Published in final edited form as:

Nat Biotechnol. ; 30(5): 405–406. doi:10.1038/nbt.2207.

Direct cloning of large genomic sequences

Ryan E. Cobb¹ and Huimin Zhao^{1,2}

Huimin Zhao: zhao5@illinois.edu

¹Department of Chemical and Biomolecular Engineering, University of Illinois at Urbana-Champaign, Urbana, IL

²Institute for Genomic Biology, Departments of Chemistry, Biochemistry, and Bioengineering, University of Illinois at Urbana-Champaign, Urbana, IL

Abstract

The discovery of an efficient mechanism of homologous recombination between two linear DNA substrates enables direct cloning of large genomic sequences.

Methods for recovering genomic sequences of interest are among the most important tools in biotechnology, but many require laborious library generation and screening, several preparatory DNA amplification and assembly steps, or cost-prohibitive direct DNA synthesis. In this issue, Fu *et al.*¹ introduce a new cloning technique that allows for the direct transfer of large regions of genomic DNA into an expression vector. They applied the approach to clone ten regions 10–52 kb in length from the genome of *Photorhabdus luminescens* into an *Escherichia coli* expression vector and used the constructs to rapidly identify two novel secondary metabolites. This direct cloning approach has many potential applications, including rapid heterologous expression of known biosynthetic gene clusters in tractable hosts, bioprospecting for novel secondary metabolites, *de novo* gene cluster assembly and construction of cDNA libraries.

The ever-increasing numbers of sequenced genomes and metagenomes represent a rich source of secondary metabolite pathways for producing novel antibiotics and chemotherapeutics. But cloning these gene clusters is not trivial because most are quite large. Traditional cloning techniques are laborious and time-consuming as they involve the construction and screening of genomic DNA libraries to identify pathways of interest. In addition, many genomes contain ‘cryptic’ gene clusters, which encode pathways whose final products are unknown or have not been detected. Most approaches for investigating cryptic gene clusters rely on eliciting their expression in the native host. For example, the ‘one strain many compounds’ method² entails the manipulation of culture parameters, such as media composition and aeration, to identify conditions that favor expression of secondary metabolite clusters. Other strategies include co-culture of the organism of interest with additional species or the use of epigenetic factors to modulate secondary metabolite regulation³. However, it is often difficult to predict which conditions, if any, will lead to expression of a particular cryptic gene cluster. A more targeted approach is over-expression of pathway-specific regulators, but this is only applicable in genetically tractable hosts.

A more recent direction is heterologous expression of cryptic clusters in a well-characterized host using sophisticated DNA manipulation techniques. For example, the Gibson assembly method⁴ enables one-step, isothermal assembly of large DNA molecules from several much smaller fragments, which can be prepared via standard PCR or direct synthesis. Similar *in vitro* assembly protocols include sequence- and ligation-independent cloning⁵ and circular polymerase extension cloning⁶. The Golden Gate assembly method⁷ uses unique Type II restriction enzymes to facilitate assembly *in vitro*, whereas the DNA assembler method⁸ uses the natural homologous recombination proficiency of *Saccharomyces cerevisiae* to assemble DNA fragments *in vivo*. Although each of these approaches bypasses the need for a traditional library generation and screening approach to isolate a gene cluster of interest, they nevertheless require multiple PCR amplification steps that may introduce random mutations. Further, mutations, insertions or deletions can also be introduced during the assembly process itself via incorrect pairing of fragments.

The cloning method presented by Fu *et al.*¹ is predicated on their discovery that there exists a functional distinction between homologous recombination mechanisms involving one linear and one circular DNA molecule and those involving two linear DNA molecules. In a ‘recombineering’ experiment, an exonuclease and a single-strand annealing protein—typically the lambda phage protein pair Red α and Red β , respectively—are used to facilitate recombination between a circular DNA molecule (i.e., the bacterial chromosome) and a linear DNA fragment. Other proteins, such as the Rac phage proteins RecE and RecT, can also serve in this capacity, albeit with lower recombination efficiency. Interestingly, in contrast to the 226-amino-acid Red α protein, RecE is a much larger protein of 866 amino acids, of which only the last 260 amino acids constitute the exonuclease domain. In their new study, Fu *et al.*¹ for the first time demonstrate that the full-length RecE protein can facilitate efficient recombination between two linear DNA substrates. Although both the truncated RecE-RecT pair (in which only the exonuclease domain of RecE is expressed) and the Red α -Red β pair are superior at standard recombineering compared with full-length RecE and RecT, the latter catalyze homologous recombination between two linear DNA molecules >20 times more efficiently.

This finding has immediate implications for the cloning of linear DNA constructs into linear vectors. The authors demonstrated the capabilities of their method by direct cloning of sequences from bacterial artificial chromosomes, cDNA and genomic DNA. To illustrate the application of direct cloning to genome mining—the discovery of novel secondary metabolites solely from genomic sequence data—they chose the Gram-negative bacterium *P. luminescens*. Bioinformatics analysis of the sequenced genome identified a total of ten megasynthetase clusters, which are of significant biotechnological interest due to their potential to produce a variety of useful compounds such as antibiotics, chemotherapeutics, immunosuppressants and insecticides. For each of the ten clusters, the genomic DNA was digested with restriction enzymes that cleave upstream and downstream of the cluster, and a standard expression vector was amplified with a unique pair of oligonucleotides to introduce ends homologous to the boundaries of the cluster. Transformation of the linearized vector and digested genomic DNA into *E. coli* expressing full length RecET enabled direct cloning of nine of the ten clusters, although two were found to be consistently mutated at their 5'

ends. Of the remaining seven, two produced detectable levels of compounds ($>0.1 \mu\text{g/ml}$), which were purified and identified as the secondary metabolites luminmycin A and luminmide A/B.

Cloning the tenth and largest (52-kb) locus required a two-step procedure. After full-length RecE-RecT cloning, most of the incorrect plasmids were recircularized empty vectors. To overcome this background, a second recombineering step was added in which a second antibiotic selection marker was introduced on a linear DNA fragment with homology arms corresponding to part of the target sequence. To facilitate recombination between this linear fragment and the correctly cloned circular plasmid, the Red α -Red β recombineering proteins were expressed, and selection was carried out in the presence of the second antibiotic. This amended protocol allowed successful cloning of the largest target gene.

Direct cloning as shown by Fu *et al.*¹ offers many advantages over other cloning and assembly methods (Fig. 1). Compared with traditional random cloning methods, it circumvents the intermediate steps of library generation and screening to identify the target cluster. Compared with direct assembly methods, it does not require PCR amplification from genomic DNA, minimizing the chance of introducing mutations. In addition, the use of a standard expression vector with a tetracycline-inducible promoter prevents constitutive expression of potentially toxic metabolites, and only a new pair of primers is needed to target any new cluster. Finally, both cloning and expression are carried out in *E. coli*, a well-studied and highly tractable host.

The method also has several limitations. First, it requires the identification of unique restriction enzyme sites upstream and downstream of the target gene cluster. In particular, the 5' site must be close to the start of the first gene to achieve the highest recombination efficiency. Second, efficiency is limited by the size of the target cluster, as seen with the 52 kb locus. Even with an additional selection step, only 29 % recombination efficiency (6/21 correct clones) was observed, suggesting that significantly larger contiguous DNA regions may be inaccessible by this method. Third, preservation of the genetic context of the gene cluster could be detrimental to heterologous expression for various reasons, such as the presence of repressors or the absence of necessary activators. Fourth, if the gene cluster is not completely operonic in structure, the heterologous host might not recognize the promoters of the native host or might express the genes at unbalanced levels. For gene clusters from eukaryotes, introns may not be processed correctly. Finally, direct cloning is not readily amenable to modification of the target gene cluster and thus requires supplemental techniques for further engineering.

As direct cloning is optimized in future studies, it would be of interest to determine whether it can be extended to multiple genomic fragments, such as additional promoters or genes not found in the target gene cluster, or even to libraries of pathway variants. It would also be interesting to investigate whether similar recombination systems can be applied in hosts other than *E. coli*. Regardless of subsequent improvements, direct cloning is clearly a valuable new method capable of mining the vast amount of new genome and metagenome sequence data. As such, it will likely prove very useful for secondary metabolite discovery and synthetic biology applications.

References

1. Fu J, et al. *Nat Biotechnol.* 2012 [PE: Update at finals; ZHANG, nbt.2183].
2. Bode HB, Bethe B, Höfs R, Zeeck A. *ChemBioChem.* 2002; 3:619–627. [PubMed: 12324995]
3. Scherlach K, Hertweck C. *Org Biomol Chem.* 2009; 7:1753–1760. [PubMed: 19590766]
4. Gibson DG, et al. *Nat Methods.* 2009; 6:343–345. [PubMed: 19363495]
5. Li MZ, Elledge SJ. *Nat Methods.* 2007; 4:251–256. [PubMed: 17293868]
6. Quan J, Tian J. *PLoS ONE.* 2009; 4:e6441. [PubMed: 19649325]
7. Engler C, Kandzia R, Marillonnet S. *PLoS ONE.* 2008; 3:e3647. [PubMed: 18985154]
8. Shao Z, Zhao H, Zhao H. *Nucleic Acids Res.* 2009; 37:e16. [PubMed: 19074487]

Figure 1.

Strategies for heterologous expression of a gene cluster from genomic DNA. The traditional approach (left) requires preparation and screening of a library, whereas direct cloning (center) bypasses these steps. Alternative approaches based on recently developed synthetic biology techniques (right) allow more sophisticated manipulations but require additional steps.