enzymes, it is desirable to be able to interrogate individual droplets by using some novel emulsion formulations (18) or by flow cytometry or microfluidics (8) in a high-throughput way, both of which have shown promise.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We sincerely thank Dr. Andrew Griffiths and members of his lab at MRC LMB and Institut de Science et d'Ingénierie Supramoléculaires, Strasbourg, France for their help in teaching us the water-in-oil emulsion technology. We thank our colleagues at NEB for their help during the course of this work. We also thank the reviewers for their helpful suggestions. This work is supported by New England BioLabs, Inc. Funding to pay the Open Access publication charge was provided by New England Biolabs.

Conflict of interest statement. None declared.

REFERENCES

1. Roberts,R.J. (2005) How restriction enzymes became the work-horses of molecular biology. *Proc. Natl Acad. Sci. U S A*, **102**, 5905–5908.
2. Alves,J. and Vennekohl,P. (2004) Protein engineering of restriction enzymes. In Pingoud,A. (ed), *Nucleic Acids and Molecular Biology*, Spring-Verlag, Berlin Heidelberg, Vol. 14, pp. 393–411.
3. Szomolanyi,E., Kiss,A. and Venetianer,P. (1980) Cloning the modification methylase gene of *Bacillus sphaericus* R in *Escherichia coli*. *Gene*, **10**, 219–225.
4. Doi,N., Kumadaki,S., Oishi,Y., Matsumura,N. and Yanagawa,H. (2004) *In vitro* selection of restriction endonucleases by *in vitro* compartmentalization. *Nucleic Acids Res.*, **32**, e95.

5. Tawfik,D.S. and Griffiths,A.D. (1998) Man-made cell-like compartments for molecular evolution. *Nat. Biotechnol.*, **16**, 652–656.
6. Roberts,R.W. and Szostak,J.W. (1997) RNA-peptide fusions for the *in vitro* selection of peptides and proteins. *Proc. Natl Acad. Sci. U S A*, **94**, 12297–12302.
7. Mattheakis,L.C., Bhatt,R.R. and Dower,W.J. (1994) An *in vitro* polysome display system for identifying ligands from very large peptide libraries. *Proc. Natl Acad. Sci. U S A*, **91**, 9022–9026.
8. Griffiths,A.D. and Tawfik,D.S. (2006) Miniaturising the laboratory in emulsion droplets. *Trends Biotechnol.*, **24**, 395–402.
9. Shimizu,Y., Inoue,A., Tomari,Y., Suzuki,T., Yokogawa,T., Nishikawa,K. and Ueda,T. (2001) Cell-free translation reconstituted with purified components. *Nat. Biotechnol.*, **19**, 751–755.
10. Parashar,V., Capalash,N., Xu,S.Y., Sako,Y. and Sharma,P. (2006) TspMI, a thermostable isoschizomer of XmaI (5'C/CCGGG3'): characterization and single molecule imaging with DNA. *Appl. Microbiol. Biotechnol.*, **72**, 917–923.
11. Kong,H., Lin,L.F., Porter,N., Stickel,S., Byrd,D., Posfai,J. and Roberts,R.J. (2000) Functional analysis of putative restriction-modification system genes in the *Helicobacter pylori* J99 genome. *Nucleic Acids Res.*, **28**, 3216–3223.
12. Williams,R., Peisajovich,S.G., Miller,O.J., Magdassi,S., Tawfik,D.S. and Griffiths,A.D. (2006) Amplification of complex gene libraries by emulsion PCR. *Nat. Methods*, **3**, 545–550.
13. Sears,A., Peakman,L.J., Wilson,G.G. and Szczelkun,M.D. (2005) Characterization of the Type III restriction endonuclease PstII from *Providencia stuartii*. *Nucleic Acids Res.*, **33**, 4775–4787.
14. Roberts,R.J., Vincze,T., Posfai,J. and Macelis,D. (2007) REBASE—enzymes and genes for DNA restriction and modification. *Nucleic Acids Res.*, **35**, D269–270.
15. van der Woerd,M.J., Pelletier,J.J., Xu,S. and Friedman,A.M. (2001) Restriction enzyme BsoBI-DNA complex: a tunnel for recognition of degenerate DNA sequences and potential histidine catalysis. *Structure*, **9**, 133–144.
16. Forster,A.C. and Church,G.M. (2007) Synthetic biology projects *in vitro*. *Genome Res.*, **17**, 1–6.
17. Samuelson,J.C., Zhu,Z. and Xu,S.Y. (2004) The isolation of strand-specific nicking endonucleases from a randomized SapI expression library. *Nucleic Acids Res.*, **32**, 3661–3671.
18. Mastrobattista,E., Taly,V., Chanudet,E., Treacy,P., Kelly,B.T. and Griffiths,A.D. (2005) High-throughput screening of enzyme libraries: *in vitro* evolution of a beta-galactosidase by fluorescence-activated sorting of double emulsions. *Chem. Biol.*, **12**, 1291–1300.