

Marquette University

e-Publications@Marquette

Biological Sciences Faculty Research and Publications

Biological Sciences, Department of

8-2003

Overproduction and Analysis of Eukaryotic Multiprotein Complexes in *Escherichia coli* Using a Dual-vector Strategy


Jeff Finkelstein
Rockefeller University

Edwin Antony
Marquette University, edwin.antony@marquette.edu

Manju M. Hingorani
Wesleyan University

Michael O'Donnell
Rockefeller University

Follow this and additional works at: https://epublications.marquette.edu/bio_fac

 Part of the [Biology Commons](#)

Recommended Citation

Finkelstein, Jeff; Antony, Edwin; Hingorani, Manju M.; and O'Donnell, Michael, "Overproduction and Analysis of Eukaryotic Multiprotein Complexes in *Escherichia coli* Using a Dual-vector Strategy" (2003). *Biological Sciences Faculty Research and Publications*. 411.
https://epublications.marquette.edu/bio_fac/411

Marquette University

e-Publications@Marquette

Biological Sciences Faculty Research and Publications/College of Arts and Sciences

This paper is NOT THE PUBLISHED VERSION; but the author's final, peer-reviewed manuscript. The published version may be accessed by following the link in the citation below.

Analytical Biochemistry, Vol. 319, No. 1 (August 1, 2003): 78-87. [DOI](#). This article is © Elsevier and permission has been granted for this version to appear in [e-Publications@Marquette](#). Elsevier does not grant permission for this article to be further copied/distributed or hosted elsewhere without the express permission from Elsevier.

Overproduction and Analysis of Eukaryotic Multiprotein Complexes in *Escherichia Coli* Using A Dual-Vector Strategy

Jeff Finkelstein

Rockefeller University and Howard Hughes Medical Institute, New York, NY

Edwin Antony

Molecular Biology and Biochemistry Department, Wesleyan University, Middletown, CT

Manju M Hingorani

Molecular Biology and Biochemistry Department, Wesleyan University, Middletown, CT

Michael O'Donnell

Rockefeller University and Howard Hughes Medical Institute, New York, NY

Abstract

Biochemical studies of eukaryotic proteins are often constrained by low availability of these typically large, multicomponent protein complexes in pure form. *Escherichia coli* is a commonly used host for large-scale

protein production; however, its utility for eukaryotic protein production is limited because of problems associated with transcription, translation, and proper folding of proteins. Here we describe the development and testing of pLANT, a vector that addresses many of these problems simultaneously. The pLANT vector contains a T7 promoter-controlled expression unit, a p15A origin of replication, and genes for rare transfer RNAs and kanamycin resistance. Thus, the pLANT vector can be used in combination with the pET vector to coexpress multiple proteins in *E. coli*. Using this approach, we have successfully produced high-milligram quantities of two different *Saccharomyces cerevisiae* complexes in *E. coli*: the heterodimeric *Msh2–Msh6* mismatch repair protein (248 kDa) and the five-subunit replication factor C clamp loader (250 kDa). Quantitative analyses indicate that these proteins are fully active, affirming the utility of pLANT+pET-based production of eukaryotic proteins in *E. coli* for in vitro studies of their structure and function.

Keywords

Msh2–Msh6, RFC, Eukaryotic protein complex, Overexpression, *Escherichia coli*, pET vector

Many key metabolic and regulatory processes of the cell are driven by the action of large protein assemblies. In *Escherichia coli* DNA metabolism, for example, the 10-subunit DNA polymerase III holoenzyme catalyzes DNA replication and is required for both recombination and repair [1]. Often, the protein machinery of eukaryotic organisms comprises even more intricate multicomponent versions of prokaryotic counterparts. For example, in the bacterial DNA mismatch repair system, a single MutS polypeptide functions as a homodimer to recognize errors in DNA and signal initiation of repair, whereas eukaryotes contain several MutS homologues (*Msh*)¹ that function as heterodimers (e.g., *Saccharomyces cerevisiae* and human *Msh2–Msh6* and *Msh2–Msh3*) [2], [3].

Detailed biochemical and biophysical studies require large-scale production of these protein complexes, which is usually accomplished by overexpression and purification of individual subunits from a host organism such as *E. coli*, followed by reconstitution of the complexes in vitro. In some cases, this method has yielded protein complexes in sufficient quantities for crystal structure determination (e.g., *E. coli* γ complex clamp loader [4] and the nucleosome core particle [5], among others). Quite often, however, in vitro reconstitution yields can be very low because one or more of the individual subunits are misfolded or are otherwise unstable/insoluble in the absence of their interacting partners. In fact, overproduction of heterologous (e.g., eukaryotic) proteins in *E. coli*—particularly components of multiprotein complexes—often results in aggregation and formation of inclusion bodies [6]. Another challenge for large-scale eukaryotic protein complex production is the low expression level of these proteins in *E. coli*, the most convenient and commonly used host organism. This problem may be largely due to differential codon usage in prokaryotes and eukaryotes (codon bias). For example, the AGA and AGG arginine codons routinely found in *S. cerevisiae* genes are rarely used in *E. coli*; therefore, the *E. coli* translational machinery is not equipped to handle such genes [7], [8].

Problems related to multiprotein complex solubility and stability can often be solved by coexpression of the individual components *in vivo* [9]. This strategy has been used successfully to produce a wide variety of protein complexes, including nuclear receptor proteins [10], immunoglobulin Fv fragment [11], *Schizosaccharomyces pombe* RPA single-stranded DNA binding protein [12], human RPA [13], and the VHL–elonginC–elonginB complex [14]. Moreover, there is an increase in production efficiency, as the entire protein complex is purified at one time rather than in a piecemeal fashion. The component proteins can be coexpressed from a single plasmid vector carrying several genes in monocistronic or polycistronic units. Depending on the size of the protein complex, however, this plasmid may well be excessively large and difficult to prepare and manipulate. An alternate strategy is to utilize multiple vectors with compatible origins of replication and different resistance markers to coexpress the proteins.

In this report, we describe the development of pLANT, an overexpression vector that is compatible with the pET vector system for co-expression of proteins in *E. coli*. In addition, we have prepared an advanced version of the pLANT vector that contains genes for tRNAs rarely used in *E. coli* but essential for good expression of many *S. cerevisiae* and other eukaryotic genes [8], [15], [16]. The pLANT+pET combination was tested for expression in *E. coli* of two large eukaryotic protein complexes, the *S. cerevisiae* *Msh2–Msh6* dimer and the 5-subunit replication factor C (RFC) clamp loader. We demonstrate that the single subunits of these complexes are insoluble when expressed individually; however, when all the subunits are expressed together (with the help of appropriate tRNAs), tens of milligrams of fully active *Msh2–Msh6* and RFC can be purified from *E. coli* with ease. Finally, we have performed active site titrations with their respective substrates to demonstrate that the pure complexes are fully active.

Materials and methods

pLANT vector design and development

pLANT vectors were constructed in a multistep procedure initiated by replacement of the ColE1 replication origin of pET-9c (Novagen [17]) with the p15A replication origin. A 3.5-kb *A/wNI/XmnI* fragment of pET-9c (excluding the replication origin but containing the T7 promoter-mediated expression unit and kanamycin resistance gene) was treated with Klenow fragment and alkaline phosphatase (New England Biolabs) to form a blunt end and then ligated to the 0.88-kb *XmnI/ClaI* blunt end fragment of pACYC184 (New England Biolabs) containing the p15A origin to yield the 4.4-kb pLANT-1. These manipulations, and others to follow below, were performed as described [18]. Next, pLANT-2 was formed by incorporating the T7 expression unit from pET-11a (Novagen) into pLANT-1. Briefly, the 3.8-kb *BglII/EcoRI* fragment from pLANT-1 was ligated to a 0.47-kb *BglII/EcoRI* fragment from pET-11a. The resulting pLANT-2 vector offers the standard restriction sites *NdeI/BamHI* for subcloning genes (as does pLANT-1) and extra control of the T7 promoter via the overlapping *lac* operator. The pLANT-3 vector was formed by incorporating the T7 expression unit from pET-16b into pLANT-1. In this case, the 0.5-kb *BglII/EcoRI* fragment from pET-16b was ligated to the 3.8-kb *BglII/EcoRI* fragment of pLANT-1. This vector allows production of proteins with a His-tag (if the *NdeI* site is used; no His tag if the *NcoI* site is used). Both pLANT-2 and pLANT-3 vectors were modified further by introduction of *argU* (for AGA and AGG codons), *ileY* (for AUA), and *leuW* (for the CUA codon) tRNA genes between the *HindIII/EcoRI* sites. The resulting overexpression vectors, pLANT-2/RIL and pLANT-3/RIL, carry a p15A replication origin, a *T7lac* promoter followed by the efficient ribosome binding site of T7 gene 1 protein, a gene insertion site (at *NdeI* or *NcoI*, which provide the translation start codon), a T7 transcription terminator, a kanamycin resistance gene, and tRNA genes to compensate for codon bias.

Cloning of *S. cerevisiae* genes into pLANT or pET vectors

Msh2 and *Msh6* genes were amplified from *S. cerevisiae* genomic DNA using oligonucleotide primers that introduced an *NdeI* site at the initiating methionine and placed a *BamHI* site 70 nucleotides past the stop codon. The PCR products for *Msh2* and *Msh6* were ligated individually into the *NdeI/BamHI* sites of pET-11a and pLANT-2/RIL vectors, respectively.

A pET plasmid encoding the RFC2, 3, and 4 subunits was prepared by first amplifying the *RFC2*, 3, and 4 genes from *S. cerevisiae* genomic DNA and cloning them individually into the *NdeI/BamHI* sites of pET-11a (*RFC3*) or the *NcoI/BamHI* sites in pET-16b (*RFC2* and 4). To prepare plasmids for coexpression of these RFC proteins in *E. coli*, *RFC2*, *RFC3*, and *RFC4* genes were combined into a single pET-11a vector. A *BglII/ClaI* fragment of pET(11a)–*RFC3* containing the gene and T7 promoter was blunted and inserted into pET(11a)–*RFC4* (cut with *AvaI*, blunted, and phosphatased). Next, pET(11a)–*RFC[3+4]* was cut with *SphI*, blunted, and phosphatased; then a *BglII/HindIII* fragment of pET(11a)–*RFC2*, containing the gene and T7 promoter, was blunted and ligated to the linearized pET(11a)–*RFC[3+4]*. The resulting 9.6-kb pET(11a)–*RFC[2+3+4]* plasmid contains three RFC subunit

genes, each in individual T7 expression units. The *RFC2*, *3*, and *4* genes were sequenced to ensure that no errors were introduced during the PCR amplification.

Next, the genes encoding RFC1 and RFC5 were amplified from *S. cerevisiae* genomic DNA and inserted individually into the *NdeI/BamHI* sites of pLANT-2/RIL (*RFC5*) or the *NcoI/BamHI* sites of pET-16b (*RFC1^s*). A modified version of RFC5 protein—RFC5^{HK}—was prepared by inserting DNA encoding the amino acid sequence MGLRRASVHHHHHSSGHIEGRH (which contains a Kinase tag for ³²P labeling with cAMP-dependent protein kinase and a six-histidine tag) into the *NdeI* site of pLANT-2/RIL-*RFC5* to yield pLANT-2/RIL-*RFC5^{HK}*. To construct the RFC[1+5]-pLANT vector, an *SgrAI/ClaI* fragment of pET(11a)-*RFC1* was blunted and inserted into pLANT-2/RIL-*RFC5^{HK}* (cut with *EcoRV* and phosphatased) to yield pLANT-2/RIL-RFC[1+5^{HK}]. To produce a truncated version of RFC1 protein, RFC1^s, the gene was truncated by PCR such that 282 amino acids from the N terminus were deleted, and the glycine at position 283 was changed to methionine. The truncated *RFC1^s* gene fragment was cut using *NdeI* and *BamHI* and inserted into *NdeI/BamHI* sites of pET-11a to yield pET(11a)-*RFC1^s*. An *SgrAI/ClaI* fragment of pET(11a)-*RFC1^s* was blunted and inserted into pLANT-2/RIL-*RFC5* (cut with *EcoRV* and phosphatased) to yield pLANT-2/RIL-RFC[1^s+5]. The *RFC1^s* and *RFC5* genes were sequenced to ensure that no errors were incurred during PCR amplification. These pLANT-based plasmids were used in combination with pET(11a)-RFC[2+3+4] to produce modified RFC complexes.

Overexpression and purification of *S. cerevisiae* proteins from *E. coli*

For overproduction of *Msh2-Msh6* protein complex, the pLANT-2/RIL-*Msh6* and pET(11a)-*Msh2* plasmids were cotransformed into BLR(DE3) cells (Novagen) and selected for resistance to both ampicillin and kanamycin; 0.5–1 µg of each plasmid is required for efficient cotransformation. A fresh transformant was grown in 16 liters of LB media containing ampicillin (100 µg/ml) and kanamycin (50 µg/ml) at 37 °C to OD₆₀₀ 0.6 and induced with 0.5 mM isopropyl-β-d-thiogalactoside for 3 h. All further steps were performed at 4 °C. The cells were harvested by centrifugation and then resuspended in 300 ml of Buffer A (25 mM Tris-HCl (pH 8.0), 1 mM EDTA, 5% glycerol, 1 mM phenylmethylsulfonyl fluoride) containing 1 M NaCl. The cells were lysed by treatment with lysozyme (0.4 mg/ml), three freeze-thaw cycles (liquid N₂ and 37 °C), and, finally, treatment in a Dounce homogenizer. The cell lysate was clarified by centrifugation and dialyzed overnight against Buffer A. *Msh2-Msh6* protein complex was purified by sequential ion-exchange chromatography over an SP-Sepharose column (Buffer A; 20 ml bed volume; 200 ml gradient of 150–400 mM NaCl), Affi-Gel heparin column (20 ml bed volume; 200 ml gradient of 250–500 mM NaCl in Buffer B (20 mM potassium phosphate (pH 7.2), 0.5 mM EDTA, 5% glycerol)), and Q-Sepharose column (5 ml bed volume; 60 ml gradient of 200–600 mM NaCl in Buffer A). Following the Q-Sepharose column, the protein was dialyzed against Buffer A to a conductivity equal to 100 mM NaCl and concentrated by centrifugation through a centricon-10 (Millipore) to a concentration of approximately 2 mg/ml. The final yield is 1–1.2 mg of *Msh2-Msh6* complex per liter of cell culture. The final preparation of *Msh2-Msh6* complex is greater than 95% pure. The *Msh2-Msh6* concentration was measured by Bradford assay and by absorbance at 280 nm in 6 M guanidinium hydrochloride (extinction coefficient: 186,970 M⁻¹ cm⁻¹). The column resins, SP-Sepharose and Q-Sepharose, were purchased from Amersham Pharmacia and Affi-Gel heparin was purchased from Bio-Rad.

For overproduction of RFC protein complex, pLANT-2/RIL-RFC[1^s+5] (or pLANT-2/RIL-RFC[1+5^{HK}]) was cotransformed with pET(11a)-RFC[2+3+4] into BLR(DE3) cells as described above. The proteins were overexpressed in *E. coli* as described above for *Msh2-Msh6* (12 liters *E. coli* cell culture). The cells were lysed similarly except Buffer C (30 mM HEPES-NaOH (pH 7.5), 0.5 mM EDTA, 7% glycerol) was used and 100 mM NaCl was included. RFC was purified by chromatography over a SP-Sepharose column (25 ml bed volume; 250 ml gradient of 150–400 mM NaCl in Buffer C) and a Q-Sepharose column (8 ml bed volume; 80 ml gradient of 150–600 mM NaCl in Buffer C). The purified protein was dialyzed against Buffer C containing 100 mM NaCl. The final yields are greater than 5 mg RFC per liter of cell culture greater than 95% pure protein was obtained for each of

these RFC preparations. Protein concentration was measured by Bradford assay and by absorbance at 280 nm in 6 M guanidinium hydrochloride (extinction coefficients: RFC^{HK5} complex: 163,520 M⁻¹cm⁻¹; RFC¹⁵ complex, 162,120 M⁻¹ cm⁻¹).

Solubility tests of Msh2–Msh6 mismatch repair and RFC clamp loader proteins

To test the solubility of *Msh2*, *Msh6*, and the *Msh2–Msh6* complex, the proteins were induced in 100 ml *E. coli* cell cultures (grown from freshly transformed colonies selected for the appropriate plasmids). Cells were harvested and lysed as described above in Buffer A containing either 0 or 1 M NaCl. The cell lysate was clarified by centrifugation and equivalent amounts (relative to the initial lysate volume) of cleared lysate (soluble protein) and cell debris pellet (insoluble protein) were analyzed by 10% SDS–PAGE. Proteins were visualized by Coomassie blue staining to assess the relative level of soluble versus insoluble recombinant protein. RFC complexes were analyzed similarly (except that cell lysis was performed in Buffer C). RFC subunits were resolved by analysis on an 18 cm × 20 cm × 1 mm, 7.5% acrylamide, 0.25% bisacrylamide, 0.1% SDS gel, cross-linked with 0.16% TEMED (with a 4% stacking gel). The gel was first chilled and then developed at 4 °C at 25 mA constant current until the dye front traversed 3/4 of the stacking gel and then at 50 mA until the dye front had entered the resolving gel by approximately 3 cm. Following this, the gel was developed at a constant voltage of 300 V until the dye was near the end of the gel [19]. The entire process required about 5 h. This is necessary to resolve the RFC2, 3, and 5 subunits which otherwise comigrate.

Activity assays of Msh2–Msh6 mismatch repair and RFC clamp loader proteins

The mismatch recognition and ATP binding activities of *Msh2–Msh6* complex were assayed by nitrocellulose membrane binding assays. Briefly, nitrocellulose membranes (0.2 μm pore size; Schleicher & Schuell) were pretreated with 0.5 N NaOH, rinsed well with H₂O, and equilibrated in Buffer D (50 mM Tris–HCl (pH 8), 5 mM MgCl₂, 5% glycerol, 110 mM NaCl). *Msh2–Msh6* (0–2 μM) was titrated into 15-μl reactions containing ³²P-labeled 37-nucleotide duplex DNA (0.5 μM), either fully matched or containing a G:T mismatch at the center, in Buffer D and 110 mM NaCl (final concentration) at 25 °C. Aliquots of 10 μl were filtered through the membrane on a single filter assembly (VWR) under vacuum. The membranes were washed before and after filtration with 150 μl Buffer D. The molar amount of DNA bound to *Msh2–Msh6* was determined by quantitating the radioactivity on the membrane with a PhosphorImager (Molecular Dynamics) and plotted versus *Msh2–Msh6* concentration.

The concentration of active sites within the *Msh2–Msh6* heterodimer was assayed by determining the stoichiometry of ATPγS binding to the protein complex. Reactions (15 μl) containing *Msh2–Msh6* (1 μM) and 0–150 μM ATPγS + 0.3 μCi ³⁵S-ATPγS were incubated for 15 min at 25 °C, and 10-μl aliquots were filtered through nitrocellulose membranes as described above. The molar amount of nucleotide bound to *Msh2–Msh6* was determined and plotted versus free nucleotide concentration. The binding isotherms were fit to an equation describing 1:2 protein–ligand interaction: $[N \cdot M]/[M_t] = K_1[N_f] + 2K_1K_2[N_f]^2 / (1 + K_1[N_f] + 2K_1K_2[N_f]^2)$, where $N \cdot M$ is the amount of nucleotide bound to *Msh2–Msh6*, M_t is total *Msh2–Msh6* concentration, N_f is free nucleotide concentration, and K_1 and K_2 are apparent association constants.

The clamp loading activity of RFC was measured by quantitating PCNA loading onto circular DNA using ³²P-labeled PCNA [20]. PCNA^{PK} (2 μM), a derivative of PCNA containing an N-terminal kinase recognition site [21], was phosphorylated in a 100-μl reaction using 50 units of cAMP-dependent protein kinase (New England Biolabs) and 20 μCi [γ-³²P]ATP in kinase buffer (50 mM Tris–HCl (pH 7.5), 10 mM MgCl₂) for 1 h at 37 °C (RFC^{HK5} complex was ³²P-labeled in similar fashion). Excess [γ-³²P]ATP was removed by filtration through Centricon-10 (Millipore) [22]. For the loading reaction, 5.5 nM [³²P]PCNA (trimer) was mixed with 8.5 nM M13mp18 ssDNA (primed with a 30-mer DNA oligonucleotide), 15 μg SSB, and 1 mM ATP in 100 μl Buffer E (30 mM Tris–HCl (pH 7.5), 8 mM MgCl₂, 5% glycerol, 0.1 mg/ml bovine serum albumin, 1 mM dithiothreitol,

0.1 mM EDTA containing 15 mM NaCl). The reaction was initiated with 4.2 nM RFC, incubated at 30 °C for 10 min, and then filtered over a 5-ml Bio-Gel A-15m column equilibrated in Buffer E containing 100 mM NaCl. Fractions of 200 µl were collected and [³²P]PCNA was quantitated by scintillation counting.

Active-site concentration of RFC was assayed by measuring the stoichiometry of DNA binding to the protein complex [20]. DNA binding reactions (15 µl) contained 0.5 µM ³²P-primed DNA (30-mer primer annealed to a DNA 81-mer template) and 0–2 µM RFC in Buffer F (30 mM Hepes–NaOH (pH 7.5), 4 mM MgCl₂, 5% glycerol) and were incubated for 10 min at 25 °C. Aliquots of 10 µl were filtered through nitrocellulose membranes (which were washed before and after filtration with 120 µl Buffer F). The molar amount of ³²P-DNA bound to RFC was quantitated and analyzed as above.

Results

Overexpression of eukaryotic protein complexes in *E. coli*

The pET overexpression system is an extremely powerful and popular method for large-scale production of bacterial and archaeal proteins, in addition to eukaryotic proteins in *E. coli*. However, problems with poor protein expression and solubility have limited its use in the production of large, multiprotein complexes—particularly from eukaryotic organisms. Codon bias is an important reason for poor eukaryotic/heterologous protein expression in *E. coli*. For example, genes encoding many *S. cerevisiae* proteins contain AGA and AGG (arginine), CUA (leucine), or AUA (isoleucine) codons that are rarely used in *E. coli*, and the correspondingly low levels of tRNAs are thought to limit translation of these proteins in *E. coli*[7]. In addition, poor protein solubility is especially common with components of multiprotein complexes, likely because the proteins misfold/aggregate when produced in the absence of their interacting partners [23]. One way to overcome these problems is to co-produce the rare tRNAs and the interacting protein partners of a complex in *E. coli*. This can be achieved by incorporating genes for the rare tRNAs, and for subunits of a multiprotein complex, into a single plasmid. However, given the typically large size of eukaryotic protein complexes, such plasmids can be unduly difficult to prepare and manipulate. It would be simpler to incorporate the genes into multiple plasmids that can be maintained stably within the same cell. To this end, we have developed plasmid vectors that are compatible with the pET vector system for simultaneous production of rare tRNAs and several different proteins in *E. coli*. As demonstrated here for *S. cerevisiae* DNA metabolic proteins, *Msh2–Msh6* and five-subunit RFC, milligram quantities of active eukaryotic protein complexes from *E. coli* can be obtained relatively simply with this approach.

Maintenance of two different plasmids in the same cell requires that they have distinct and compatible replication origins, as competition between plasmids containing the same origin during replication and segregation results in plasmid loss [18]. In addition, the two plasmids must have different antibiotic resistance genes for selection of cells containing both plasmids. Thus, a pET-compatible vector was prepared with a p15A origin of replication (ColE1 replication origin compatible) and a kanamycin resistance gene. This vector was named pLANT to reflect the general compatibility between plants and pets in contrast to two pets which may fight (Fig. 1). The pLANT-2 vector diagrammed in Fig. 1 contains a T7 expression unit with a T7 promoter, ribosome binding sequence, T7 terminator, and multiple restriction site sequence for gene insertion derived from pET(11a) or pET(16b). The pLANT-3 vector (not shown) is similar to pLANT-2, except that it contains the pET(16b)-derived T7 expression unit, which allows production of His-tagged target proteins. In addition, genes encoding rare tRNAs for arginine, isoleucine, and leucine (termed “RIL”) were incorporated into pLANT (forming pLANT-2/RIL (Fig. 1) or pLANT-3/RIL) to facilitate overexpression of heterologous proteins in *E. coli*.

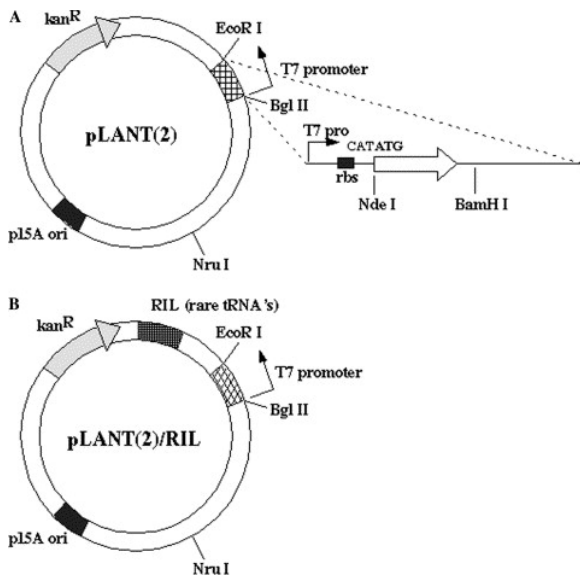


Fig. 1. Schematic of pLANT vector construction. (A) The pLANT-2 vector is derived from pET-9c (Kan^R) by replacement of the ColE1 replication origin with the p15A origin from pACYC184. The pLANT-2 vector also incorporates a T7 expression unit derived from pET-11a. This region is expanded in the figure to illustrate the sites within it and the initiating codon (boldface). (B) The pLANT-2/RIL vector includes genes encoding tRNAs for rare arginine, isoleucine, and leucine codons (denoted as RIL in the figure).

S. cerevisiae Msh2–Msh6 mismatch repair protein was chosen to test the pLANT+pET system for large-scale production of eukaryotic protein complexes in *E. coli*. Msh2 and Msh6 form a 1:1, 248-kDa heterodimer that recognizes basepair mismatches and small insertion/deletion loops in DNA and initiates postreplication DNA mismatch repair [24]. These proteins have been purified previously from *S. cerevisiae*, but in very low quantities, and co-expression of Msh2 and Msh6 is necessary because Msh6 is insoluble when overproduced alone [6], [25] (although Msh2 appears to be partially soluble). Msh2 and Msh6 genes were inserted into the NdeI/BamHI sites of pET-11a and pLANT-2/RIL vectors, respectively, and coexpressed in *E. coli* BLR(DE3) cells (Fig. 2). Both proteins are overproduced in *E. coli* and Fig. 2A shows that Msh2 is expressed in similar quantities from the low-copy pLANT plasmid or the high-copy pET plasmid. Fig. 2B shows that individual expression of the proteins yields partially soluble Msh2 but insoluble Msh6 protein. Combined expression of Msh2 and Msh6, using pET(11a)–Msh2+pLANT-2/RIL–Msh6 results in an apparent Msh2–Msh6 complex in the soluble fraction of the cell lysate (Fig. 2B). To determine whether soluble Msh2–Msh6 complex is produced, the cell lysate was fractionated over three different ion-exchange columns. Fig. 2C shows the three-column chromatography purification profile of Msh2–Msh6, which yields approximately 1 mg of >95% pure protein complex per liter of *E. coli* cells.

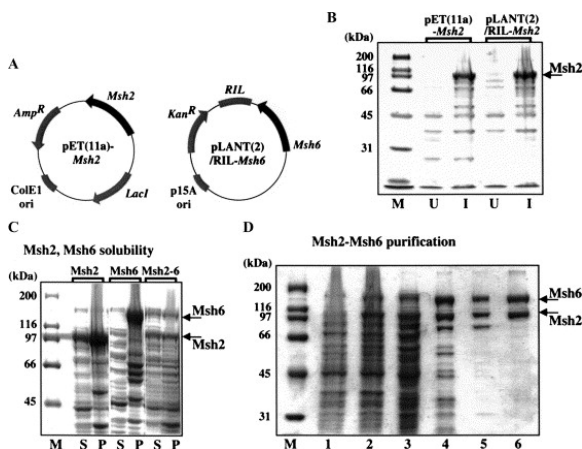


Fig. 2. Overproduction of *S. cerevisiae* Msh2–Msh6 complex in *E. coli*. (A) The two expression plasmid used to produce Msh2–Msh6 complex in *E. coli*. (B) *S. cerevisiae* DNA mismatch repair proteins Msh2 (pLANT-2/RIL–Msh2 and pET(11a)–Msh) and Msh6 (pLANT-2/RIL–Msh6) were expressed individually or together in *E. coli* and tested for solubility, as described under Materials and methods. The SDS–PAGE analysis of Msh2 expression demonstrates that similar levels of protein are produced using either the pET or pLANT vector; U, uninduced cell extract, I, induced cell extract. (C) The solubility profile of Msh2 (partially soluble), Msh6 (insoluble), and Msh2–Msh6 (partially soluble); P, pellet from cell lysate; S, supernatant from cell lysate. (D) A purification profile for Msh2–Msh6 complex: lane 1, uninduced cell extract; lane 2, induced cell extract; lane 3, cleared cell lysate; lane 4, SP-Sepharose eluate; lane 5, heparin eluate; and lane 6, Q-Sepharose eluate.

The general utility and flexibility of this approach to eukaryotic protein complex production was tested with another multiprotein complex, the five-subunit *S. cerevisiae* replication factor C. RFC is a critical component of the DNA replication and repair machinery, as it is responsible for loading circular PCNA clamps at primed DNA sites for processive DNA synthesis [26]. Five proteins, RFC1, 2, 3, 4, and 5 together form the 250-kDa RFC clamp loader (RFC1, 95 kDa; RFC2, 40 kDa; RFC3, 38 kDa; RFC4, 36 kDa; RFC5, 40 kDa) [27]. These proteins are largely insoluble when expressed individually in *E. coli* or in *S. cerevisiae* (Fig. 3A); however, co-expression of all five genes yields soluble RFC complex (Fig. 3B) [19], [28]. The genes encoding RFC2, 3, and 4 proteins were cloned into pET-11a, and genes encoding RFC1 and 5 proteins were placed into the pLANT-2/RIL plasmid containing the rare tRNAs for arginine, isoleucine, and leucine. In the example shown here we use a truncated version of RFC1 (RFC^{1s}) in which the N-terminal 282 amino acids that appear nonessential for DNA replication activity and cell viability are deleted [19]. This RFC complex has been expressed previously in *E. coli* using a single plasmid and has been demonstrated to retain clamp-loading activity [19]. Indeed, studies in the human system show that the corresponding deletion in human RFC1 leads to fivefold greater activity compared to wild-type RFC [29]. Using this pLANT+pET expression system, the five RFC subunits appear in both the soluble and the pellet fractions after cell lysis (Fig. 3C). The smaller subunits (RFC2-5) are expressed in stoichiometric excess over the RFC1; the reason for this is not clear. The fact that the excess small RFC subunits remain soluble is likely based in the fact that they form subassemblies such as RFC34, RFC234, and RFC2345 complex (unpublished). During purification of RFC complex, these subassemblies are well separated from RFC. We obtain approximately 5 mg of 95% pure *S. cerevisiae* RFC^{1s} protein from 1 liter of induced cells (Fig. 3D). This amount is approximately five times the amount of truncated RFC complex reported earlier, which was obtained using a pET plasmid containing all five genes [19]. Our results indicate that use of the low-copy number pLANT plasmid for RFC[1^s+5] expression does not reduce protein synthesis compared to use of pET plasmid, as illustrated above for Msh2. Moreover, the smaller the plasmid (e.g., containing only one or two genes), the simpler it is to introduce specific changes or mutations, thereby facilitating the production of multiprotein complexes with only one subunit modified. For example, the RFC5 gene was easily modified with an N-terminal tag containing a kinase site and multiple histidine residues and combined with the RFC1 gene. This allowed production of RFC containing RFC-5 with a tag (RFC^{HK5} complex), which could be labeled with ³²P (Fig. 3D). Hence, the pLANT+pET vector system facilitates production of mutant multiprotein complexes.

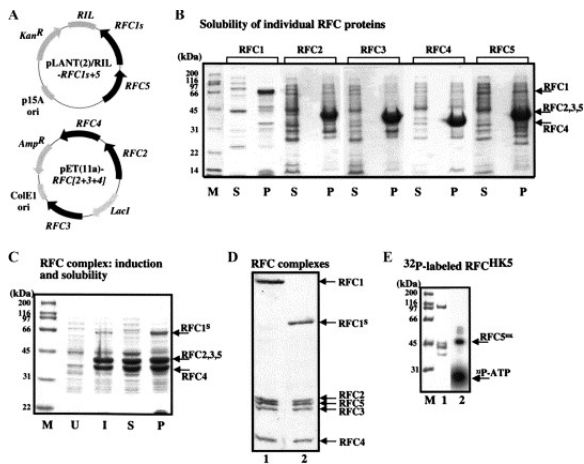


Fig. 3. Overproduction of *S. cerevisiae* RFC complex in *E. coli*. (A) The two expression plasmids used to produce RFC. (B) When expressed individually, the subunits of replication factor C complex are insoluble. (C) Expression and solubility of RFC complex from pLANT-2/RIL-RFC[1⁵+5] + pET(11a)-RFC[2+3+4]. (D) Purified RFC complex (full-length) and RFC¹⁵ complex (truncated RFC1) with all five subunits resolved. (E) RFC^{HK5} complex stained with Coomassie blue (lane 1) and an autoradiogram of the gel demonstrating that the RFC^{5HK} subunit can be labeled with ³²P (lane 2).

S. cerevisiae Msh2–Msh6 and RFC complexes purified from *E. coli* are fully active

Next, we tested *Msh2–Msh6* and RFC¹⁵ complex for activity to determine whether this strategy for eukaryotic multiprotein complex production yields biologically active protein complexes. The *Msh2–Msh6* complex recognizes basepair mismatches in DNA and is known to bind duplex DNA containing a G:T mismatch with high affinity relative to fully matched DNA [30]. Fig. 4A shows a titration of ³²P-labeled 37-nucleotide duplex DNA with increasing concentrations of *Msh2–Msh6*, assayed by nitrocellulose membrane filtration. *Msh2–Msh6* binds G:T DNA with high affinity (apparent $K_d=38$ nM, from experiments performed at lower DNA concentration). Further, saturation is reached at a 1:1 stoichiometric ratio of G:T DNA to *Msh2–Msh6* complex (0.32 μ M DNA:0.34 μ M *Msh2–Msh6*). In contrast, there is very little interaction between *Msh2–Msh6* and fully matched G:C DNA. The active-site concentration of *Msh2–Msh6* was measured by ATP binding assays (ATP γ S was used as a nonhydrolyzable ATP analog). Both *Msh2* and *Msh6* contain Walker A (ATP binding) and B (Mg²⁺ binding) motifs, suggesting that the *Msh2–Msh6* complex binds two molecules of ATP. As illustrated in Fig. 4B, titration of *Msh2–Msh6* (1 μ M) with increasing concentrations of ATP γ S yields a stoichiometry of approximately 2 ATP γ S molecules bound to the *Msh2–Msh6* heterodimer. The two ATP γ S molecules bind *Msh2–Msh6* with differing affinities (apparent $K_d=4$ and 20 μ M), consistent with reports of asymmetry in *Msh2* and *Msh6* function [31]. Thus, the *S. cerevisiae* *Msh2–Msh6* complex purified from *E. coli* appears to be fully active with respect to nucleotide binding and mismatch recognition.

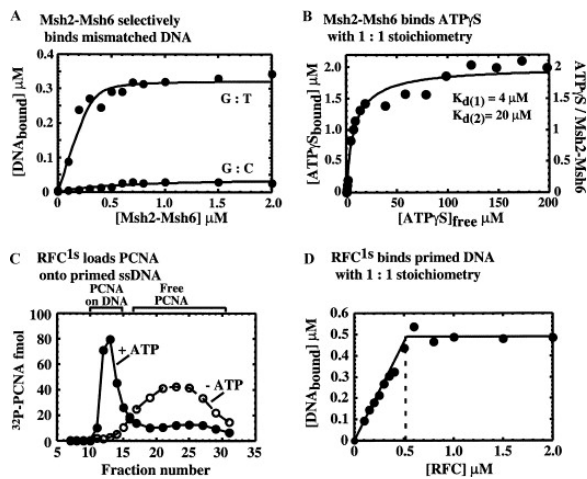


Fig. 4. Activity of *S. cerevisiae* *Msh2–Msh6* and RFC complexes produced in *E. coli*. (A) *Msh2–Msh6* binds with high affinity to duplex DNA containing a G:T mismatch while binding to matched G:C DNA is nearly undetectable, as measured by nitrocellulose membrane assays. (B) Active-site concentration of *Msh2–Msh6* was determined by measuring the stoichiometry of ATP γ S binding to *Msh2–Msh6* (both proteins have Walker ATP-binding motifs). Two ATP γ S molecules bind per *Msh2–Msh6*, indicating that the recombinant protein complex is fully active ($K_d=4$ and $20 \mu\text{M}$ for the two binding sites). (C) Clamp loading activity of RFC¹⁵ complex assayed by assembly of ³²P-labeled PCNA onto circular, primed M13mp18 single-stranded DNA. (D) The active-site concentration of RFC¹⁵, determined by measuring the stoichiometry of primer–template DNA binding to RFC by nitrocellulose membrane assays. RFC¹⁵ binds DNA with near 1:1 stoichiometry, indicating that the recombinant protein is >95% active.

To determine whether the recombinant RFC complex is active, we tested the ability of truncated RFC¹⁵ complex to load ³²P-labeled PCNA onto a large circular M13mp18 ssDNA primed with a single oligonucleotide (Fig. 4C). Assembly of ³²P-PCNA onto the DNA is monitored by analysis of the reaction on a large pore gel filtration resin such as BioGel A-15m. In this analysis, the topologically linked ³²P-PCNA–DNA complex is large compared to ³²P-PCNA alone and elutes early from the column (fractions 7–12), while free ³²P-PCNA elutes later (fractions 14–25). No PCNA assembly occurs in the absence of ATP (Fig. 4C). The stoichiometry of RFC¹⁵ binding to DNA was measured by nitrocellulose membrane binding assays in which $0.5 \mu\text{M}$ ³²P-labeled primer–template DNA (31/81 nucleotides) was mixed with increasing concentrations of RFC¹⁵. Fig. 4D shows that the binding isotherm reaches saturation near a 1:1 ratio of DNA:RFC¹⁵ (0.49:0.51, DNA:RFC¹⁵), indicating that the complex is >95% active for binding the DNA substrate.

Discussion

Eukaryotic proteins are generally less well characterized than their prokaryotic counterparts, especially with respect to analysis of structure and detailed mechanism of action. One important contributing factor is the high-milligram quantity of protein required for crystallographic or kinetic studies, and the difficulty in producing such large amounts of protein, especially multicomponent protein complexes, in eukaryotic organisms. *E. coli* is used widely as a host organism to overproduce bacterial proteins because it is easy and inexpensive to grow, and an in-depth understanding of this organism has led to development of many useful strains and cloning vectors. In recent years, a growing list of examples suggests that *E. coli* may be just as useful for eukaryotic/heterologous protein production [9], [15]. Several of the problems associated with eukaryotic protein production in *E. coli* are now being addressed successfully. For example, heterologous proteins produced in high amounts tend to misfold and aggregate, but their solubility can be enhanced by coexpression of molecular chaperones [32]. In the case of multiprotein complexes, coexpression of interacting protein partners may be sufficient for proper folding and assembly [10], [12], [13]. In another example, differences in codon usage in prokaryotes and

eukaryotes (which can substantially reduce protein production) can be resolved by replacement of rare codons in the genes, but more easily by co-expression of rare tRNA genes in *E. coli* [16]. Thus, the strategy of co-expression can be useful in many cases and has been utilized here to enhance production of large, multiprotein complexes from *S. cerevisiae* in *E. coli* (Fig. 1, Fig. 2). The vectors developed in this study make it easy to co-express multiple proteins and rare tRNAs in *E. coli*, and our success with production of two very different protein complexes, *Msh2–Msh6* and replication factor C, indicate that this strategy is likely to be generally applicable.

For the *Msh2–Msh6* complex, the yields have improved from 0.1–0.2 mg pure protein/g *S. cerevisiae* cells [30] (which are a lot more cumbersome and expensive to grow than *E. coli*) to 0.5–0.8 mg pure protein/g *E. coli* cells. More importantly, the recombinant protein complex appears to be fully active when assayed for mismatch recognition and ATPyS binding activity (Fig. 4). The steady state ATPase activity of *E. coli*-produced *Msh2–Msh6* is also consistent with earlier reports [31] ($k_{cat}=0.5\text{ s}^{-1}$ at 30 °C, data not shown).

For the RFC complex, the yield is approximately 2.5 mg pure protein/g *E. coli* cells which is a substantial improvement over the previously reported yield of 0.02 mg/g *S. cerevisiae* cells [28] and the more recently reported yield of 0.5 mg/g *E. coli* cells of a truncated version of RFC [19] (In the previous study, the truncated *RFC1* gene was combined with genes for *RFC2*, 3, 4, 5 in a single pET-based expression vector and co-expressed with the rare *argU* tRNA in *E. coli*). As observed in the case of *Msh2–Msh6*, the five-subunit RFC complex also appears to be fully active for binding its substrates, DNA and PCNA, and is therefore suitable for further crystallographic and kinetic analysis.

In addition to improving wild-type eukaryotic protein complex production in *E. coli*, the pLANT+pET coexpression system facilitates modification/mutation of specific subunits within the complex. As demonstrated for His/kinase-tagged RFC^{HK5} complex, it is relatively easy to modify one gene in a plasmid containing only two or three genes (pLANT-2/RIL–RFC[1+5]), rather than in a plasmid containing all five genes. This two-vector strategy also facilitates systematic co-expression of various subunit proteins for identification and production of soluble sub-complexes, which can be extremely valuable in characterizing the structure and function of multiprotein assemblies [33] such as RFC (M. O'Donnell and M.M. Hingorani, unpublished work).

In summary, a better understanding of protein expression and folding in *E. coli* and concomitant improvements in the tools of recombinant protein production are making *E. coli* an attractive host for eukaryotic protein production, especially for hard-to-purify multicomponent protein complexes. Availability of the *Msh2–Msh6* and RFC complexes and other proteins produced by the pLANT+pET strategy (e.g., *S. cerevisiae* MCM proteins [34]) is expected to greatly enhance our investigation into the workings of DNA metabolic proteins in eukaryotic organisms. We anticipate that this strategy will prove useful for a variety of protein complexes, from *S. cerevisiae* and other eukaryotic organisms as well.

Acknowledgements

This work was supported by NIH Grants GM38839 (M.O'D) and GM64514-01 (M.M.H). We thank Lee Coryell, Max Stadler, Magda Coman, and Nina Yao for help with experiments. We will be happy to provide the pLANT vectors described herein upon request.

References

- [1] Z. Kelman, M. O'Donnell. *Annu. Rev. Biochem.*, 64 (1995), pp. 171-200
- [2] P. Modrich, R. Lahue. *Annu. Rev. Biochem.*, 65 (1996), pp. 101-133
- [3] S. Jacob, F. Praz. *Biochimie*, 84 (2002), pp. 27-47
- [4] D. Jeruzalmi, M. O'Donnell, J. Kuriyan. *Cell*, 106 (2001), pp. 429-441
- [5] K. Luger, A.W. Mader, R.K. Richmond, D.F. Sargent, T.J. Richmond. *Nature*, 389 (1997), pp. 251-260

- [6] H. Lilie, E. Schwarz, R. Rudolph. *Curr. Opin. Biotechnol.*, 9 (1998), pp. 497-501
- [7] K. Zahn. *J. Bacteriol.*, 178 (1996), pp. 2926-2933
- [8] M.D. Forman, R.F. Stack, P.S. Masters, C.R. Hauer, S.M. Baxter. *Protein Sci.*, 7 (1998), pp. 500-503
- [9] F. Baneyx. *Curr. Opin. Biotechnol.*, 10 (1999), pp. 411-421
- [10] C. Li, J.W. Schwabe, E. Banayo, R.M. Evans. *Proc. Natl. Acad. Sci. USA*, 94 (1997), pp. 2278-2283
- [11] A. Skerra, A. Pluckthun. *Science*, 240 (1988), pp. 1038-1041
- [12] M. Ishiai, J.P. Sanchez, A.A. Amin, Y. Murakami, J. Hurwitz. *J. Biol. Chem.*, 271 (1996), pp. 20868-20878
- [13] L.A. Henriksen, C.B. Umbricht, M.S. Wold. *J. Biol. Chem.*, 269 (1994), pp. 11121-11132
- [14] S. Tan. *Protein Expr. Purif.*, 21 (2001), pp. 224-234
- [15] S.C. Makrides. *Microbiol. Rev.*, 60 (1996), pp. 512-538
- [16] G. Dieci, L. Bottarelli, A. Ballabeni, S. Ottonello. *Protein Expr. Purif.*, 18 (2000), pp. 346-354
- [17] F.W. Studier, A.H. Rosenberg, J.J. Dunn, J.W. Dubendorff. *Methods Enzymol.*, 185 (1990), pp. 60-89
- [18] J. Sambrook, D.W. Russell. **Molecular Cloning: A Laboratory Manual**. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY (2001)
- [19] X.V. Gomes, S.L. Gary, P.M. Burgers. *J. Biol. Chem.*, 275 (2000), pp. 14541-14549
- [20] M.M. Hingorani, M.M. Coman. *J. Biol. Chem.*, 277 (2002), pp. 47213-47224
- [21] R. Onrust, J. Finkelstein, V. Naktinis, J. Turner, L. Fang, M. O'Donnell. *J. Biol. Chem.*, 270 (1995), pp. 13348-13357
- [22] Z. Kelman, V. Naktinis, M. O'Donnell. *Methods Enzymol.*, 262 (1995), pp. 430-442
- [23] F. Baneyx. J.E. Davies, A.L. Demain, G. Cohen, C.L. Hershberger, L.J. Forney, I.B. Holland, W.-S. Hu, J.-H. Wu, D.H. Sherman, R.C. Wilson (Eds.), *Manual of Industrial Microbiology and Biotechnology*, American Society for Microbiology, Washington DC (1999), pp. 551-565
- [24] J. Jiricny. *EMBO J.*, 17 (1998), pp. 6427-6436
- [25] E. Alani. *Mol. Cell Biol.*, 16 (1996), pp. 5604-5615
- [26] S. Waga, B. Stillman. *Annu. Rev. Biochem.*, 67 (1998), pp. 721-751
- [27] G. Cullmann, K. Fien, R. Kobayashi, B. Stillman. *Mol. Cell Biol.*, 15 (1995), pp. 4661-4671
- [28] K.J. Gerik, S.L. Gary, P.M. Burgers. *J. Biol. Chem.*, 272 (1997), pp. 1256-1262
- [29] F. Uhlmann, J. Cai, E. Gibbs, M. O'Donnell, J. Hurwitz. *J. Biol. Chem.*, 272 (1997), pp. 10058-10064
- [30] G.T. Marsischky, R.D. Kolodner. *J. Biol. Chem.*, 274 (1999), pp. 26668-26682
- [31] B. Studamire, T. Quach, E. Alani. *Mol. Cell Biol.*, 18 (1998), pp. 7590-7601
- [32] J.G. Thomas, A. Ayling, F. Baneyx. *Appl. Biochem. Biotechnol.*, 66 (1997), pp. 197-238
- [33] S. Fribourg, C. Romier, S. Werten, Y.G. Gangloff, A. Poterszman, D. Moras. *J. Mol. Biol.*, 306 (2001), pp. 363-373
- [34] M.J. Davey, C. Indiani, M. O'Donnell. *J. Biol. Chem.*, 11 (2002), p. 11

¹Abbreviations used: Msh, MutS homologue; RFC, replication factor C.