ELSEVIER

Gene 282 (2002) 33–41

GENE
AN INTERNATIONAL JOURNAL ON
GENES AND GENOMES

www.elsevier.com/locate/gene

ORF-FINDER: a vector for high-throughput gene identification

Irene T. Rombel*, Kathryn F. Sykes, Simon Rayner, Stephen Albert Johnston

Center for Biomedical Inventions, Departments of Internal Medicine and Biochemistry, University of Texas-Southwestern Medical Center,
5323 Harry Hines Boulevard, Dallas, TX 75390-9185, USA

Received 27 July 2001; accepted 10 November 2001
Received by V. Paces

Abstract

We have developed a simple and efficient system (ORF-FINDER) for selecting open reading frames (ORFs) from randomly fragmented genomic DNA fragments. The ORF-FINDER vectors are plasmids that contain a translational start site out of frame with respect to the gene for green fluorescent protein (GFP). Insertion of DNA fragments that bring the initiating ATG in frame with GFP and that contain no stop codons (that is, ORFs) results in the expression of ORF–GFP fusion proteins. In addition, we have developed software (GeneWorks and GenomeAnalyzer) to predict the optimal insert size for maximizing the number of gene-coding ORFs and minimizing unintentionally selected non-coding ORFs. To demonstrate the feasibility of using the ORF-FINDER system to screen genomes for ORFs, we cloned yeast genomic DNA and succeeded in enriching for ORFs by 25-fold. Furthermore, we have shown that the vector can effectively isolate ORFs from the more complex genomes of eukaryotic parasites. We envision that ORF-FINDER will have several applications including genome sequencing projects, gene building from oligonucleotides and construction of expression libraries enriched for ORFs. © 2002 Elsevier Science B.V. All rights reserved.

Keywords: Open reading frame; Open reading frame selection vector; Functional genomic screen; Expression library immunization; Parasite

1. Introduction

Progress in functional genomics is currently hampered on a practical level by the extremely large number of clones which must be generated and screened for the relevant phenotype or function. A representative random genomic expression library must contain sufficient members to span the genome multiple times in order to ensure that each protein-coding segment is present and cloned in its correct frame and orientation. For a simple virus or bacterium, in which most of the genomic DNA encodes proteins and we assume no cloning biases, this corresponds minimally to a 6-fold excess in the size of a library over the size of the genome that will need to be screened. Even greater expression library size over genome is necessary when screening eukaryotic genomes, since only a small portion of the DNA contains genes. For example, only 20% of the 30,000 kb genome of the parasite *Plasmodium falciparum* is predicted to encode genes (Gardner et al., 1998). These considerations have rendered many functional screens of eukaryotic genomes untenable, particularly those requiring animal models for testing, including ones to which we are applying expression library immunization (Tang et al., 1992; Barry et al., 1995).

An ORF selection vector is a plasmid that allows identification of clones containing open reading frames (ORFs) that correspond to gene fragments. These vectors allow ORF-containing clones to be distinguished from clones that contain non-coding DNA or ORFs that are cloned in the wrong frame or orientation. Several ORF selection vectors have been described that are based on fusing DNA inserts to reporter genes that encode selectable enzymatic functions. Most of these vectors were not designed to select ORF-containing DNA fragments from a genome but rather randomly generated fragments from a single specific gene as a means of facilitating antibody production against epitopes.

The first described ORF vector, pUK230, contains a *lacZ* reporter gene located downstream and out of frame with respect to an initiating ATG codon, with the two being separated by restriction enzyme sites (Koenin et al., 1982). Insertion of an ORF of the correct length allows expression of β-galactosidase, conferring a LacZ$^+$ phenotype on the host. This vector was shown to be useful for isolating ORF clones from a randomly cleaved 10 kb geno-

Abbreviations: ORF, open reading frame; GORF, gene open reading frame; FORF, fortuitous open reading frame; GFP, green fluorescent protein; ELI, expression library immunization

* Corresponding author. Tel.: +1214-648-1805; fax: +1-214-648-4156.
E-mail address: irene.rombel@utsouthwestern.edu (I.T. Rombel).

PII: S0378-1119(01)00819-8

mic DNA fragment that contained two exons. However, while 13% of clones exhibited a LacZ$^+$ phenotype (Ruther et al., 1982), only one out of 18 of clones (5.6%) generated from an entirely coding DNA fragment should be expected, indicating a large number of false positives. Moreover, given that only ~300 bp (3%) of the 10 kb fragment is coding DNA, there was clearly a large overrepresentation of positives. Like pUK230, ORF vectors PORF1 and PORF2 also contain a *lacZ* reporter gene (Weinstock et al., 1983; Weinstock, 1987). In addition, these plasmids contain the 5′ end of the *Escherichia coli ompF* gene, including the promoter and translational start site, located upstream and out of frame with the promoterless *lacZ* gene. As with pUK230, insertion of ORF fragments of the correct length to bring the reporter gene in frame with the initiating ATG gives rise to a LacZ$^+$ phenotype. In this case, the resultant polypeptide is a tribrid protein with the ORF translation product sandwiched between OmpF and β-galactosidase. The PORF vectors were tested with subfragments of isolated genes and shown to be useful for generating protein fusions that could be used to raise antibodies. However, the efficacy of using these vectors to distinguish ORFs from non-ORFs was not determined.

More recently, an ORFTRAP vector that contains an intein embedded within a kanamycin resistance gene was described (Daugelat and Jacobs, 1999). The ORFTRAP system relies upon insertion of an ORF to allow the intein to be translated in its correct frame, resulting in splicing and hence expression of the KanR gene. However, the ORFTRAP vector was not tolerant of a wide range of fusions, as evidenced by an intolerance for most fragments larger than ~250 bp. Furthermore, a genomic screen of *Haemophilus influenzae* yielded only 0.5% of KanR colonies, rather than the predicted 5.5%, indicating that more than 90% of the protein fusions were unstable.

Clearly, one of the main limitations of ORF screens predicated on enzymatic activity is that this functional property is likely to be perturbed by many ORF fusions. As a result of this instability, all of the afore-mentioned ORF selection vectors suffer from the same major disadvantage in that they are unlikely to tolerate a wide repertoire of protein fusions, and therefore would not produce representative libraries. In our experience, this has been the case. Consequently, these systems have not been useful for genomic screening. To address this problem, we have developed a system that does not rely on enzymatic activity, using the enhanced green fluorescent protein (GFP; Crameri et al., 1996). GFP is a non-enzymatic monomeric protein that is less likely to be adversely affected by protein fusions than the enzymatic reporters used in previous ORF screens. It is a small (27 kDa), unusually stable protein which has anecdotally been shown to be widely tolerant of many protein fusions (Prasher, 1995; Cubitt et al., 1995; Tsien, 1998). Detection of GFP is relatively trivial because it can be detected on irradiation using a standard long-wave UV light source. Furthermore, introduction of a substrate is not required for GFP, unlike other commonly used reporter genes.

By constructing, screening and characterizing a yeast genomic library in our GFP-based ORF-FINDER vector, we demonstrated that this system enriches for ORF-containing clones by approximately 25-fold, and is tolerant of approximately 50% of all ORF fusions. We have also conducted a preliminary analysis of the ORF-FINDER system in a complex parasitic genome, with encouraging results.

2. Materials and methods

2.1. Construction of ORF selection vectors

The parental open reading frame selection vector pORF–GFP was derived from plasmid pCMVi-UB (Sykes and Johnston, 1999). The T7 promoter nested within the CMV promoter was removed, as was the SV40 origin of replication. The T7 promoter and cognate ribosomal-binding site from pET-3a (Stratagene, La Jolla, CA) was PCR-amplified and cloned into pCMVi-UB as a *BgI* I to *Bam* HI fragment, with the latter site remaining unique for subsequent cloning purposes. The enhanced GFP gene from pBAD-GFPc3 (Crameri et al., 1996) was PCR amplified and inserted downstream and out of frame with respect to the initiating ATG of the T7 transcription/translation region and the unique *Bam*HI site. A T7 transcriptional terminator was amplified from pET3 (Stratagene) and inserted into a *Pst*I site downstream of the GFP gene. In addition, intron acceptor and donor sites derived from pCI (Promega, Madison, WI) were added to immediately span the respective 5′ and 3′ sites of the GFP gene. Further details are available from the authors upon request. The final parental plasmid vector pORF–GFP is shown in Fig. 1. Plasmid ORF-FINDER1 (Fig. 2) was constructed from pORF–GFP by (1) inserting unique restriction sites for *Pac*I and *Asc*I on either side of the *Bam*HI site, (2) replacing the region immediately upstream of the GFP gene with an alanine-rich linker (with the concomitant removal of the 5′ intron splice site), and (3) substituting the initiating ATG codon of GFP with a GCG codon for alanine. Plasmid ORF-FINDER2 (Fig. 3) was derived from ORF-FINDER1 by replacing the *Bam*HI restriction site with a *Nar*I site.

2.2. Cloning of genomic DNA and selection of ORF–GFP fusions

Genomic DNA from *Saccharomyces cerevisiae* was prepared using standard techniques (Sambrook et al., 1989). *Neospora caninum* and *Trypanosoma cruzi* genomic DNA preparations were kind gifts from Tobias Schlupp and Rick Tarleton, respectively. Insert DNA was prepared by partial digestion with the appropriate enzyme (typically *Sau*3AI), followed by size-fractionation on 1% agarose gels and purification on Qiaquick gel extraction columns

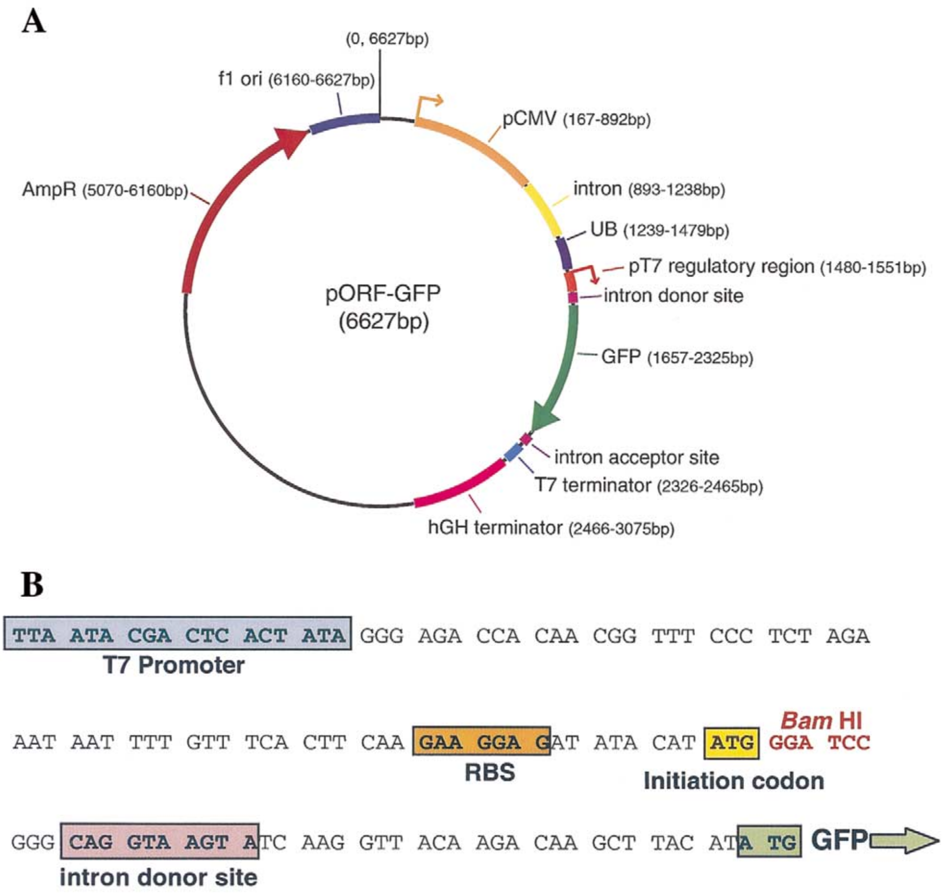
A
(0, 6627bp)
f1 ori (6160-6627bp)
pCMV (167-892bp)
AmpR (5070-6160bp)
intron (893-1238bp)
UB (1239-1479bp)
pT7 regulatory region (1480-1551bp)
intron donor site
pORF-GFP
(6627bp)
GFP (1657-2325bp)
intron acceptor site
T7 terminator (2326-2465bp)
hGH terminator (2466-3075bp)
B
TTA ATA CGA CTC ACT ATA GGG AGA CCA CAA CGG TTT CCC TCT AGA
T7 Promoter
AAT AAT TTT GTT TCA CTT CAA GAA GGA GAT ATA CAT ATG GGA TCC
Bam HI
RBS
Initiation codon
GGG CAG GTA AGT ATC AAG GTT ACA AGA CAA GCT TAC ATA TG GFP
intron donor site

Fig. 1. (a) pORF–GFP contains the transcription/translation regulatory region from bacteriophage T7 located upstream from a GFP reporter gene, which is spanned by intron donor and acceptor sites. Additional features of the vector include a mammalian pCMV promoter and intron, a ubiquitin gene that is upstream and in frame with the T7 start site, a T7 transcriptional terminator and an hGH translational terminator. In addition, pORF–GFP contains the β-lactamase gene for selection and an f1 origin of replication. (b) The transcription/translation regulatory region of pORF–GFP that is involved in ORF selection contains the T7 promoter and a Shine–Dalgarno consensus sequence (RBS). The translational start site (ATG) is located out-of-frame with respect to the downstream GFP reporter gene, and the two are separated by a unique *Bam*HI cloning site. An intron donor site is also located between the *Bam*HI site and GFP.

(Qiagen, Valencia, CA). Insert DNA was ligated with the ORF selection plasmid vectors using standard cloning techniques and transformed into *E. coli* host strain HMS174(DE3) (Novagen, Madison, WI) by electroporation (BioRad, Hercules, CA). Transformants were spread onto LB agar plates supplemented with ampicillin (75 μg/ml) and IPTG (40 μM), and grown at 30 °C for 40–48 h, at which time GFP expression could by visualized upon irradiation with a standard long-wavelength UV light source.

2.3. Insert analysis

Plasmid DNA was isolated from clones using a Wizard Kit (Promega, Madison, WI). Inserts were sequenced using

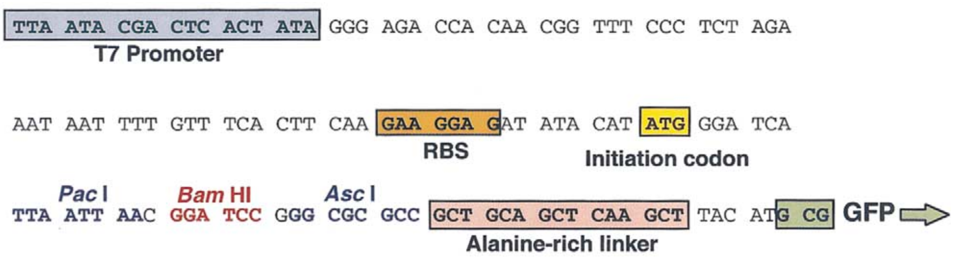
TTA ATA CGA CTC ACT ATA GGG AGA CCA CAA CGG TTT CCC TCT AGA
T7 Promoter
AAT AAT TTT GTT TCA CTT CAA GAA GGA GAT ATA CAT ATG GGA TCA
RBS
Initiation codon
Pac I
Bam HI
Asc I
TTA ATT AAC GGA TCC GGG CGC GCC GCT GCA GCT CAA GCT TAC ATG CG GFP
Alanine-rich linker

Fig. 2. The transcription/translation regulatory region of ORF-FINDER1 is similar to that of pORF–GFP (Fig. 1b), apart from the following differences: the *Bam*HI cloning site is spanned by unique sites for *Pac*I and *Asc*I, an alanine-rich linker is present immediately upstream of the GFP gene, the natural ATG initiation codon of GFP has been substituted with GCG, and the intron donor site has been removed.

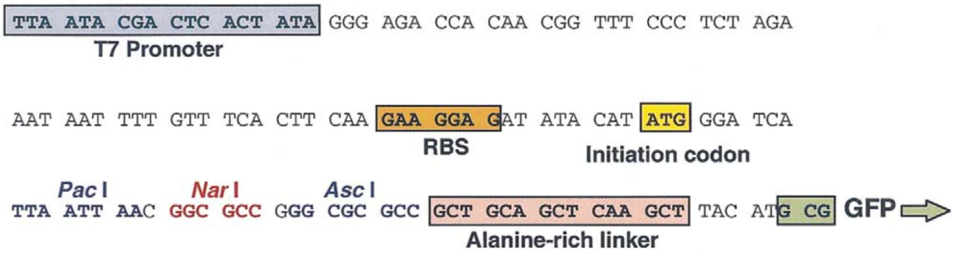
TTA ATA CGA CTC ACT ATA GGG AGA CCA CAA CGG TTT CCC TCT AGA
T7 Promoter
AAT AAT TTT GTT TCA CTT CAA GAA GGA GAT ATA CAT ATG GGA TCA
RBS
Initiation codon
PacI
NarI
AscI
TTA ATT AAC GGC GCC GGG CGC GCC GCT GCA GCT CAA GCT TAC ATG CG
GFP
Alanine-rich linker

Fig. 3. The ORF-FINDER2 vector is identical to ORF-FINDER1 (Fig. 2), except that the *Bam*HI cloning site has been replaced with a unique site for *Nar*I.

the BigDye Terminator Cycle Sequencing Ready Reaction kit (PE Applied Biosystems, Foster City, CA) and analyzed on an automated sequencer (ABI, Foster City, CA). Homology searches against the yeast genome were performed using the BLAST software (http://genome-www.stanford.edu/Saccharomyces).

For statistical analysis of ORF frequencies, the gene ORF (GORF) size distribution was generated from annotated sequence data obtained from the NCBI web site and analyzed using our software (available from our web site at www.cbi.swmed.edu/tech.html). The size distribution of fortuitous non-coding ORFs (FORFs), which correspond to the distances between two stop codons, was generated as described above. Generating the ORF distribution requires an annotated file either in Genbank or EMBL format. Our parsing software (GeneWorks) identifies all of the annotated coding regions and extracts the corresponding sequence, writing the information to an output file. This data can then be binned according to sequence size to generate the GORF distribution. Where appropriate, the software also provides the option of keeping each exon as a discrete feature or joining them together to form the final product. For our analysis, we kept the exons as separate features. A single GORF distribution was generated from annotated files for all 16 yeast chromosomes. The FORF distribution was generated by running the raw sequence file (either flat file or FASTA format) through our second analysis program (GenomeAnalyzer). This searches all reading frames for regions delimited by successive stop codons. Once all six reading frames have been analyzed, we have a FORF distribution that also contains all of the GORFs since these will also be identified in the process. As a final step, it is necessary to remove all true coding regions (GORFs) from the FORF distribution. This is achieved by a BLAST type similarity search between the file generated from GeneWorks (the GORF distribution) and the file generated from GenomeAnalyzer (the FORF distribution). In this manner, the yeast FORF distribution was generated by identifying all sequences between adjacent stop codons in all six reading frames from the raw data for all 16 chromosomes and subtracting out the GORF distribution. The data from the 16 chromosomes were then combined into a single FORF distribution. To determine if the ORF frequencies observed with different vectors were statistically different, a Student *t*-test was performed using Microsoft Excel.

3. Results

3.1. Construction of a selection vector for open reading frames

We had experimented with several ORF vectors that relied on fusion of genomic fragments to antibiotic resistance proteins (ampicillin, kanamycin). However, we found that these vectors produced very poor libraries; we obtained a very limited number of transformants, which contained plasmids with 10–25 bp genomic inserts. These non-useful fusions suggested that the enzymatically selected function was intolerant of library fusions. Therefore, we decided to develop the GFP vector as it would not depend on enzymatic activity.

To construct the pORF–GFP vector (Fig. 1a), we inserted the GFP reporter gene downstream and out of frame with respect to the ATG initiation codon adjacent to a bacteriophage T7 transcription/translation region (Fig. 1b). The reporter gene is separated from the initiating codon by a linker that contains a unique *Bam*HI cloning site. Insertion of DNA fragments with a length of $3n + 1$ between these two sequences is required to allow translation of an ORF–GFP fusion. The presence of the T7 promoter allows high levels of expression to occur upon IPTG induction; conversely expression can be minimized during subsequent amplification steps by omission of IPTG to limit possible mutation and/or loss of plasmid clones. To confirm that the pORF–GFP vector could indeed provide a distinguish-

Table 1
Phenotypic distribution of putative yeast ORF-containing clones

Total no. of colonies	No. of green colonies	Bright green phenotype	Medium green phenotype	Pale green phenotype
3120	129 (4.1%)	1.0%	1.1%	2.0%

able phenotype, a thymine residue was inserted upstream of the GFP gene to bring it in frame with the initiating ATG. Colonies of *E. coli* that contained this construct fluoresced strongly when grown in the presence of IPTG, whereas those containing the pORF–GFP vector were white (not shown).

3.2. Testing the pORF–GFP selection vector with yeast genomic DNA

To accurately determine the efficacy of pORF–GFP as an open reading frame screening vector, libraries were prepared from *S. cerevisiae* genomic DNA. *Sau*3AI-partially digested, size-selected *S. cerevisiae* DNA fragments (ranging from 100 to 600 bp) were cloned into the *Bam*HI site of pORF–GFP. To identify ORF-containing clones, transformants were screened in the presence of IPTG for green fluorescence. In preliminary experiments, it was found that the conditions of IPTG induction affected the number of false positives (colonies containing non-ORF inserts that had stop codons or were out-of-frame with GFP but nevertheless fluoresced). This number could be reduced by (1) lowering the IPTG concentration from the standard 100 μM to 40 μM and (2) incubating the plated bacteria at 30 °C rather than 37 °C (results not shown).

Using the optimized conditions, four independently constructed yeast libraries were screened for ORFs and the results of the observed phenotypes are summarized in Table 1. Of the total 3120 transformants screened, 129 colonies (4.1%) had a green fluorescent phenotype. The intensity of fluorescence varied between the colonies, so that the putative ORF-containing green colonies were classified as bright, medium or pale green, as indicated in Table 1. An inverse relationship between insert length and intensity of fluorescence was observed, with bright, medium and pale green colonies carrying inserts with respective average lengths of 208, 336 and 529 bp. Longer non-ORF inserts were more likely to generate false positives since they are more likely to contain an internal promoter and/or Shine–Dalgarno sequence and thereby allow GFP expression to occur.

In order to measure the efficacy of the ORF screen, we determined whether there is a relationship between insert identity and intensity of fluorescence. The cloned inserts from 90 green colonies were sequenced (Table 2) and analyses showed that 49 (54%) were ORFs based on the criteria that (1) they linked the initiating ATG codon of pORF–GFP in frame with the GFP reporter gene and (2) they contained no stop codons. Given that 4.1% of the total number of colonies were fluorescent green, this means that 2.2% of the total clones contained ORFs. Since one coding fragment in 18 should theoretically be cloned in the correct orientation and reading frame, and 68% of the *S. cerevisiae* genome encodes genes (Patthy, 1999), the predicted frequency of true gene ORF-containing colonies is 3.8%. Therefore, assuming no cloning biases, approximately 58% of the expected ORF–GFP fusions appear to be functional.

To determine which of the 49 ORFs identified with pORF–GFP corresponded to the ORFs of predicted genes, the translated genomic database of *S. cerevisiae* was searched with each of the translated ORF sequences. Of the 49 ORFs identified in the screen, 22 (45%) were found to correspond to gene ORFs (GORFs). Most of the fortuitous non-gene ORFs (FORFs), that is, non-gene sequences that were stop codon-free and in frame with GFP, contained relatively small inserts. This indicated that the proportion of GORFs could be greatly increased by selection of a longer insert size range (see below).

In order to ascertain whether there was any overt bias in cloning or selection, the identity (and function, where known) of each gene fragment was determined from the yeast genome database. Of the 22 ORFs that were identified as genes, 17 were unique clones, while two clones appeared to map to different positions within the same gene of unknown function. Curiously, three of the gene ORFs corresponded to the 25srRNA gene, as did seven of the 27 noncoding FORFs. This frequency exceeds the expected number of such clones which would be anticipated (since only ~100 of the ~6000 genes in the yeast genome code for 25srRNA).

3.3. Optimizing ORF vector design

Based on the results of the yeast pORF–GFP test library screen, a number of modifications were made to the vector to optimize its selectivity and versatility for genomic screening. The ATG start codon of the GFP gene was changed to CGC to reduce the incidence of spurious readthrough from Shine–Dalgarno-like sequences within the insert in order to reduce the number of false positive clones. To increase the stability of fusion proteins, the sequence immediately upstream of the GFP gene was modified to encode a short flexible peptide linker (Margolin, 2000). For subse-

Table 2
Size distribution of cloned yeast ORFs

Phenotype	No. sequenced	No. of ORFs	Average insert length (bp)	Average ORF length (bp)	Gene ORFs
Bright green	26	22 (85%)	182	250	4 (18%)
Medium green	35	15 (43%)	244	307	12 (80%)
Pale green	29	12 (41%)	523	553	6 (50%)
Total	90	49 (54%)	316	342	22 (45%)

Table 3
Frequency and sizes of ORFs selected from eukaryotic parasite genomes

Parasite	Vector	Total no. of colonies	No.of green colonies	Average insert length (bp)	No. sequenced	No. of ORFs
N. caninum	pORF–GFP	330	32 (10%)	144	32	22 (85%)
T. cruzi	pORF–GFP	409	8 (2%)	224	6	5 (83%)
N. caninum	ORF-FINDER1	422	36 (9%)	143	10	10 (100%)
T. cruzi	ORF-FINDER1	675	26 (4%)	125	3	3 (100%)

quent excision of DNA inserts, sites for restriction enzymes *Pac*I and *Asc*I (which recognize 8 bp sequences) were introduced to flank the *Bam*HI site. The resultant vector, ORF-FINDER1, is shown in Fig. 2. To increase the cloning repertoire of the vector, the *Bam*HI site of ORF-FINDER1 was replaced with a site for *Nar*I (compatible with the enzymes *Taq*I, *Mae*II, *Msp*I, *Aci*I and *Hin*P1I) to produce ORF-FINDER2 (Fig. 3).

To test the efficacy of the ORF-selection vectors for identifying ORFs from complex genomic DNA, libraries were prepared in pORF–GFP with *Sau*3AI-digested size-fractionated (~100–300 bp) genomic DNA from two eukaryotic parasites, *Neospora caninum* and *Trypanosoma cruzi*. In addition, comparable libraries were prepared in ORF-FINDER1 in order to compare the efficacy of the original pORF–GFP vector with the modified form. The resultant *N. caninum* and *T. cruzi* libraries contained approximately 10% and 2–4% of fluorescent green colonies, respectively (Table 3). Sequence analysis of a number of randomly chosen fluorescent green clones from each of the four libraries revealed that a relatively high percentage of these clones contained ORFs. Statistical analysis of these sequences showed that the ORF-FINDER1 vector gave rise to a significantly greater fraction of ORFs than the parental pORF–GFP vector ($P = 0.002$). Although it was not possible to ascertain the number of GORFs (since neither genome has been sequenced to date), all of the cloned inserts were unique, indicating that there was no overt bias for the selection of certain gene sequences.

3.4. Statistical analysis for insert size optimization for gene ORF selection

Analysis of the putative ORF-containing clones identified in the yeast pORF–GFP test library (Table 1) showed that optimizing the insert size range is critical to gene ORF (GORF) selection. Fragments that are too small are more likely to contain non-gene ORFs (FORFs). Conversely, fragments that are too large are more likely to encompass the natural stop codon of genes, thus leading to the under-representation of small genes. From the results described above, we empirically determined the optimal insert size for selecting ORFs to be approximately 250–500 bp. To ascertain whether this is indeed the optimal size range to use routinely for ORF screens, we performed a computational analysis of the annotated yeast database in order to determine the size distribution of GORFs and FORFs (Fig. 4a). Consistent with our experimental observations, we calculated that removing fragments that are ≤ 250 bp would eliminate 98.6% of FORFs and 3.4% of GORFs, while removing fragments that are ≤ 300 bp would eliminate 99.4% and 4.5% of FORFs and GORFs (Fig. 4b).

Given that *S. cerevisiae* has introns in only 4% of its genome, one would predict greater overlap between the sizes of GORF (exon) and FORF fragments in more complex genomes. To determine if this is the case, and to ascertain if this would preclude GORF selection from these larger genomes, we analyzed the annotated sequences of chromosomes II and III of *Plasmodium falciparum* (see Section 2). The estimated gene density is less than half of that found in yeast, and 43% of the genes contain at least one intron (Patthy, 1999). The cumulative ORF size distribution for *P. falciparum* (not shown) showed that removal of fragments that are ≤ 250 bp would respectively eliminate 99% of FORFs and only 33% of GORFs. It should be noted that the *P. falciparum* genome has an unusually low $G + C$ content of only 20% and a large number of long non-coding (AT) repeats (Gardner et al., 1998). This should result in an under-representation of random stop codons and hence an over-representation of FORFs. Therefore, one would predict that most genomes would contain fewer long FORFs, thus allowing a greater distinction between GORFs and FORFs.

4. Discussion

In order to streamline functional genomic screens, we have devised a system for selecting open reading frames from genomes, including those that contain large amounts of non-coding DNA. Specifically, we have designed and optimized an efficient ORF selection vector that enriches for ORF-containing cloned DNA and is amenable to high-throughput screening. In addition, it requires no prior knowledge of the organism's genomic composition.

Previous versions of ORF expression vectors, such as the *lacZ*-based pUK230 (Koenin et al., 1982; Ruther et al., 1982), PORF1 and PORF2 plasmids (Weinstock et al., 1983), and the intein-based ORFTRAP plasmid (Daugelat and Jacobs, 1999) use enzymatic reporter genes. However, such enzyme-based systems are not tolerant of fusions, presumably owing to conformational demands such as ensuring that the substrate(s) is correctly positioned relative to the catalytic residues, and that the catalytic center is not sterically perturbed. Consequently, enzyme-based ORF selection vectors are limited in their usefulness for genome-wide screening since they rely upon production of enzymatically functional ORF fusions for detection. In fact, these vectors were used only on gene-wide scales. By contrast, the ORF-FINDER vectors we describe in this study utilize the non-enzymatic GFP reporter gene, which has anecdotally been found to be widely tolerant of fusions (Prasher, 1995; Cubitt et al., 1995; Tsien, 1998). It should be noted, however, that not all fusions are successful and the failures are almost never published, so it has been difficult to assess

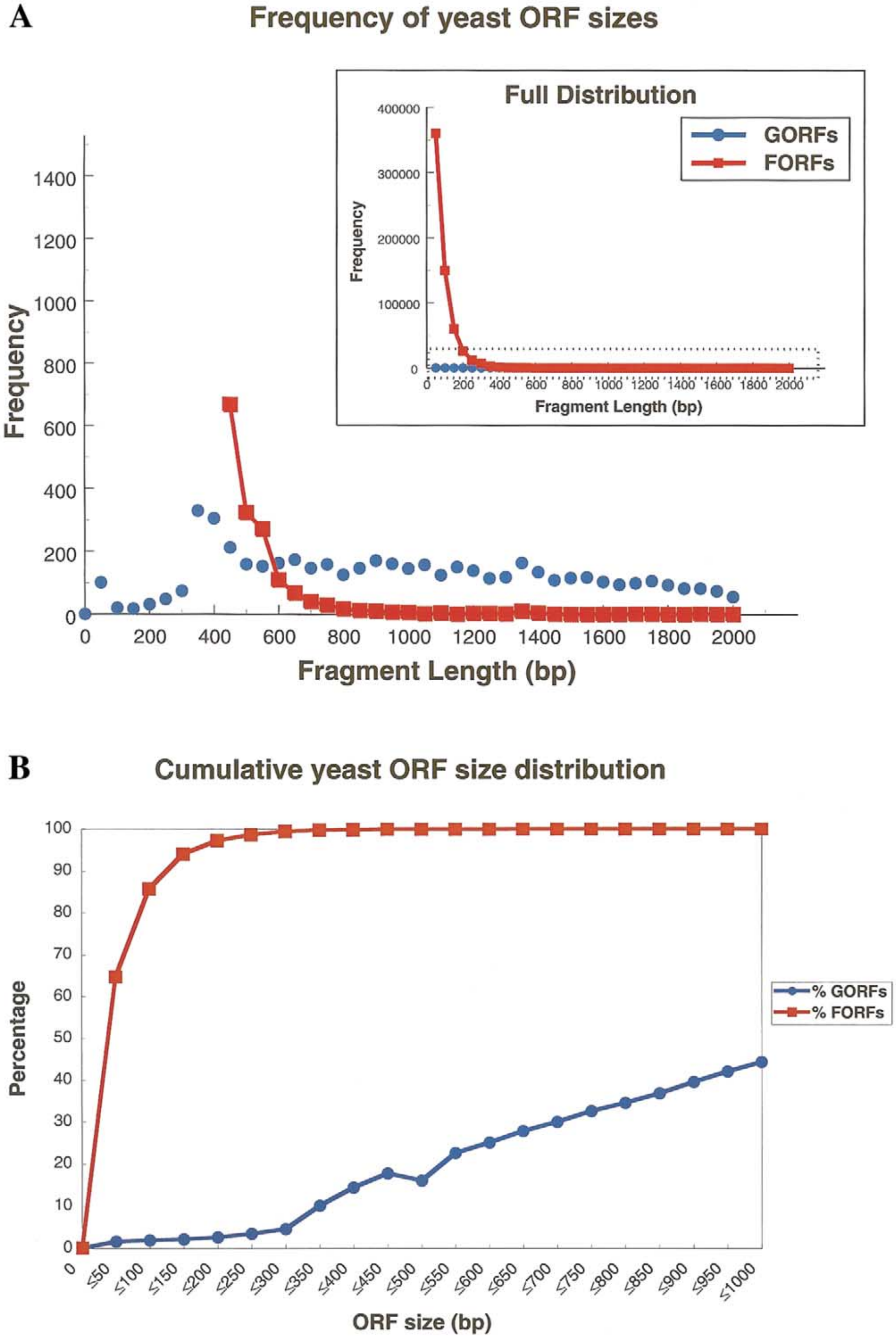

Fig. 4. GORF and FORF distributions for the yeast genome. (a) The frequency of gene ORFs (GORFs; blue circles) shows the number of DNA fragments of a particular length that correspond to protein-coding sequences. The frequency of non-coding ORFs (FORFs; red squares) shows the number of fragments that fall between two stop codons but that do not correspond to protein-coding DNA. The main graph corresponds to the small dotted region within the inset. (b) The cumulative yeast ORF size distributions represent the percentage of the total numbers of GORFs or FORFs (shown in a) that fall below any given size.

the overall success rate. One report in fact indicates that GFP can be sensitive to the improper folding of sequences fused to the N terminus, allowing fluorescence to discriminate folded from non-folded GFP fusions proteins (Waldo et al., 1999). By constructing, screening and characterizing a yeast genomic library in plasmid pORF–GFP, we identified more than 50% of the expected number of ORF-containing clones based on fluorescence, indicating that our ORF-FINDER system is fairly tolerant of a wide range of fusions. Furthermore, we observed an inverse relationship between the intensity of the colony color and the corresponding insert size, suggesting that larger ORF–GFP fusion proteins are less likely to fold correctly than their smaller counterparts or suffer more aggregation, resulting in diminished fluorescence. In the future, a feature that could be incorporated into this system to assist with protein folding is the coexpression of chaperones, which have previously been used to assist in the folding of recalcitrant proteins (Makrides, 1996). Interestingly, it has been observed that proteins fused with GFP can maintain fluorescence even when they are insoluble and trapped within inclusion bodies (S. Russell, S.A. Johnston, unpublished results). This may have contributed to the relatively wide tolerance to fusion that we observed. Furthermore, this feature offers an advantage over *lacZ*-based vectors, since in the latter case the exogenous substrate for β-galactosidase would likely be excluded from inclusion bodies.

The ORF-FINDER vectors that we described in this study can be used for effectively discriminating between ORF- and non-ORF-containing DNA fragments from any simple or complex mixture. However, no ORF selection vector can distinguish between an ORF that encodes part of a gene (GORF) from a fortuitous ORF (FORF) that encodes a random sequence that is free of STOP codons and happens to bring a reporter gene in frame with the initiating codon. In order to increase the number of GORFs relative to FORFs that are identified with ORF-FINDER, we found that delimiting the insert size range is critical for success, as evidenced by experimental data as well as computational analysis of genomic data. By carefully regulating the size distribution of the insert DNA fragments, it is possible to optimize the number of GORF-containing fragments and minimize the occurrence of FORF-containing fragments. For this purpose, we have written two programs (Geneworks and Genome Analyzer) which assist with optimization of insert size.

By using the ORF-FINDER system to construct and screen several genomic libraries, we have demonstrated that GFP is tolerant to a wide range of heterologous protein fusions, including ORF-encoded polypeptides of yeast and parasitic origin. Furthermore, we observed an approximately 25-fold enrichment for ORFs with our screening system, thus allowing reduction of subsequent functional screening by the same factor. The ORF-FINDER selection system is amenable to a wide range of genomic applications, including identification of genes for eukaryotic sequencing projects without the mRNA biasing and redundancies that are inherent with cDNA-based approaches. The 250–500 bp size range we defined for the yeast genome may be small for genome sequencing projects but it would nevertheless identify gene regions for further cloning. On the other hand, this size range is ideal for functional genomics projects for which single protein domains would be analyzed. In preliminary screens of the uncharacterized *T. cruzi* and *N. caninum* genomes, we demonstrated that this can be used as a no-knowledge genome approach. The ORF-FINDER vector can also be used in other cloning techniques. For example, gene building from oligonucleotides can be facilitated by allowing selection of clones that contain the correct full-length sequence (Stemmer et al., 1995). An important application of ORF-FINDER is to facilitate the isolation of vaccine candidates from eukaryotic parasites by decreasing the number of library members that must be screened by expression library immunization (ELI) (Tang et al., 1992; Barry et al., 1995). ELI is a functional screen for protective antigen-encoding DNA fragments from genomic libraries, using appropriate animal models to assay for protection. At present, ELI has been used to screen bacterial and viral genomic libraries, which have relatively small and compact genomes. Since the ORF-FINDER system is carried out in *E. coli*, the outcome is a relatively inexpensive primary genomic screen for any subsequent ORF screen in more expensive and time-consuming animal experiments.

In summary, we have developed a vector, protocols for its use and programs to assist in its application to screen for open reading frames.

Acknowledgements

We would like to thank Drs. Tobias Schlupp and Rick Tarleton for the kind gifts of *N. caninum* and *T. cruzi* genomic DNA, respectively. This work was supported by a grant from DARPA to S.A.J. and an Advanced Technology Program grant from the Texas Higher Education Coordinating Board to I.T.R.

References

Barry, M.A., Lai, W.C., Johnston, S.A., 1995. Protection against mycoplasma infection using expression-library immunization. Nature 377, 632–635.

Crameri, A., Whitehorn, E.A., Tate, E., Stemmer, W.P.C., 1996. Improved green fluorescent protein by molecular evolution using DNA shuffling. Nat. Biotechnol. 14, 315–319.

Cubitt, A., Heim, R., Adams, S., Boyd, A., Gross, L., Tsien, R., 1995. Understanding, improving and using green fluorescent proteins. Trends Biochem. Sci. 20, 448–455.

Daugelat, S., Jacobs, W.R., 1999. The *Mycobacterium tuberculosis recA* intein can be used in an ORFTRAP to select for open reading frames. Protein Sci. 8, 644–653.

Gardner, M.J., et al., 1998. Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*. Science 282, 1126–1132.

Koenin, M., Ruther, U., Muller-Hill, B., 1982. Immunoenzymatic detection of expressed gene fragments cloned in the *lacZ* gene of *E. coli*. EMBO J. 1, 509–512.

Makrides, S., 1996. Strategies for achieving high-level expression of genes in *Escherichia coli*. Microbiol. Rev. 60, 512–538.

Margolin, W., 2000. Green fluorescent protein as a reporter for macromolecular localizations in bacterial cells. Methods 20, 62–72.

Patthy, L., 1999. Genome evolution and the evolution of exon-shuffling – a review. Gene 238, 103–114.

Prasher, D.C., 1995. Using GFP to see the light. Trends Biochem. Sci. 11, 320–323.

Ruther, U., Koenen, M., Sippel, A.E., Muller-Hill, B., 1982. Exon cloning: immunoenzymatic identification of exons of the chicken lysozyme gene. Proc. Natl. Acad. Sci. USA 79, 6852–6855.

Sambrook, J., Fritsch, E.F., Maniatis, T., 1989. Molecular Cloning: A Laboratory Manual, 2nd Edition. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Stemmer, W.P.C., Crameri, A., Ha, K., Brennan, A., Heyneker, H., 1995. Single-step assembly of a gene and entire plasmid from large numbers of oligodeoxyribonucleotides. Gene 164, 49–53.

Sykes, K.F., Johnston, S.A., 1999. Genetic live vaccines mimic the antigenicity but not the pathogenicity of live viruses. DNA Cell Biol. 18, 521–531.

Tang, D., DeVit, M., Johnston, S.A., 1992. Genetic immunization is a simple method for eliciting an immune response. Nature 356, 152–154.

Tsien, R.Y., 1998. The green fluorescent protein. Annu. Rev. Biochem. 67, 509–544.

Waldo, G.S., Standish, B.M., Berendzen, J., Terwilliger, T.C., 1999. Rapid protein-folding assay using green fluorescent protein. Nat. Biotechnol. 17, 691–695.

Weinstock, G.M., 1987. Use of open reading frame expression vectors. Methods Enzymol. 154, 156–163.

Weinstock, G.M., Rhys, C., Berman, M.L., Hampar, D., Jackson, D., Silhavy, T.J., Weisemann, J., Zweig, M., 1983. Open reading frame expression vectors: A general method for antigen production *in Escherichia coli* using protein fusions to β-galactosidase. Proc. Natl. Acad. Sci. USA 80, 4432–4436.