High-throughput cloning of eukaryotic open reading frames (ORFs) using Gateway™ site-specific recombination

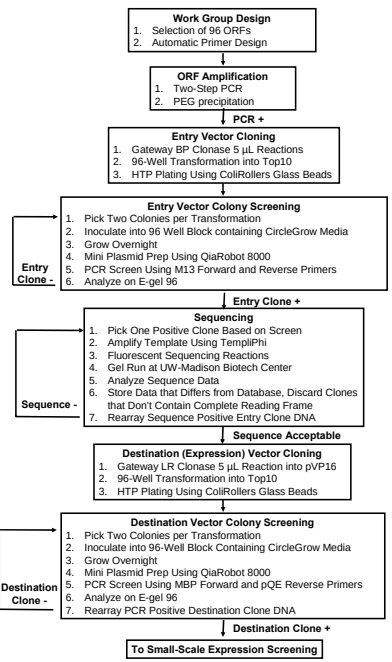
Russell L. Wrobel, Sandy Thao, Craig S. Newman, Qin Zhao, Todd Kimball, Eric Steffen,
Paul G. Blommel, Megan Riters, Zachary T. Eggers, and Brian G. Fox

University of Wisconsin-Madison, 433 Babcock Drive, Madison, Wisconsin, USA 53706-1549, http://www.uwstructuralgenomics.org

Abstract

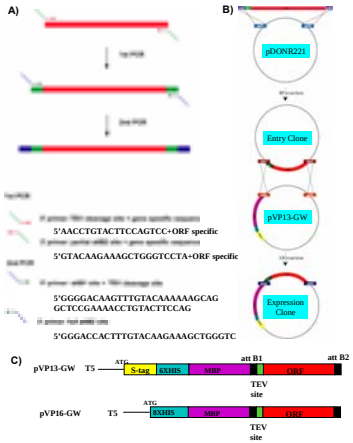
The Center for Eukaryotic Structural Genomics (CESG) was founded as a collaborative effort to develop technologies for the rapid and economic determination of protein three-dimensional structures. The initial focus was on the genome of the model plant *Arabidopsis thaliana*. Open reading frames (ORFs) were chosen on the likelihood that they would represent important unknown regions of protein conformation and fold space or that they would elucidate novel fold-function relationships. Protocols were developed for the high-throughput cloning of *Arabidopsis* ORFs into *Escherichia coli* expression vectors. The chosen ORFs were amplified from a bulk cDNA pool created by reverse transcription of RNA isolated from an *Arabidopsis* callus culture. A novel Gateway™ protocol was developed to insert the amplified open reading frames into an entry vector for storage and sequence determination. Sequence verified entry clones were then used to create expression vectors again via the Gateway™ recombination. We have developed and tested several different Gateway™ compatible *E. coli* expression vectors that contain different solubility and purification tags. To expand the prospects of finding high value targets we have begun to clone ORFs from other model eukaryotic organisms. The Integrated Molecular Analysis of Gene Expression (IMAGE) consortium's Mammalian Gene Collection (MGC) provides over 12,000 human and 10,000 mouse non-redundant full-length ORF clones from which to choose high value targets. We used our cloning strategy to make expression clones from several hundred of these purchased mammalian ORF clones. We also have successfully cloned rice ORFs using a bulk cDNA pool similar to that used for Arabidopsis ORFs, thus attesting to the flexibility of our cloning protocols. Comparative analysis of over 3,200 cloning experiments from these different cDNA sources will be presented. The expression and purification of proteins expressed from these clones will be presented on other posters.

Cloning protocol flowchart



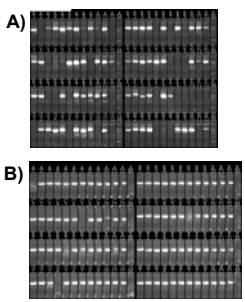
Gateway cloning at CESG

CESG uses Gateway technology (Invitrogen) to generate expression vectors that provide simple swapping of expression systems and protein tags. A) Diagrammed is the two-step amplification of an ORF by PCR with the addition of recombination (*att*) and TEV protease cleavage sites. The template for the first PCR can be genomic DNA for intronless genes, bulk cDNA from reverse transcribed cellular RNA, or plasmid DNA containing cloned cDNAs; B) Recombinational cloning of the PCR-generated insert to give Entry and Expression clones; C) CESG has developed multiple Gateway compatible expression vectors. Diagrammed here are the vectors we rely on most frequently. They are T5, LacI repressible promoter (pQE) based vectors. pVP13-GW has an S-tag (for visualization) and a 6XHIS-tag (for purification). pVP16-GW has an 8XHIS tag (for purification) and both have the MBP (maltose-binding protein, for solubilization) fused to the N-terminal of the target protein. When required, the entire fusion is cleavable from the protein target by TEV protease. We are in the process of testing other solublization tags and expression systems.



Examples of the two-step PCR amplification

Shown are A) the amplified products using a template of reverse transcribed RNA isolated from the *Arabidopsis* T87 callus culture line. Using this template we are able to amplify about 60-80% of the targeted ORFs. This success rate reflects our NimbleGen's chip analysis that showed about 60-80% of all *Arabidopsis* ORFs are expressed in this callus cell line. B) Plasmid DNA containing human cDNA clones purchased from the Mammalian Gene Collection and used as template in our PCR process. We obtain a greater number of PCR+ using cDNA clones as our template. Failures are usually due to receiving the wrong clone.



An altered Gateway *att*B1 site is more efficient

Shown below is the nucleotide sequence of the *att*B1 site we use (QZ) compared to Invitrogen's recommended *att*B1 site (INV) with and without the nucleotides that encode the TEV recognition site (TEV). The table shows the number of colony forming units per nanogram DNA obtained from BP reactions of 8 ORFs containing the different *att*B1 and TEV sites transformed into competent Top10 cells. When our altered *att*B1 site is used in conjunction with the TEV site, we get much better cloning efficiency than with the *att*B1 site recommended by Invitrogen. However, if the TEV site is not included we do not observe this increase in cloning efficiency.

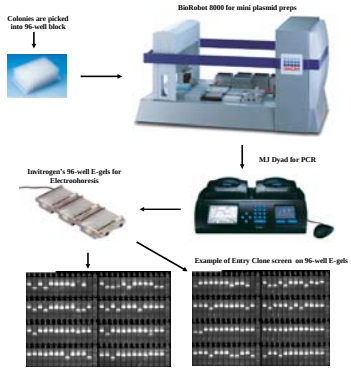
attB1 "TEV site"

QZ-TEV: 5'GGGG ACA A**TG** TTG TAC AAA AAA GCA GGC TCC GAA AAC CTG TAC CAG TCC
INV-TEV: 5'GGGG ACA AGT TTG TAC AAA AAA GCA GGC TCC GAA AAC CTG TAC CAG TCC
QZ-GW: 5'GGGG ACA A**TG** TTG TAC AAA AAA GCA GGC TCC
INV-GW: 5'GGGG ACA AGT TTG TAC AAA AAA GCA GGC TCC

2	QZ-TEV	INV-TEV	QZ-GW	INV-GW
1	27	1	71	267
2	31	2	30	50
3	26	2	34	53
4	128	1	47	54
5	29	1	40	62
6	25	1	33	51
7	21	1	50	53
8	53	2	94	18
Average	42.5	1.38	49.88	76

High-throughput plasmid DNA screen

To determine the presence of insert in both our Entry and Expression plasmids, the following robot aided protocol was developed. The colonies are picked into 96-well growth blocks containing CircleGrow media and the appropriate antibiotic. This is grown overnight at 37ºC with vigorous shaking. The next morning plasmid DNA is isolated with the use of QiaRobot 8000. This DNA is then used as a template in PCR using universal vector primers that flank the insertion site. The PCR products are analyzed on E-gels 96 (Invitrogen) and positive clones moved down the pipeline.



Efficiency of CESG cloning strategies

Tabulated below is the relative and overall efficiency of four steps in our cloning process. *Arabidopsis* and rice ORFs were amplified from bulk cDNA reverse transcribed from RNA isolated from a callus culture and whole seedlings, respectively. The mammalian ORFs were amplified from purchased cDNA containing plasmids. The relative efficiency of the PCR step most likely reflects the probability of the targeted ORF being included in the cDNA source. The failures in PCR using plasmids containing mammalian cDNA are usually due to receiving the wrong plasmid from the vendor. Once a PCR product is obtained, the efficiency of the two Gateway recombination steps is about the same at 92-97%.

		Arabidopsis			Mammalian			Rice		
Step	Database Action	Database Instances	Relative Efficiency (%)	Overall Efficiency (%)	Database Instances	Relative Efficiency (%)	Overall Efficiency (%)	Database Instances	Relative Efficiency (%)	Overall Efficiency (%)
Target Gene Selection	Selected	2401			768			120		
PCR Amplification	PCR +	1739	72.4	72.4	740	96.4	96.4	67	55.8	55.8
Entry Clone	Entry Clone +	1691	97.2	70.4	713	96.4	92.8	63	94.0	52.5
Destination Clone	Destination Clone +	1558	92.1	64.9	654	91.7	85.2	60	95.2	50.0

Direct cloning of ORFs from bulk cDNA often uncovers differences from the predicted gene model

Tabulated below are the results of the DNA sequence analyses of our *Arabidopsis* entry clones. The sequence actions are the actions entered into our LIMS. The sequence + action means there is no difference between the sequence of our clone and the sequence of the gene model prediction. "Sequence silent" means there is a silent mutation in our clone. "Sequence missense" means there are no more than three missence mutations. "Sequence -" means there is a frame shift mutation in our clone, usually caused by nucleotide deletion in the primer regions. "Sequence annotation" means that the ORF clone is different from the predicted gene model annotation. This is usually due to changes in the intron/exon composition of the gene. Included in this are eight cases where we have found two slice variants of the same gene. All sequences that differ from the gene model are placed in our LIMS, and we routinely report the annotation differences to GenBank. If bulk cDNA is used to clone mammalian genes we would expect more of these annotation differences since mammalian genes contain on average more introns than plant genes and have a greater propensity for alternative splicing. We consider the clone as acceptable if its sequence action is designated +, silent, missense, or annotation. If a clone is "Sequence –", we will usually sequence another clone to find one that is acceptable.

Sequence Action	Database Instances	Percent total
Sequence +	730	49.5%
Sequence Silent	44	3.0%
Sequence Missense	207	14.0%
Sequence Annotation	350	23.7%
Sequence -	145	9.8%
Total	1476	100.0%

Summary

The Gateway method provides an efficient method for cloning ORFs for structural genomics. Our methods are sufficiently robust and allow us to use a variety of templates, including genomic DNA, bulk cDNA, and plasmid DNA, to amplify a chosen ORF. Once the ORFs are amplified, the Gateway reactions are highly efficient at inserting them into our expression vectors. We have also developed a robotically aided PCR screen to identify clones containing inserts. Our sequence analysis of the clones obtained from bulk cDNA uncovered that a moderate level (23.7%) of the cloned ORFs differ from the gene model predictions. We think these annotation differences will become more prevalent if bulk cDNA is used to amplify mammalian targets. The expression of some of the clones constructed using this system are presented in other posters presented by CESG members.

CESG is supported by the National Institute of General Medical Sciences through the Protein Structure Initiative (P50 GM 64598).