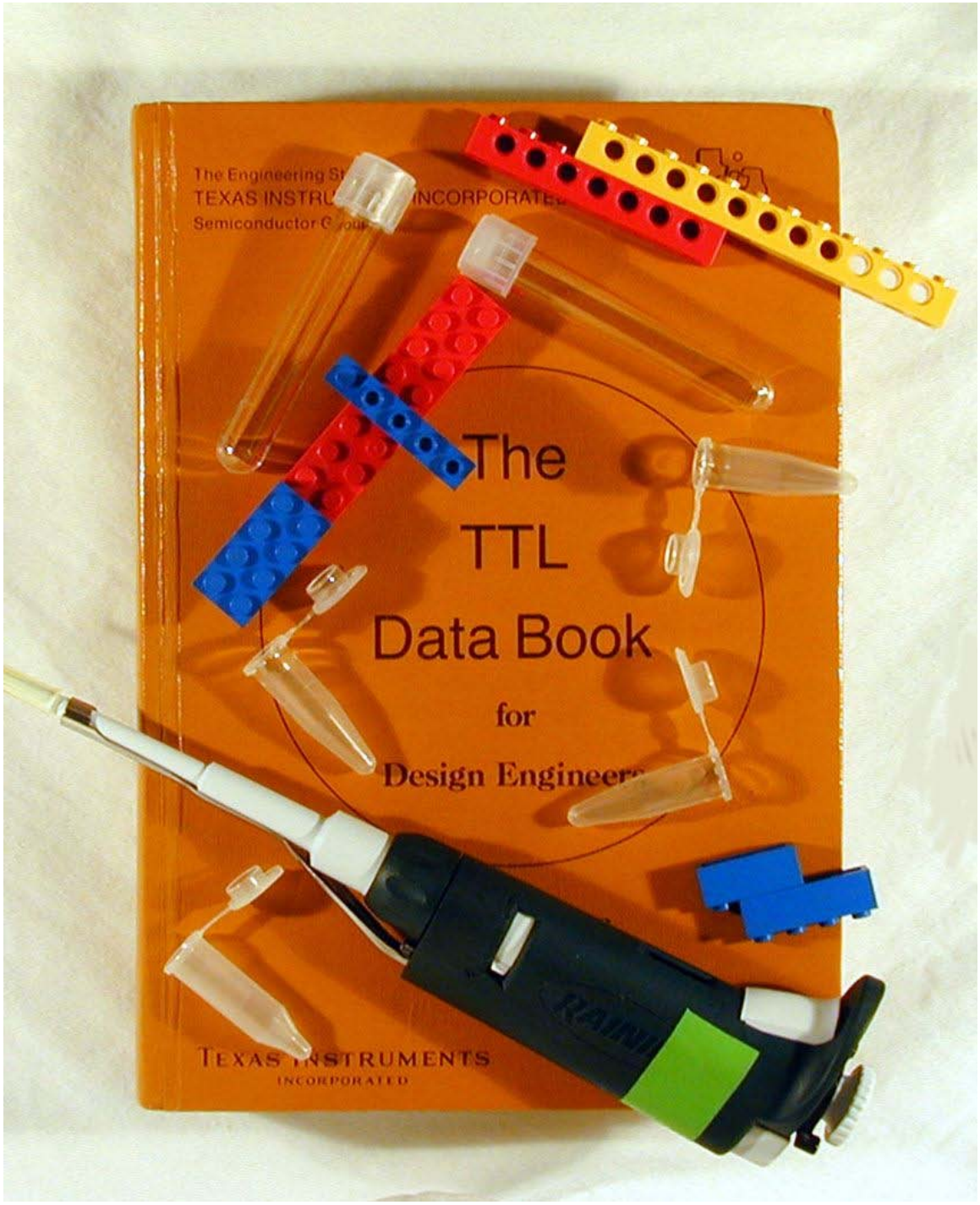
Idempotent Vector Design for Standard Assembly of Biobricks

Tom Knight
MIT Artificial Intelligence Laboratory



Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2003		**00-00-2003 to 00-00-2003**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Idempotent Vector Design for Standard Assembly of Biobricks	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Massachusetts Institute of Technology,77 Massachusetts Avenue,Cambridge,MA,02139-4307**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**		**11**	

0.1 Motivation

The lack of standardization in assembly techniques for DNA sequences forces each DNA assembly reaction to be both an experimental tool for addressing the current research topic, and an experiment in and of itself. One of our goals is to replace this *ad hoc* experimental design with a set of standard and reliable engineering mechanisms to remove much of the tedium and surprise during assembly of genetic components into larger systems.

William Sellers, in a speech *On a Uniform System of of Screw Threads* at the Franklin Institute in Philadephia on April 21, 1864, remarked that "In this country, no organized attempt has as yet been made to establish any system, each manufacturer having adopted whatever his judgement may have dictated as best, or as most convenient for himself" (Surowiecki 02). He argued forcefully and successfully for the standardization of pitch, diameter, and form of screw threads, providing the infrastructure which allowed the industrial revolution to take off. The machinists of the Franklin Institute built the lathes, drills, taps, and dies necessary to make the standard a success.

Figure 1: A US Standard Screw Thread

We anticipate advantages similar to those which accompany the standardization of screw threads in mechanical design – the widespread ability to interchange parts, to assemble sub-components, to outsource assembly to others, and to rely extensively on previously manufactured components. Here, we present a simple sequence and assembly standard as part of an experiment to see how far this idea of standardized interface technology can be applied.

The key notion in the design of our strategy is that the transformations performed on component parts during the assembly reactions are *idempotent* in a structural sense. That is, each reaction leaves the key structural elements of the component the same. The output of any such transformation, therefore, is a component which can be used as the input to any subsequent manipulation. It need never be constructed again – it can be added to the permanent library of previously assembled components, and used as a compound structure in more complex assemblies.

0.2 Standard Biobrick Sequence Interface

We have chosen a very simple but powerful sequence standard for biobrick components. Each component consists of a circular vector of double stranded DNA containing the comonent regulatory sequence, flanked on the upstream end by EcoRI and XbaI restriction sites, and on the downstream end by SpeI and PstI restriction sites. We require that the component vector not contain other occurences of these restriction sites. Any such sites must be removed by point mutation prior to assembly with this technique. Techniques for inserting the flanking cut sites and performing point mutations are described below, but we will assume this sort of manipulation for the time being.

Thus a typical component vector consists of a sequence of the following form:

```
5' --gca GAATTC GCGGCCGC T TCTAGA G --insert-- T ACTAGT A GCGGCCG CTGCAG gct--
   --cgt CTTAAG CGCCGGCG A ACATCT C ---------- A TGATCA T CGCCGGC GACGTC cga--
         EcoRI   NotI       XbaI                  SpeI     NotI    PstI
```

The single bases between restriction sites are carefully chosen to eliminate the accidental generation of EcoBI and EcoKI methylation sites, which could prevent cutting with each of these enzymes in mB+ and mK+ strains (most laboratory strains), such as DH5α.

Each vector may be usefully cut in four distinct ways yielding four useful fragments. Cutting with EcoRI and SpeI creates a front insert (FI). Cutting with XbaI and PstI creates a back insert (BI).

Cutting with EcoRI and XbaI creates a front vector (FV). Cutting with SpeI and PstI creates a back vector (BV).

Because of the compatible overhangs of the XbaI and SpeI recognition sequences (TCTAGA and ACTAGT), back inserts can ligate with back vectors to add components to the back of existing constructs. Similarly, front inserts can ligate with front vectors to add components to the front of existing constructs. In the ligation process, a mixed SpeI/XbaI site, with sequence ACTAGA is formed at the junction of the insertions. Since this site is not a recognition site for any of the enzymes, it cannot be further cut.

Importantly, then, the result of either such insertion is a construct which is identical in form to our standard component, having exactly the same restriction sites as the parent components. The ability to prefix or postfix components to other components, creating components of the same form provides much of the power of this technique.

An example of a standard component is the vector pSB103-ECFP, containing a base vector with an insertion of the ribosomal binding site and coding sequence for the Clontech ECFP cyan fluorescent protein, with the appropriate restriction sites, as shown in figure 2.

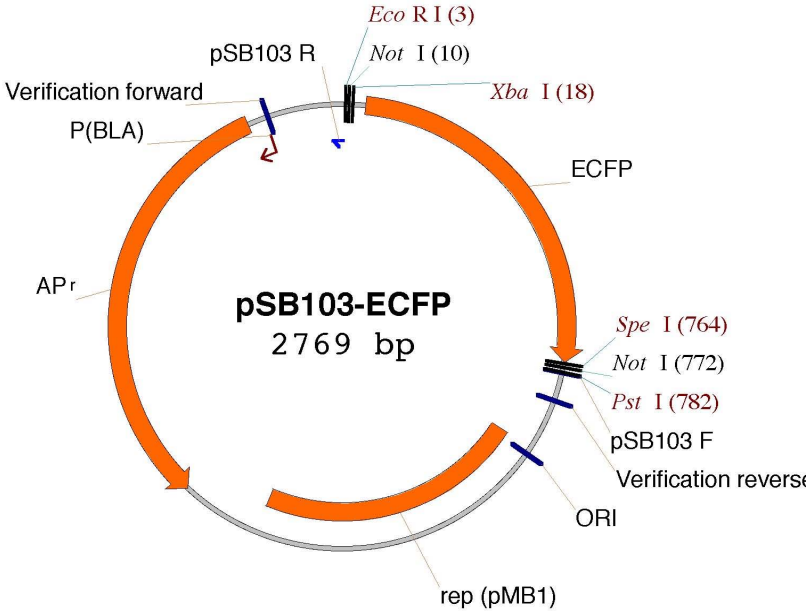


Figure 2: Standard form of the pSB103-ECFP component

A second example is the vector containing the Plac promoter, pSB103-Plac, shown below, having an identical structure (figure 3).

0.3 Standard Assembly Vector Sequence Interface

We have constructed an assembly vector, pSB103, which is a high copy number ampicillin resistant plasmid, derivative of pUC18 (Yanisch-Perron 1985), missing the LacZ gene, promoter, and MCS, which are replaced by a cloning site of the following sequence:

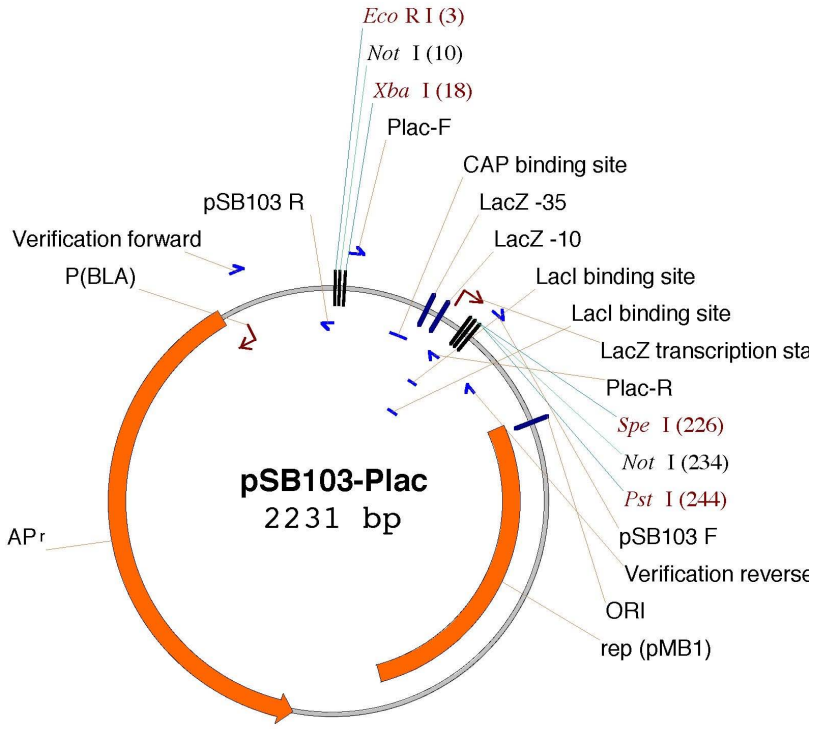


Figure 3: Standard Component form of the Plac promoter sequence

```
5' --gca GAATTC GCGGCCGC T TCTAGA GT ACTAGT A GCGGCCG CTGCAG gct---  3'
3' --cgt CTTAAG CGCCGGCG A ACATCT CA TGATCA T CGCCGGC GACGTC cga---  5'
         EcoRI   NotI       XbaI       SpeI     NotI    PstI
```

0.4 Biobrick Composition Techniques

Composition of biobrick components is always performed in one of two ways, either by prefixing one component with another, or by postfixing one component with another. In each case, the result is a new, compound component, which can then be used in the same way, as either an insert or a vector, in more complex reactions.

0.4.1 Prefixing a recipient with a donor component

To prefix a recipient component, the component is cut with EcoRI and XbaI enzymes. This yields a front vector (FV) with cut sites of the following form:

```
5' --gca G                *CTAGA G---- 3'
3' --cgt CTTAA*                T C---- 5'
         EcoRI              XbaI
```

The donor component is cut with EcoRI and SpeI, yielding a front insert (FI) of the following form:

```
5' *AATTC GCGGCCGC T TCTAGA G --Insert-- T A      3'
3'      G CGCCGGCG A ACATCT C --Insert-- A TGATC* 5'
   EcoRI   NotI       XbaI                  SpeI
```

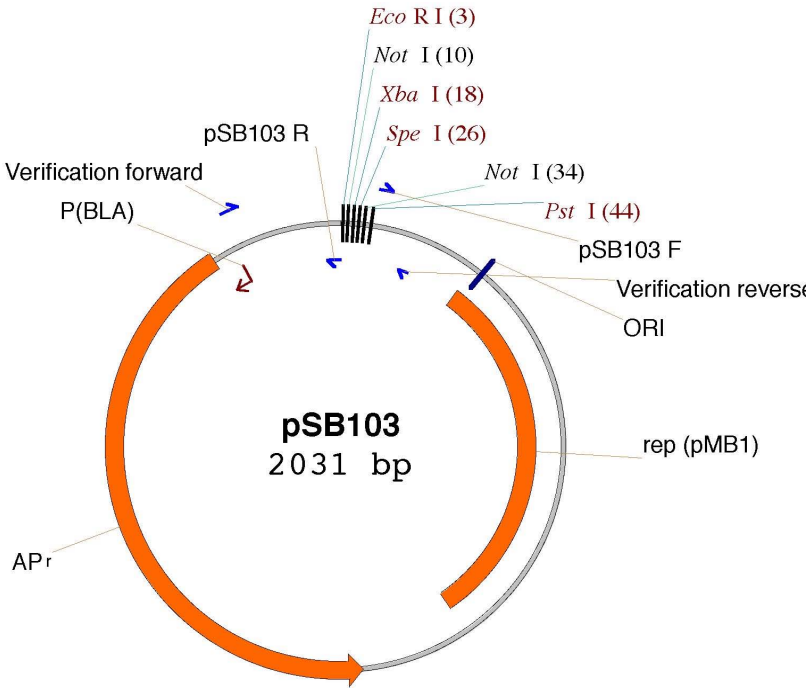


Figure 4: Structure of the pSB103 vector

Ligation links the EcoRI sites and the mixed SpeI/XbaI site with compatible overhangs, in the following way:

```
5' --gca G  *AATTC GCGGCCGC T TCTAGA G--insert--T A   *CTAGA G---- 3'
3' --cgt CTTAA*  G CGCCGGCG A ACATCT C--insert--A TGATC*   T C---- 5'
         EcoRI       NotI       XbaI                  Mixed
```

This process recreates the EcoRI and XbaI sites at the beginning of the component, and creates the uncuttable mixed SpeI/XbaI site at the junction. The vector continues to carry a SpeI and PstI site following any previous insertions.

0.4.2 Postfix insertion of a donor fragment into a recipient vector

Similarly, to insert a donor fragment after a component on a vector, we create the back vector (BV) with SpeI and PstI, yielding a vector of the following form:

```
5' --T A                   *G gct---  3'
3' --A TGATC*            ACGTC cga---  5'
       SpeI              PstI
```

Cutting the donor component with XbaI and PstI yields the following back insert (BI):

```
5' *CTAGA G --insert-- T ACTAGT A GCGGCCG CTGCA 3'
3'      T C --insert-- A TGATCA T CGCCGGC G*    5'
    XbaI                 SpeI     NotI    PstI
```

Ligation of these products yields the following structure:

```
5' --T A  *CTAGA G --insert-- T ACTAGT A GCGGCCG CTGCA  *G gct---  3'
3' --A TGATC*  T C --insert-- A TGATCA T CGCCGGC G*  ACGTC cga---  5'
         Mixed                    SpeI     NotI      PstI
```

Which restores the SpeI and PstI sites, and creates the same uncuttable mixed SpeI/XbaI site the junction of the two inserts. In fact, the exact sequence created is independent of the order in which the front and back inserts are assembled. The vector continues to carry an upstream EcoRI and XbaI sites prior to any previous component insertions.

0.5 Example

We can construct a functional gene expressing ECFP protein from the Plac promoter and ECFP coding sequence found in the vectors described in figure 2 and figure 3. Cutting pSB103-ECFP with EcoRI and XbaI creates a front vector. Cutting pSB103-Plac with EcoRI and SpeI creates a front insert. Ligation of these two components creates pSB103-Plac-ECFP (figure 5), a functional gene expressing cyan fluorescent protein.

Similarly, we can cut pSB103-Plac with SpeI and PstI, creating a back vector, and pSB103-ECFP with XbaI and PstI, creating a back insert. Ligation produces exactly the same sequence, pSB103-Plac-ECFP (figure 5).

The complete gene can be cut out and used as a component in another more complex assembly, by cutting it out to create a front insert (EcoRI and SpeI) or back insert (XbaI and PstI).

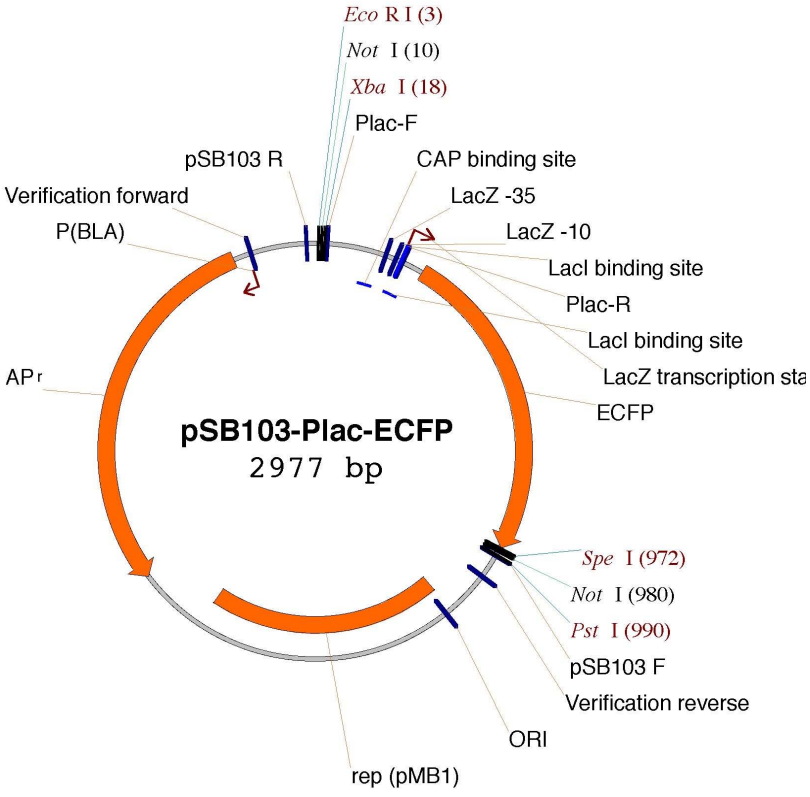


Figure 5: Structure of the pSB103-Plac-ECFP vector

0.6 Other approaches

This approach is the third generation of our simple assembly vectors, each adding functionality to the assembly process. The NOMAD system (Rebatchouk 1996) provides very similar capability, in perhaps an even more elegant form. The major potential advantage of the NOMAD system is the lack of distinction between front and back inserts – there is only a single type of insert. The enzymes used (StyI, BsaI and BsmBI) are substantially less user-friendly than the pSB103 enzyme choices, but the advantages may outweigh the disadvantages. We were unaware of its existence at the time of this work, and will be looking closely at that approach as a possible competitor. DNA and reference information concerning the NOMAD system is unavailable, and we have found no other groups utilizing their approach.

Meanwhile, the library of assembled components in the pSB103 vector is rapidly growing, and this vector systems seems, in our work, to be reliable, easy to use, and time efficient.

0.7 Standards Summary

0.7.1 Vector Sequences

All vectors inserts contain restriction sites at the upstream end with the following sequence:

```
5' --gca GAATTC GCGGCCGC T TCTAGA G --- 3'
3' --cgt CTTAAG CGCCGGCG A ACATCT C --- 5'
         EcoRI   NotI       XbaI
```

At the downstream end, all vector inserts have the following restriction sites and sequence:

```
5' --- T ACTAGT A GCGGCCG CTGCAG gct---  3'
3' --- A TGATCA T CGCCGGC GACGTC cga---  5'
          SpeI     NotI    PstI
```

The vector and insert must not contain any other EcoRI, XbaI, SpeI, or PstI cut sites. This excludes the following hexamer sequences:

```
EcoRI:   GAATTC

XbaI:    TCTAGA

SpeI:    ACTAGT

PstI:    CTGCAG
```

The gap between the EcoRI and XbaI sites, and the gap between the SpeI and PstI sites both contain the octameric recognition site for NotI, GCGGCCGC, which may be useful in some applications. We recommend that other occurences of that sequence be avoided in the vectors and insertions, although the presence of such sites will not affect the assembly operations described here. A minimum number of bases (undetermined) between the EcoRI and XbaI sites and the SpeI and PstI sites may be required to allow complete cutting with both enzymes. We have not investigated the exact overhang base requirements, but the eight bases of the NotI recognition sequence is more than adequate.

0.7.2 Verificiation Primers

Standard PCR and sequencing primers have been chosen for pSB103 to use for colony PCR for insert length selection, and for sequencing of inserts. The two primers are Verification Forward (`5' ttg tct cat gag cgg ata ca 3'`) and Verification Reverse (`5' att acc gcc ttt gag tga gc`). These primers give a 270 bp fragment on the null-insert pSB103 vector in colony PCR. Inserts are easily sized by colony PCR. Both primers are designed to be approximately 100 bp from the restriction sites, placing the restriction sites and insert starts at high quality read locations in the sequencing reactions.

0.8 Considerations in the selection of the restriction enzymes

The choice of restriction enzymes was a significant issue in the design of the pSB103 vector and assembly plan. We wanted restriction enzymes which were easy to use and reliable, which functioned in compatible buffer systems and at compatible temperatures, which could be heat killed, provided complete digestion, with few required bases outside of their recognition site, and exhibited low star activity. In addition, we wanted four base overhangs to enhance ligation efficiency.

The sequence of the recognition site was also an issue. Avoiding the accidental creation of ATG start codons at awkward places in combined sequences was one goal. Another challenge was the avoidance of methylation sensitive sequences. Choice of enzymes which ignore DNA methylation was one approach, but other requirements forced the choice of some enzymes which were methylation sensitive. Then, avoiding the accidental creation of DNA methylation sites in common cloning strains, such as DH5α, was a goal. The EcoBI and EcoKI methylases are still active in these, and most other laboratory cloning strains, potentially methylating sites which we must be able to cut. By careful choice of flanking bases, we eliminated the possible creation of EcoBI and EcoKI methylation sites at the critical sequences we required for our assembly technique to reliably function.

0.9 Experimental

0.9.1 pSB103 vector

The pSB103 vector was constructed by PCR from a pUC18 template, using a pair of primers `5' gct tctaga g t actagt a gcggccg ctgcag gcttcctcgctcactgactc 3'` and `5' agtac tctaga a gcgg ccgc gaattc t gcctcgtgatacgcctattt 3'`. These primers introduced the required multiple cloning site, and eliminated much of the non-essential overhead of the pUC18 vector, including all of the Plac/LacZα sequence and the unnecessary surrounding DNA from the original creation of pUC18 from the pBR322 vector. The resulting vector was substantially smaller than the original, although further reductions are both possible and desirable.

The PCR product was Qiagen PCR purified, and cut with XbaI in a 100μl reaction. The product was run on a 1% agarose gel, and the approximately 2000 bp band was cut from the gel, purified, and restriction cut with XbaI. The cut PCR product was ligated with the NEB Quick Ligation kit, and transformed by heat shock into competent DH5α (Invitrogen library efficiency) cells. Picked colonies were grown overnight in 250 ml of LB-Amp and maxipreped using the Biorad Quantum maxiprep kit. ABI BigDye sequencing reactions, run on an ABI 310 sequencer confirmed the correct sequence of one of the two prepared colonies.

0.9.2 lacZα CDS

The lacZα gene was prepared as a standard component for use in testing the effectiveness of the cloning technique. The standard LacZα fragment, part of pUC18, contains the multiple cloning

site, and is unsuitable for use as a standard component. Instead, we used PCR to isolate the lacZα portion of the wild type lac operon from E. coli genomic DNA. Genomic DNA from E. coli MG1655 was prepared and used as a template for the PCR reaction. Primers used were lacZ-F-X, `5' ccgc t tctaga g cag gaa aca gct atg acc atg a 3'` and lacZ-R-PS, `5' gaagc ctg cag cggccgc t actagt a tta aaa gcg cca ttc gcc att 3'`. These primers captured the ribosomal binding site and ATG start from the wild type lacZ gene, prefixing an XbaI cut site upstream. They also added a stop codon, and the appropriate SpeI and PstI cut sites to the downstream end. The lacZα fragment was selected to be the same size as the portion of the wild type lacZ gene found in pUC18. The pUC18 sequence includes a substantial tail of junk amino acid sequence, from the original pBR322 vector, which was eliminated. The PCR product was gel purified, cut with XbaI/PstI, and ligated into an XbaI/PstI cut pSB103 backbone. The result was length verified by colony PCR, clones picked, and sequenced.

0.9.3 Plac promoter

The Plac promoter was PCR'd from pUC18 DNA using primers Plac-F-EX (`5' ggca gaattc gcggccgc t tctaga gca ata cgc aaa ccg cct ct 3'`) and Plac-R-S (`5' ccgct actagt a tgt gtg aaa ttg tta tcc gct ca 3'`). These primers add an EcoRI and XbaI site upstream, and an SpeI site downstream. The PCR product was gel purified, cut with EcoRI and SpeI, and ligated into the pSB103 vector which had been cut with EcoRI and SpeI. Colonies were selected by colony PCR and length identification, grown up, minipreped, and sequenced to identify correct clones.

0.9.4 ECFP CDS

The ECFP coding sequence was PCR'd from the Clontech pECFP vector using forward primer `5' ccgc t tctaga g AAAGAGGAGAAATTAAGC atg gtg agc aag ggc gag gag c 3'` and reverse primer `5' gaagc ctgcag cggccgc t actagt a TTA CTT GTA CAG CTC GTC CAT GC`. The forward primer added a prokaryotic ribosomal binding site (caps) and an XbaI site. The reverse primer added an SpeI and PstI site. The PCR product was cut with XbaI and PstI, gel purified, and ligated into an SpeI and PstI cut pSB103 vector. Colony PCR was used to locate twp out of six correctly sized inserts. Minipreps were performed on the selected colonies, and the results sequenced, isolating one clone with the correct sequence.

0.9.5 EYFP CDS

The EYFP coding sequence was PCR'd from the Clontech pECFP vector using forward primer `5' ccgc t tctaga g AAAGAGGAGAAATTAAGC atg gtg agc aag ggc gag gag c 3'` and reverse primer `5' gaagc ctgcag cggccgc t actagt a TTA CTT GTA CAG CTC GTC CAT GC`. The forward primer added a prokaryotic ribosomal binding site (caps) and an XbaI site. The reverse primer added an SpeI and PstI site. The PCR product was cut with XbaI and PstI, and ligated into an SpeI and PstI cut pSB103 vector. Colony PCR was used to locate three out of twelve correctly sized inserts. Minipreps were performed on the selected colonies, and the results sequenced, producing the expected results.

Subsequent cutting of the resulting vector disclosed that the EYFP sequence has a PstI site, which was removed with the insertion of a cryptic mutation changing the `ctgcag` recognition site to `ctgcaa` while retaining the Gln translation. While this is a less common codon (14% vs. 86%), the modified version performs well. Mutation primers used were `5' cct tcg gct acg gcc tgc aAt gct tcg ccc gct acc` and `ggt agc ggg cga agc aTt gca ggc cgt agc cga agg` . The protocol of the Stratagene Quik-Change mutation kit was followed closely, and functioned well. Transformed colonies were verified by colony PCR, using a mutant-selective primer (`5' tcg ggg tag cgg gcg aag cat 3'`) which failed to match at the 3' end for non-mutated clones. Colony PCR showed all

twelve of the picked colonies to be correct, and subsequent sequencing and restriction digests with PstI verified the mutation.

0.9.6 rrnB-T1 terminator

The transcriptional terminator T1 from the E. coli rrnB ribosomal RNA gene was PCR'd from a previously cloned vector, pSB102-T1. The primers used, T1-F-X (`5' ctct tctaga g cat caa ata aaa cga aag g 3'`) and T1-R-PS (`5' ctct ctgcag cggccgc t actagt a gtc tag ggc ggc gga ttt 5'`) added an XbaI site upstream and an SpeI/PstI site downstream. Cutting the prepared product with XbaI/PstI and ligating to a similarly cut pSB103 vector, transforming, and selecting colonies with colony PCR and subsequent sequencing gave the desired clone.

0.9.7 Plac-lacZ gene

0.9.8 Plac-ECFP gene Plac-EYFP genes

0.9.9 T1-ECFP and T1-EYFP

0.9.10 Plac-EYFP-ECFP and Plac-ECFP-EYFP constructs

0.9.11 Plac-EYFP-T1-ECFP and Plac-ECFP-T1-EYFP

0.10 Acknowledgements

I would like to thank many people who made it possible for me to perform this research by educating me, creating the environment which allowed this research to be done, and sharing the vision of what is possible. Among these are Sonny Maynard, Nick Papadakis, Roger Brent, Drew Endy, Michael Elowitz, Rodney Brooks, Tom McKenna, Erik Winfree, George Church, and Jonathan King. Thanks to NTT and DARPA for their continued financial support. This work is funded by DARPA/ONR contract N00014-01-1-1060 "Computing with Synthetic Biology."

0.11 Component List

0.11.1 Plac promoter

0.11.2 LacZα CDS

0.11.3 ECFP CDS

0.11.4 EYFP CDS

0.11.5 rrnB-T1 terminator

0.12 References

Yanisch-Perron C, Vieira J, and Messing J, Improved M13 phage cloning vectors and host strains: nucleotide sequences of the M13mp18 and pUC19 vectors, Gene 33:103-119 (1985).

Rebatchouk D, Daraselia D, and Narita JO, NOMAD: a versatile strategy for in vitro DNA manipulation applied to promoter analysis and vector design, PNAS 93(20):10891-6 (1996).

Surowiecki, James, The Turn of the Century, Wired, January 2002.