SHORT TECHNICAL REPORTS

Manuscripts published in the Short Technical Reports section are shorter and less comprehensive in scope than full Research Reports.

Semi-Multiplex PCR Technique for Screening of Abundant Transcripts During Systematic Sequencing of cDNA Libraries

BioTechniques 21:644-649 (October 1996)

ABSTRACT

The systematic sequencing of cDNA libraries is an efficient approach for the identification of new genes, but the presence of abundant mRNAs is often a major problem. This paper describes a very simple method of "semi-multiplex PCR" that allows specific identification of such abundant transcripts before DNA sequencing without using nonrepresentative subtracted libraries. The PCR utilizes a series of forward primers specific for abundant transcripts with a pair of universal primers used for template generation. cDNA clones corresponding to abundant mRNAs are then revealed by double bands in agarose gel.

INTRODUCTION

The systematic sequencing of libraries of complementary DNA (cDNA) is becoming a common approach for the rapid identification and cataloguing of genes in a variety of organisms (1,6,18, 20,21). The general strategy of such approaches is based on "single-pass" systematic sequencing of random cDNA clones, which results in the production of short partial sequences generally known as expressed sequence tags (ESTs) (2). In some cases (e.g., human brain, Reference 3), the number of new genes identified can be very high due to a reduced number of redundant transcripts, which may be more apparent than real as a result of the adopted strategy. But in other cases, the presence of highly represented tissue-specific mRNAs in strictly committed organs like muscle (16), heart (10), liver (11) and pancreas (17,19) would greatly hamper the identification of rare transcripts by a random sequencing approach. To overcome this obstacle, several strategies have been developed such as the construction of normalized libraries (7,12), the subtraction of common mRNA before copying it into cDNA (14,15) or the prescreening of libraries with probes for the most abundant mRNAs by large-scale dot blot analysis (9,10). The use of normalized libraries seems a good solution, but this will not allow the study of the relative expression levels of the genes active in a certain tissue or developmental stage. Subtraction protocols are normally based on several rounds of hybridization of the mRNA to be cloned with a competitor antisense single-stranded cDNA followed by the elimination of cDNA-mRNA hybrids (5,15). Often, the amount of the remaining non-hybridized mRNA is too little to generate a library and can only be used as probe for screening a preexisting full-scale cDNA library (14). Furthermore, in theory, using subtractive or normalizing strategies, one cannot be sure that during the hybridization process, rare transcripts or slightly different forms of abundant mRNAs would be lost. Finally, the screening of large number of cDNA clones by hybridization with probes is time-consuming and demands a good calibration of hybridization conditions to avoid misleading results.

We have prepared a cDNA library from human adult skeletal muscle in order to study its transcription profile by a random sequencing approach, and the results of this analysis are presented elsewhere (9). After having sequenced the first 1000 clones, we realized that in this tissue there is a small group of highly transcribed genes. This paper presents a prescreening test for such most-abundant trancripts and its application to our project of systematic sequencing of ESTs from human skeletal muscle. The screening method is based on a semi-multiplex polymerase chain reaction (PCR), which uses a combination of "interference" primers identifying abundant transcripts with a pair of universal primers adopted for template production. In agarose gel, the PCR of abundant cDNAs show double bands (one generated by the universal primers and the second by the specific interference primer), allowing their immediate identification and recording before

sequencing.

MATERIALS AND METHODS

Primer Selection

All the primers were derived from the consensus sequence obtained by aligning all the EST homologues to the same transcript, and the sequences were taken from a region close to the 3′ end of the consensus in order to maintain the hybridization with the shortest cDNA clones of the library. The set of primers used for the semi-multiplex PCR has been designed with the help of a computer program called SM-PCR, especially written for this purpose and available by anonymous file transfer protocol (ftp) at eos.bio.unipd.it. Given a set of *n* target sequences, the program tries to find acceptable sets of oligonucleotides (one per each target sequence) that satisfy the restrictions indicated in the following list, where *magic, oligolen, unbal, e2, e3, e4, e5, d* and *s* are parameters defined by the user: *(i)* the last *magic* bases at the 3′ end of each oligonucleotide should not find any inverse-complementary sequence in any oligonucleotide of the set; *(ii)* the length of the oligonucleotides should be *oligolen* bases; *(iii)* the maximal number of G+C or A+T should not be more than *unbal*; *(iv)* the last *e2* bases should not contain more than *e3* G+C; *(v)* the last *e4* bases should not contain more than *e5* G+C; *(vi)* not more than *d* contiguous G+C or A+T are allowed; and *(vii)* not more than *s* contiguous repeat of the same bases are allowed. The primer sets are then checked manually for homogenicity of melting temperatures using the program OLIGO™ Version 4.0 (National Biosciences, Plymouth, MN, USA). All the oligonucleotides described in this paper were purchased from Genset (Paris, France).

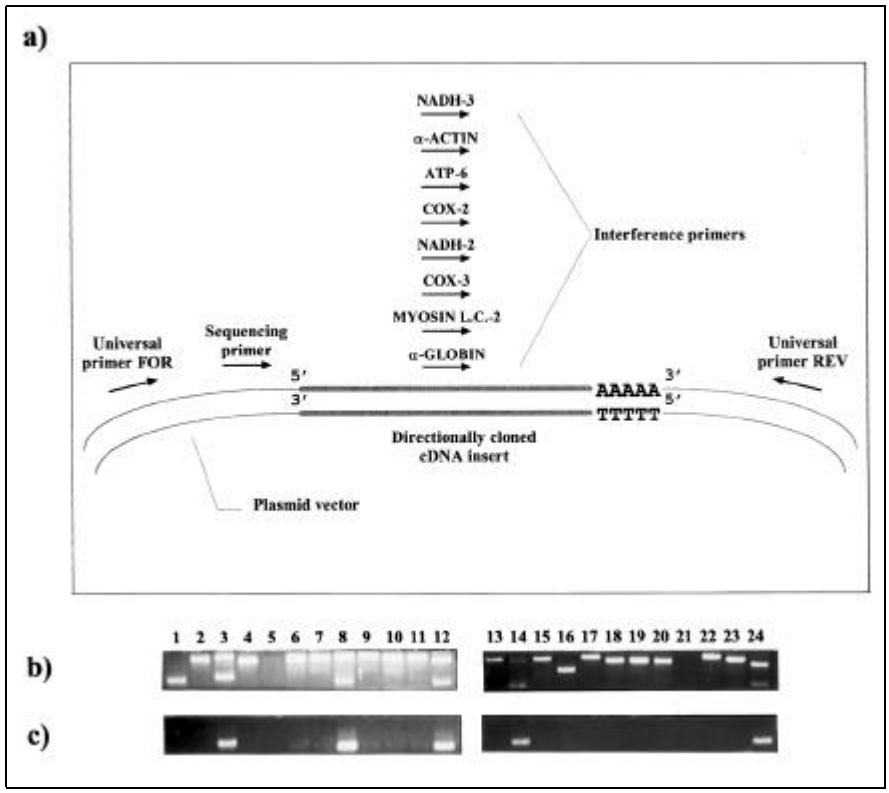

Figure 1. PCR strategy and agarose gel test of semi-multiplex PCR. (a) A schematic representation of the semi-multiplex PCR technique. Arrows indicate the approximate position and orientation of the eight interference primers, of the two universal primers and of the primer used for the sequencing reaction relative to a generic cDNA insert of the human skeletal muscle library. (b) Agarose gel test of the semi-multiplex PCR. Different cDNA clones are amplified using the two universal primers FOR and REV plus the eight interference primers shown in (a). The amplifications of cDNA clones belonging to the most abundant transcripts (lanes 3, 8, 12, 14 and 24) result in two bands: the longer is generated by the universal FOR and REV primers and the shorter by the specific internal interference primer and the universal primer REV. Lane 1 is a PCR of a self-ligated vector used as a marker control, and lanes 5 and 21 contain unsuccessful amplifications. (c). The same cDNA clones were amplified with the primer mixture of (b) but omitting the universal primer FOR to control the specificity of the interference primers for the generation of double bands. Note that the specific bands appearing in lanes 3, 8, 12, 14 and 24 correspond to the faster bands present in the same lanes of Panel b.

Insert Amplification, Template Purification and Sequencing

The human skeletal muscle cDNA library consists of 3′-end specific cDNA fragments sonicated and size-selected in the range of 450–550 bp. These cDNA tags are directionally cloned in the plasmid vector pcDNA II (Invitrogen, Leek, The Netherlands) and carried by the TOP10F′ bacterial strain (Invitrogen). Individual recombinant bacterial clones are picked up after overnight growth on Petri dishes containing SOB medium with 1.5% Bacto® Agar (Difco, Detroit, MI, USA) and 100 µg/mL ampicillin and transferred to Thermowell® Model G 96-well plates (Corning Costar, Cambridge, MA, USA). Fifty microliters of 1× PCR buffer (20 mM Tris-HCl, pH 8.3, 50 mM KCl, 2 mM $MgCl_2$ and 0.1% vol/vol Tween®-20) are then added to each well, and bacteria are lysed at 95°C for 10 min. Ten microliters of such "microlysates" are transferred to fresh 96-well plates containing 40 µL of amplification mixture composed by 0.126 mM of each of the four dNTPs, 3.5 pmol of each of the ten primers (one forward and one reverse universal primer plus the eight interference primers, see the Results section for details) and 1 U of *Taq* DNA Polymerase (Pharmacia Biotech, Uppsala, Sweden) in 1× PCR buffer. Amplifications are carried out in the GeneAmp® PCR System 9600 Thermal Cycler (Perkin-Elmer, Norwalk, CT, USA) set at the following parameters: 95°C for 2 min, then 30 cycles of denaturation (95°C), annealing (52°C) and elongation (72°C) steps each held for 20 s and a final incubation step at 72°C for 3 min. Ten-microliter samples are then checked for the presence of double bands by electrophoresis on 2% agarose gels containing ethidium bromide. Ten microliters of successful PCRs are

Table 1. Redundant Transcripts in Human Skeletal Muscle mRNA and Set of PCR Primers

Sequence Name	Frequency		Interference Primer	T_m
NADH-dehydrogenase chain	10.0	FOR	5′-GTCTGGCCTATGAGTGACTACAAA-3′	52.1°C
α-Actin	8.5	FOR	5′-AACTTCCGTTGCTGCCATCGTAAA-3′	60.6°C
ATPase-6	7.3	FOR	5′-CGCTGTCGCCTTAATCCAAGCCTA-3′	62.3°C
Cytochrome oxydase chain	4.4	FOR	5′-GGGGTATACTACGGTCAATGCTCT-3′	54.6°C
NADH dehydrogenase chain	3.6	FOR	5′-GCCTTACCACGCTACTCCTACCTA-3′	54.9°C
Cytochrome oxydase chain	3.1	FOR	5′-ACTATCTGCTTCATCCGCCAACTA-3′	56.1°C
α-Globin	2.6	FOR	5′-CCGCACATCCTACAGTTGGAAATC-3′	58.5°C
Myosin light chain-2	2.1	FOR	5′-TGAGCACCGTGCTGACCTCCAAAT-3′	63.5°C
Total most frequent ESTs	41.6		UNIVERSAL PRIMER	
Other ESTs	58.4	FOR	5′-GTTGTAAAACGACGGCCAGTGAAT-3′	58.1°C
Total	100.0	REV	5′-TCCGGCTCGTATGTTGTGTGGAAT-3′	61.2°C

The eight most frequent transcripts and their relative abundance in the human skeletal muscle cDNA library used in this work are shown in columns 1 and 2, respectively. Five mitochondrial genes and three nuclear genes account for the 41.6% of the total muscle mRNA population. Sequences and melting temperatures (T_m) of the eight specific interference primers and of the two universal primers used in the semi-multiplex PCR technique are reported in columns 3 and 4, respectively. This combination of primers was generated by the computer program SM-PCR set at the following parameters: *magic* = 3, *unbal* = 11, *oligolen* = 24, *e2* = 3, *e3* = 1, *e4* = 5, *e5* = 3, *d* = 5 and *s* = 4 (see Materials and Methods).

then purified either by ultrafiltration through Microcon™-100 concentrators (Amicon, Beverly, MA, USA) according to the product specifications or by digestion with 4 U of shrimp alkaline phosphatase and 10 U of exonuclease I (Amersham International, Little Chalfont, Bucks, England, UK) at 37°C for 15 min followed by heat-inactivation of the enzymes at 80°C for 15 min. Purified templates are processed in a cycle-sequencing reaction with AmpliTaq® DNA Polymerase (Perkin-Elmer) using the Taq DyeDeoxy Terminator Kit or the Taq Dye Primer Cycle Sequencing Kit (Perkin-Elmer/Applied Biosystems Division [PE/ABI], Foster City, CA, USA). The sequencing protocol consists of a step at 95°C for 2 min, 25 cycles each composed by three steps at 96°C for 15 s (denaturation), 50°C for 1 s (annealing), 72°C for 4 min (elongation) and a final incubation at 72°C for 3 min. The sequencing primer (5′-CTCGGATCCACTAGTAACG-3′) is located 21 bases upstream from the first nucleotide of the cDNA insert and 113 bases downstream from the position of the forward universal primer used for amplification. The sequencing reactions are analyzed by Models 373A and 377 Automatic DNA Sequencers (PE/ABI). Homology searches of the edited sequences are done using BLAST (4) against the current GenBank® database and FASTA (13) against our skeletal muscle EST database.

RESULTS

The strategy used for the construction of the human muscle cDNA library and the result of its large-scale sequencing are presented elsewhere (9). Briefly, this library contains inserts 450–550-bp-long corresponding to the 3′ end of the muscle mRNAs cloned directionally into plasmid vectors. cDNA clones are randomly chosen and sequenced, and the resulting ESTs are searched against current databases and assembled in groups identifying diverse transcripts. The relative frequencies of the eight most abundant transcripts as they result from the analysis of 4500 independent ESTs are listed in Table 1. As can be seen, five mRNAs of mitochondrial origin and three mRNAs transcribed from the nuclear DNA account for the 41.6% of the total skeletal muscle mRNA population. To avoid the continuous sequencing of such ESTs, we designed a semi-multiplex PCR technique that allows the identification of abundant inserts in agarose gel electrophoresis. With the help of a computer program specially designed for this purpose (see Materials and Methods), we designed a set of eight primers specific for the most abundant mRNAs. Their sequence is reported on Table 1 together with the sequence of the universal primers used for template production.

The rationale of this semi-multiplex PCR technique is illustrated in Figure 1a. Since all the specific interference primers are forward primers, they will compete with the universal forward primer in the amplification reaction. The amplification of a cDNA clone belonging to the most abundant groups will therefore result in two bands of different molecular weight: *(i)* the heavier generated by the universal forward and reverse primers and *(ii)* the lighter generated by the specific forward primer and the universal reverse primer. Figure 1b shows a typical result of this semi-multiplex PCR test on an agarose gel. In Figure 1c, as a control for the shorter specific bands, the same templates were amplified under identical conditions but omitting the forward universal primer. A test for the validation of the tech-

nique was done on 192 independent cDNA clones. They were amplified with the mixture of ten primers as described in Materials and Methods, aliquots of the PCR products were run in agarose gels containing ethidium bromide and the results (no bands, single band corresponding to self-ligated vector, single band corresponding to non-abundant ESTs and double bands corresponding to abundant ESTs) were recorded. Aliquots of the 181 successful PCRs were used for sequencing. Table 2 summarizes the results of this test. As can be seen from the comparison between the agarose gel and sequencing results, the method is quite reliable because only one abundant EST did not show double bands and, more important, all the PCRs appearing as single bands represented non-abundant sequences. The single false-negative amplification should not be considered as a major problem because it will be processed in any case for sequencing.

Finally, we wanted to assess the possibility of increasing the number of primers in the amplification reaction to see if the detection of an abundant EST by its specific primer would still be possible. The results of such a test are presented in Figure 2a, where two different abundant cDNA clones were amplified using the specific primer mixture described in Table 1 with the addition of 10 or 20 nonspecific 21-mers. As can be seen, even in the largest primer mixture, the generation of double bands is still detectable, and the control lanes (Figure 2b) demonstrate that the smaller specific band is generated by the specific primer.

Table 2. Validation of the Semi-Multiplex PCR Technique

PCR Results	
Recombinant clones amplified	192
Negative PCR	11
Positive PCR	181
PCR band of 342 bp (= plasmid without insert)	29
PCR with double bands	38
PCR with single bands	114
Sequencing Results	
Positive PCR templates sequenced	181
Sequences corresponding to plasmid vector	29
Sequences corresponding to the top eight ESTs	39
Sequences corresponding to other ESTs	113
Correspondence Between PCR and Sequencing Results	
Sequences corresponding to plasmid vector whose PCRs result in the expected 342-bp bands	29/29
Sequences corresponding to the top eight ESTs whose PCRs result in double bands	38/39
Sequences corresponding to other ESTs whose PCRs result in single bands	113/113

The accuracy of the technique was assessed on 192 independent cDNA clones. Each clone was amplified, checked on agarose gel and sequenced. See Results for details.

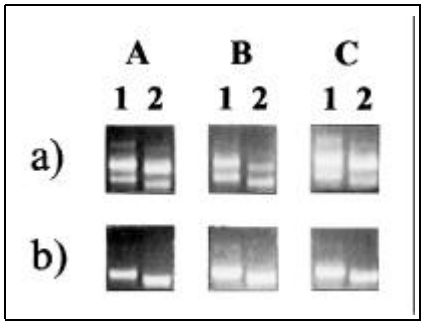

Figure 2. Potentiality of the semi-multiplex PCR. (a) To test the possibility of using a higher number of interference primers in the amplification reactions, the PCR products of two different abundant cDNA clones were run in lanes 1 (ATPase 6) and 2 (COX 2) after amplification with the mixture of the ten primers shown in Figure 1a alone (lanes A) or with the addition of 10 (lanes B) or 20 (lanes C) nonspecific 21-mers. Specific interference bands are still visible also when a mixture of thirty 21-mers are used. (b) Control of binding of the specific interference primers. The same samples were used for PCR as in (a) but omitting the universal FOR primer.

DISCUSSION

We have developed a semi-multiplex PCR protocol for the rapid identification of specific clones during sequencing of cDNA libraries. The method is simple and reliable and is performed simultaneously with the amplification step for template generation before sequencing. We tested several different sources of thermostable DNA polymerases to obtain unequivocal and reproducible results in agarose gel, and the enzyme described in Materials and Methods is now routinely used in our screening protocol. Surprisingly, the use of DNA polymerases with no 5′–3′ exonuclease activity (8) resulted in poor amplification of double bands, even though theoretically these should protect the binding and elongation from the internal interference primer. A normal thermostable DNA polymerase works well in this test probably because the band generated by the universal primers and the band generated by the specific primers are similarly short and do not compete effectively in the amplification reaction.

The result shown in Figure 2 suggests that the technique could allow the use of a higher number of interference primers. We are now testing a combination of 30 specific primers generated by our computer program SM-PCR. In our hands, the presence of such a high number of primers in the PCR is not a problem for the production of good quality templates for the sequencing reaction. Using the two procedures described in Materials and Methods for cleaning the amplification templates, we routinely obtain 350–400-bp readable sequences with less than 1% of ambiguous positions using both the dye-dideoxy terminator or the dye-primer chemistry on PE/ABI sequencers.

The method presented in this paper should be used for the detection of

abundant transcripts during systematic sequencing of cDNA libraries, but it could also be useful for the simultaneous identification of many specific cDNAs as an alternative to several rounds of library screening with specific probes or antibodies. Finally, we think that this method could be fully automated by using fluorescent-specific interference primers generated with the Applied Biosystems PRISM™ multicolor labeling technology (PE/ABI). In this case, the agarose gel test would be substituted by an automated detection with a luminescence spectrometer, thus reducing the possibility of errors in the manual inspection of PCR results.

REFERENCES

1. **Adams, M.D., M. Dubnick, A.R. Kerlavage, R. Moreno, J.M. Kelley, T.R. Utterback, J.W. Nagle and J.C. Venter.** 1992. Sequence identification of 2,375 human brain genes. Nature *355*:632-634.
2. **Adams, M.D., J.M. Kelley, J.D. Gocayne, M. Dubnick, M.H. Polymeropoulos, H. Xiao, C.R. Merril, A. Wu, B. Olde, R. Moreno, A.R. Kerlavage, W.R. McCombie and J.C. Venter.** 1991. Complementary DNA sequencing: expressed sequence tags and human genome project. Science *252*:1651-1656.
3. **Adams, M.D., A.R. Kerlavage, C. Fields and J.C. Venter.** 1993. 3,400 new expressed sequence tags identify diversity of transcripts in human brain. Nature Genet. *4*:256-267.
4. **Altschul, S.F., W. Gish, W. Miller, E.W. Myers and D.J. Lipman.** 1990. Basic local alignment search tool. J. Mol. Biol. *215*:403-410.
5. **Gieser, L. and A. Swaroop.** 1992. Expressed sequence tags and chromosomal localization of cDNA clones from a subtracted retinal pigment epithelium library. Genomics *13*:873-876.
6. **Höfte, H., T. Desprez, J. Amselem, H. Chiapello, M. Caboche, A. Moisa, M-F. Jourjon, L. Charpenteau, et al.** 1993. An inventory of 1,152 expressed sequence tags obtained by partial sequencing of cDNA from *Arabidopsis thaliana*. Plant J. *4*:1051-1061.
7. **Ko, M.S.H.** 1990. An "equalized cDNA library" by the reassociation of short double-stranded cDNAs. Nucleic Acids Res. *18*:5705-5711.
8. **Kong, H.M., R.B. Kucera and W.E. Jack.** 1993. Characterization of a DNA polymerase from the hyperthermophile archaea *Thermococcus litoralis*. J. Biol. Chem. *268*:1965-1975.
9. **Lanfranchi, G., T. Muraro, F. Caldara, B. Pacchioni, A. Pallavicini, D. Pandolfo, S. Toppo, S. Trevisan, S. Scarso and G. Valle.** 1996. Identification of 4,370 expressed sequence tags (ESTs) from a 3′-end specific cDNA library of human skeletal muscle by DNA sequencing and filter hybridization. Genome Res. *6*:35-42.
10. **Liew, C.C., D.M. Hwang, Y.W. Fung, C. Laurenssen, E. Cukerman, S. Tsui and C.Y. Lee.** 1994. A catalogue of genes in the cardiovascular system as identified by expressed sequence tags. Proc. Natl. Acad. Sci. USA *91*:10645-10649.
11. **Okubo, K., N. Hori, R. Matoba, T. Niiyama, A. Fukushima, Y. Kojima and K. Matsubara.** 1992. Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. Nature Genet. *2*:173-179.
12. **Patanjali, S.R., S. Parimoo and S.M. Weissman.** 1991. Construction of a uniform-abundance (normalized) cDNA library. Proc. Natl. Acad. Sci. USA *88*:1943-1947.
13. **Pearson, W.R. and D.J. Lipman.** 1988. Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. USA *85*:2444-2448.
14. **Sargent, T.D.** 1987. Isolation of differentially expressed genes. Methods Enzymol. *152*:423-431.
15. **Sargent, T.D. and I.B. Dawid.** 1983. Differential gene expression in the gastrula of *Xenopus laevis*. Science *222*:135-139.
16. **Sikela, J.M. and C. Auffrey.** 1993. Finding new genes faster than ever. Nature Genet. *3*:189-191.
17. **Takeda, J., H. Yano, S. Eng, Y. Zeng and G.I. Bell.** 1993. A molecular inventory of human pancreatic islets: sequence analysis of 1,000 cDNA clones. Hum. Mol. Genet. *2*:1793-1798.
18. **Uchimiya, H., S. Kidou, T. Shimazaki, S. Aotsuka, S. Takamatsu, R. Nishi, H. Hashimoto, Y. Matsubayashi, N. Kidou, M. Umeda and A. Kato.** 1992. Random sequencing of cDNA libraries reveals a variety of expressed genes in cultured cells of rice (*Oryza sativa* L.). Plant J. *2*:1005-1009.
19. **Velculescu, V.E., L. Zhang, B. Vogelstein and K.W. Kinzler.** 1995. Serial analysis of gene expression. Science *270*:484-487.
20. **Wan Kim, C., P. Markiewicz, J.J. Lee, C.F. Schierle and J.H. Miller.** 1993. Studies of hyperthermophile *Termotoga maritima* by random sequencing of cDNA and genomic libraries. J. Mol. Biol. *231*:960-981.
21. **Waterson, R., C. Martin, M. Craxton, C. Huynh, A. Coulson, L. Hillier, R.K. Durban, P. Green, et al.** 1992. A survey of expressed genes in *Caenorhabditis elegans*. Nature Genet. *1*:960-981.

This work was financed by Telethon-Italy (Grant B.30).

Address correspondence to Gerolamo Lanfranchi, CRIBI Biotechnology Centre, Università degli Studi di Padova, Via Trieste 75, 35121 Padova, Italy. Internet: lanfra@eos.bio.unipd.it

Received 30 October 1995; accepted 14 May 1996.

B. Pacchioni, S. Trevisan, S. Gomirato, S. Toppo, G. Valle

and G. Lanfranchi
CRIBI Biotechnology Centre
Università degli Studi di Padova
Padova, Italy

Cell Surface Display of a Single-Chain Antibody for Attaching Polypeptides

BioTechniques 21:650-658 (October 1996)

ABSTRACT

To provide an efficient means of coupling proteins, peptides and other suitable moieties to cells, we have constructed a retroviral expression vector for cell surface display of a single-chain antibody (scFv) against the hapten 4-ethoxymethylene-2-phenyl-oxazoline-5-one (phOx). The hapten phOx can be easily conjugated to primary amino and sulfhydryl groups, thus providing points of attachment for the cell surface-bound anti-phOx scFv. This universal cell coupling system could prove to be particularly useful for anchoring monoclonal antibodies for tumor targeting and to present co-stimulatory molecules and other ligands (even mixtures) at the cell surface for gene therapy.

INTRODUCTION

To develop a tumor vaccine, we have been exploring generally applicable methods for attaching co-stimulatory molecules to the surface of tumor cells in order to increase their immunogenicity. In a previous approach, for example, we constructed a single-chain antibody (scFv) containing a free cysteine near its C-terminus against hemagglutinin-neuraminidase in the envelope of Newcastle disease virus (9). After biotinylating the free cysteine, the scFv can be coupled to any molecules linked to streptavidin. Such antibody conjugates can then be bound to the surface of cells that have been incubated with the relatively harmless Newcastle disease virus. However, one disadvantage of this method is the number of different components.

A simpler method for linking any proteins and other moieties to cell surfaces is presented in this paper. It involves anchoring a scFv antibody directed against the hapten 4-ethoxymethylene-2-phenyl-oxazoline-5-one (phOx) to the cell surface by means of a eukaryotic membrane binding domain fused with its C-terminus. The phOx hapten reacts readily with primary amino and sulfhydryl groups, thus providing an efficient means for coupling any proteins and peptides to the cell surface.

MATERIALS AND METHODS

Amplification of DNA Coding for a