Darwin Assembly: fast, efficient, multi-site bespoke mutagenesis

Christopher Cozens1 and Vitor B. Pinheiro1,2,*

1University College London, Gower Street, London WC1E 6BT, UK and 2Institute of Structural and Molecular Biology, Birkbeck College, University of London, Malet Street WC1E 7HX, UK

Received October 23, 2017; Revised December 21, 2017; Editorial Decision January 20, 2018

ABSTRACT

Engineering proteins for designer functions and biotechnological applications almost invariably requires (or at least benefits from) multiple mutations to non-contiguous residues. Several methods for multiple site-directed mutagenesis exist, but there remains a need for fast and simple methods to efficiently introduce such mutations – particularly for generating large, high quality libraries for directed evolution. Here, we present Darwin Assembly, which can deliver high quality libraries of \(>10^8\) transformants, targeting multiple (>19) distal sites with minimal wild-type contamination (<0.25% of total population) and which takes a single working day from purified plasmid to library transformation. We demonstrate its efficacy with whole gene codon reassignment of chloramphenicol acetyl transferase, mutating 19 codons in a single reaction in KOD DNA polymerase and generating high quality, multiple-site libraries in T7 RNA polymerase and Tgo DNA polymerase. Darwin Assembly uses commercially available enzymes, can be readily automated, and offers a cost-effective route to highly complex and customizable library generation.

INTRODUCTION

Biology, through natural enzymes, has explored and exploited a vast repertoire of chemical reactions, including many that can be harnessed for the manufacture of clinically, biotechnologically and culturally relevant molecules (1). Nevertheless, natural enzymes are optimised to their in vivo setting and are often unsuited for the synthesis of biological or synthetic compounds in vitro or in heterologous host organisms. Consequently, expression and functional optimization or more radical engineering is often required to generate the desired enzymatic activity, whether boosting an existing activity or changing enzyme function altogether. Those needs have led to the flourishing of directed evolution and protein engineering and it has repeatedly proven possible to enhance a number of protein properties, including expression, folding (2), thermostability (3), substrate specificity (4–6) and catalytic efficiency (7,8). While single amino acid mutations can have profound effects, they are rarely sufficient to generate the desired function and often multiple, distal mutations are required in the protein of interest, e.g. T7 RNA polymerase mutants with altered promoter recognition (R96L, K98R, E207K, E222K, N748D, P759L (9)), archaeal DNA polymerase variants capable of efficiently synthesising RNA (Y409G E664K (10)), and aminoacyl transferases used for the expansion of the genetic code (T107C, P254T, C255A (11)).

Sequential cycles of individual site mutagenesis and screening can be effective in navigating from natural to engineered catalyst (4,5,7,12,13) but such an approach requires a sequence landscape where improvement can be detected in each of the intermediates and pre-defines the evolutionary path followed (14,15). Significantly, such iterative approaches cannot uncover phenotypes requiring epistatic mutations as multiple mutations are never made in the same round (16–18). Simultaneous multiple site mutagenesis bypasses this limitation but shifts the bottleneck to the generation of such enzyme libraries. PCR-based methods for introducing mutations to single codons, or a short patch of contiguous codons, in a gene of interest are efficient and well-established, for example iPCR (inverse PCR (19)). In iPCR, a single back-to-back primer pair is used to amplify the whole plasmid containing the gene of interest, and the PCR product is then ligated to regenerate the full-length circular plasmid. Mutations can easily be incorporated into the primers to change the sequence in the PCR product, but this method is limited to the number of mutations that can be incorporated into a single IPCR primer pair.

Simultaneous introduction of multiple mutations at distal sites in a gene is significantly more challenging. A number of strategies have been developed to introduce multiple mutations into plasmid DNA (Table 1) that cover a wide spectrum in specificity (precision over which sites are targeted for mutagenesis), efficacy (fraction of mutated population), efficiency (number of transformants modified at all targeted positions) and complexity (number of experi-
mental steps involved per mutagenesis cycle)—with at least one of these being inevitably compromised.

We set out to develop a new library assembly method that could uncompromisingly address all four quality bottlenecks of simultaneous multiple site saturation mutagenesis. Here, we describe Darwin Assembly—a simple, fast, low-cost and flexible platform capable of delivering large (>10^6 transformants), user-defined libraries with any desired combination of mutations anywhere in a gene of interest, or indeed in multiple genes/features in a plasmid. Darwin Assembly uses commercially available reagents, is automation-compatible and the protocol is not affected by changes in the number or density of mutation sites of a given target. It takes a single working day (from plasmid isolation to library transformation) and is both highly effective (<0.25% wild-type sequences in the generated libraries) and efficient (all clones mutated at all positions targeted).

MATERIALS AND METHODS

Reagents

Enzymes were from New England Biolabs (NEB; Ipswich, MA, USA) unless otherwise stated. MyTaq HS DNA Polymerase was from Bioline (London, UK). KOD Xtreme Hot Start DNA Polymerase was from EMD Millipore (Watford, UK). All oligonucleotides were from Integrated DNA Technologies (IDT; Coralville, IA, USA). ATP and NAD were from NEB. DTT was from Promega (Madison, WI, USA). PEG 8000 was from Fisher Scientific (Loughborough, UK). DNTPs were from Bioline. PCR products were purified using GeneJET PCR Purification Kits (Thermo Scientific, USA). Purification was carried out using Monarch DNA Gel Extraction Kit (NEB).

Unless otherwise stated, all site-directed mutagenesis was carried out by iPCR (19) followed by PCR purification column cleanup and blunt end ligation. All oligonucleotide sequences are provided in Supplementary Table S1 and all plasmids sequences are provided as Supplementary Information.

Plasmids and cells

Plasmid pSB1C3A2 was made from pSB1C3 (http://parts.igem.org/Part:pSB1C3) by inserting a beta-lactamase gene for ampicillin resistance (plasmid pSB1C3A) and removing an unwanted Nb.BsiSI restriction site by site-directed mutagenesis, so that there were only nicking sites on one strand. Plasmid pRST.AS11B.AS3.4 was mutated to remove BsaI sites from exonuclease degradation during the assembly reaction. The first consisted of two oligonucleotides: the 5′-oligonucleotide containing no modifications and the 3′-oligonucleotide having a protected 3′-end (3′inverted-dT) to protect the single stranded 3′ overhang from exonuclease degradation during the assembly reaction. This method was efficient for short assemblies (<1.0 kb) but longer assemblies (2–3 kb) required modifications.

The second variant tested was based on purification of the mutated strand generated during isothermal assembly (Figure 1). The boundary oligonucleotides were designed to anneal to the same strand at each end of the region being assembled and to harbor overhangs. Mutations can be introduced in the annealing sequences (same design as above for ‘inner’ mutagenic primers), but usually were not. Overhangs typically include restriction sites and outnesting PCR priming sites. Three variants of assembly boundary primers were used in the work. The first consisted of two oligonucleotides: the 5′-oligonucleotide containing no modifications and the 3′-oligonucleotide having a protected 3′-end (3′inverted-dT) to protect the single stranded 3′ overhang from exonuclease degradation during the assembly reaction. This method was efficient for short assemblies (<1.0 kb) but longer assemblies (2–3 kb) required modifications.

The second variant tested was based on purification of the mutated strand generated during the assembly reaction by biotin-streptavidin pulldown. Here, two oligonucleotides were used: a 5′-oligonucleotide harboring a biotin-TEG modification at its 5′-end and a 3′-oligonucleotide with a protected 3′-end (Figure 1 and Supplementary Figure S1). We presume that the gains in efficiency observed are
<table>
<thead>
<tr>
<th>Method</th>
<th>Principle</th>
<th>Mutated codons</th>
<th>Efficacy&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Efficiency&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>epPCR (41,42)</td>
<td>PCR introduces errors in region being amplified</td>
<td>Random</td>
<td>n.a.</td>
<td>n.a.</td>
</tr>
<tr>
<td>MMR (43)</td>
<td>Multiple mutagenic primers are used in a single PCR</td>
<td>9</td>
<td>60%</td>
<td>100%</td>
</tr>
<tr>
<td>RECODE (44)</td>
<td>Multiple mutagenic primers are used in a single PCR</td>
<td>4</td>
<td>58%</td>
<td>100%</td>
</tr>
<tr>
<td>Nicking mutagenesis (multi site mutation) (46)</td>
<td>Enzymatic generation of ssDNA template (nicking endonuclease and exonIII) Multiple mutagenic primers are extended and ligated Remaining template removed enzymatically (nicking endonuclease and exonIII)</td>
<td>3</td>
<td>50%</td>
<td>Up to 100%</td>
</tr>
<tr>
<td>QuickChange</td>
<td>Multiple mutagenic primers are extended and ligated directly on a dsDNA template</td>
<td>3</td>
<td>84%</td>
<td>55%</td>
</tr>
<tr>
<td>PFunkel (multi site mutation) (47)</td>
<td>Template removed enzymatically (DpnI) Phagemid synthesis of dU-containing ssDNA template Multiple mutagenic primers are extended and ligated Template removed enzymatically (UDG and exonIII)</td>
<td>Random</td>
<td>n.a.</td>
<td>23–80%&lt;sup&gt;c&lt;/sup&gt;</td>
</tr>
<tr>
<td>MegAnneal (48)</td>
<td>Uses epPCR to create primer library Phagemid synthesis of dU-containing ssDNA template Mutagenized primers are extended and ligated</td>
<td>4</td>
<td>50–100%</td>
<td>50–100%</td>
</tr>
<tr>
<td>OmniChange (49)</td>
<td>Pairs of mutagenic primers are used in short PCRs Compatible overhangs chemically generated (iodine/phosphorothiate) and PCR-amplified fragments ligated</td>
<td>5</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>POEP (50)</td>
<td>Pairs of mutagenic primers are used in short PCRs and PAGE purified</td>
<td>8</td>
<td>100 (digest)</td>
<td>100%</td>
</tr>
<tr>
<td>Golden Gate (51)</td>
<td>PCR products are used in an overlap extension PCR</td>
<td>13</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td>MF OE-PCR (50)</td>
<td>Pairs of mutagenic primers are used in short PCRs and purified PCR products are used in an overlap extension PCR epPCR or pairs of mutagenic primers are used in short PCRs and purified</td>
<td>Random</td>
<td>n.a.</td>
<td>n.a.</td>
</tr>
<tr>
<td>TAMS (52)</td>
<td>Linear PCR to generate ssDNA template Multiple mutagenic primers are extended and ligated Outnest used for PCR amplification of assembled fragment</td>
<td>4</td>
<td>NR</td>
<td>100%</td>
</tr>
<tr>
<td>OD SPM (53)</td>
<td>Multiple mutagenic primers are extended and ligated directly on a denatured dsDNA template Outnest used for PCR amplification of assembled fragment</td>
<td>11</td>
<td>NR</td>
<td>Up to 100%</td>
</tr>
<tr>
<td>Darwin Assembly (this study)</td>
<td>Enzymatic generation of ssDNA template (nicking endonuclease and exonIII) Multiple mutagenic primers are extended and ligated Outnest used for PCR amplification of assembled fragment</td>
<td>19</td>
<td>98–100%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Error-prone PCR (41) is simple but mutations introduced cannot be targeted to particular codons and its effectiveness is dependent on amplification biases and random incorporation of errors (42). In these circumstances efficacy and efficiency are not applicable concepts (n.a.). Some methods are not effective at mutating the whole population (43,44) and/or incorporating all mutations (45,46). Methods that can introduce mutations effectively and efficiently tend to require complex and time-consuming steps, such as phagemid propagation to generate dU-containing DNA (47,48), chemical degradation steps (49), multiple PCR reactions followed by overlap extension PCR reactions to assemble the mutated gene (44,50,51) or modularization of the library into PCR-tractable libraries that can be later re-assembled via Golden Gate assembly (6). Other methods, such as TAMS (52) and OD SPM (53), are simple and effective. Both introduce diversity by using primers containing the target mutations for primer extension and ligation against single-stranded templates followed by PCR and cloning (52,53). However, neither of these methods remove the second strand of the original plasmid and both use T4 DNA ligase, which can ligate across gaps, mispairs and cannot tolerate high temperatures favourable for specific annealing (28)—with resulting compromises in efficiency of library assembly.

<sup>a</sup> Efficacy is defined as the fraction of the population containing mutations.

<sup>b</sup> Efficiency is defined as the fraction of mutated clones where all targeted sites are mutated. Some methods do not report (NR) their efficacy or efficiency.

due to improved purification of the assembled DNA from mutagenic primers and single-stranded plasmid template.

The third variant tested used a single, long oligonucleotide harboring both 5' and 3' binding sites, akin to a padlock probe (23), with sites for post-assembly amplifications (i.e. restriction sites and outnest PCR priming sites) linked by a short flexible linker (dT5). This topology, referred to as the θ oligonucleotide, generates a closed circle after isothermal assembly, allowing exonuclease cleanup and hence removal of partial assembly products, unincorporated oligonucleotides and original plasmid (Figure 2 and Supplementary Figure S1).

Outnest PCR primers, designed to anneal onto the assembly boundary primer overlaps, were typical PCR primers used to amplify the synthesised libraries prior to cloning by Golden Gate assembly (24). Annealing temperatures were calculated using default settings in SnapGene 3.0.3 (GSL Biotech LLC), apart from oligonucleotides containing deletions which were calculated using MELTING 5.1 (http://www.ebi.ac.uk/biomodels/tools/...
**Figure 1.** Principles of Darwin Assembly. Plasmid DNA (black, with the gene of interest in orange) is nicked by a nicking endonuclease (at the purple dot) and the cut strand degraded by exonuclease III (1). Boundary and inner (mutagenic) oligonucleotides are annealed to the ssDNA plasmid (2). Key features of the oligonucleotides are highlighted: 5′-boundary oligonucleotide is 5′-biotinylated; non-complementary overhangs are shown in blue with Type IIS endonuclease recognition sites shown in white; mutations are shown as red X in the inner oligonucleotides; 3′-boundary oligonucleotide is protected at its 3′-end. After annealing, primers are extended and ligated in an isothermal assembly reaction (3). The assembled strand can be isolated by paramagnetic streptavidin-coated beads (4) and purified by alkali washing prior to PCR using outnested priming sites (5) and cloning (6) using the type IIS restriction sites (white dots). The purification step (4) is not always necessary but we found it improved PCR performance, especially for long assembly reactions (>1 kb).

---

**Single strand plasmid generation**

Plasmid DNA, isolated from overnight bacterial cultures using GeneJET Plasmid Miniprep Kits (Thermo Fisher Scientific), was made single stranded by co-incubation with a nicking endonuclease and exonuclease III (exoIII). Typically, DNA was digested at 50–60 nM (90–235 ng/μl, depending on plasmid size) with 3–10 U of nicking enzyme and 40–60 U of exoIII per μg of DNA. Plasmid DNA was vacuum concentrated (SpeedVac, ThermoFisher Scientific) if necessary prior to digestion. Reactions, in appropriate restriction enzyme buffer (typically 1 x CutSmart buffer or 1x NEBuffer 3.1), were carried out for 2 h at 37°C, followed by 20 min at 80°C to inactivate the enzymes. Nt.BspQI and Nb.BtsI were used successfully but Nt.BbvCI was consistently less active under the conditions used. No other nicking endonucleases have been tested. Digestion was confirmed by agarose gel electrophoresis followed by SYBR Gold (Thermo Fisher Scientific) staining. ssDNA was used for assembly without further purification and was not re-quantified.

**Oligonucleotide phosphorylation**

Mutagenic primers and the 3′ boundary assembly oligonucleotides were phosphorylated at their 5′-ends using T4 polynucleotide kinase (T4 PNK, NEB), to enable ligation during isothermal assembly. Oligonucleotides, at 4–20 μM final oligonucleotide concentration, were phosphorylated in 1x CutSmart buffer supplemented with 1 mM ATP and 0.1 U/μl T4 PNK. Typically, the inner mutagenic oligonucleotides were mixed and phosphorylated in a single reaction and the 3′ boundary oligonucleotide phosphorylated separately. Reactions were carried out for 2 h at 37°C, followed by 20 min at 80°C, to inactivate the enzymes. Phosphorylated oligonucleotides were added to subsequent assembly without further purification.
Isothermal assembly

Firstly, single-stranded plasmid DNA (typically 0.1–0.2 pmol per reaction) was mixed with mutagenic oligonucleotides (at 100–250-fold molar excess) and boundary assembly oligonucleotides (at 10–50-fold molar excess over plasmid DNA). Plasmid and oligonucleotide mixtures, typically between 3 and 5 μl, were then annealed by freezing samples at −20°C for at least 15 min and subsequently returning to room temperature. Alternatively, efficient annealing was also obtained by heating samples for 5 min to 95°C and cooling them at 0.1°C/s to 4°C.

After annealing, 1 volume 2× Darwin Assembly enzyme mix (2× DA mix) was added and the reaction incubated at 50°C for 10 min to 1 h, depending on the length of the assembly (60°C was also successfully tested, data not shown). Ten minutes is sufficient even for long assemblies (e.g. T7 RNA polymerase; 2.7 kb) although slightly higher PCR yields are seen after longer reaction times. 2× DA mix consists of: 0.05 U/μl Q5 High-Fidelity DNA polymerase (not hot start), 8 U/μl Taq DNA ligase, 2 mM NAD+, 0.4 mM each dNTP, 10% (w/v) PEG 8000, 2 mM DTT and 1× CutSmart buffer (recipe available in SI). This reaction mix is closely modeled on the Gibson Assembly (25), with Phusion replaced by Q5 DNA polymerase and the reaction buffer replaced with CutSmart buffer. It is stored frozen (and like commercially available Gibson Assembly mixes, will freeze solid).

Assembly purification with streptavidin-coated paramagnetic beads

Streptavidin-coated paramagnetic beads were used to clean up assemblies carried out using 5′-biotinylated 5′-boundary oligonucleotides.

Initially, 5 μl Dynabeads MyOne Streptavidin C1 beads (Thermo Fisher Scientific) were blocked in 2× BWBS-T (20 mM Tris–HCl pH 7.4, 2 M NaCl, 0.2% (v/v) Tween-20, 2 mM EDTA) for over an hour on a spinning wheel rotator at room temperature. The beads were isolated on a magnetic stand and resuspended in 50 μl 2× BWBS-T.

Once the assembly reaction was completed, its volume was adjusted to 50 μl with PCR-grade water and transferred to a 1.5 ml tube. The pre-blocked beads were added to the reaction and mixed thoroughly. Assembly binding to beads was carried out for 3 h at room temperature, although it was later found that 10 min were sufficient (albeit at slight cost to yield). After capture, beads were isolated on a magnetic stand and washed twice in 200 μl 37°C 30 mM NaOH and once in 200 μl EB-T (10 mM Tris–HCl pH 8.8, 0.1 mM EDTA, 0.01% Tween-20). Alkali denaturation is carried out at 37°C to improve bead handling – this is not essential for plasmid denaturation. The wash with EB-T is included to neutralise the beads and to reintroduce Tween-20, to prevent the beads adhering to tube walls or aggregating. After washing, the beads were resuspended in 10 μl EB (10 mM Tris–HCl pH 8.8) and used directly for PCR. A more detailed version of this approach is provided in the accompanying Supplementary Information (Supplementary Protocol 1).

Assembly purification by exonuclease

Exonuclease cleanup was used in conjunction with the θ oligonucleotide method for one-pot purification of assembly products. After assembly, a master mix consisting of exonuclease I (exoI), exoIII, a targeting oligonucleotide (see below) and an appropriate restriction endonuclease is added to the assembly reaction. The targeting oligonucleotide is designed to anneal efficiently to a suitable en-
donuclease restriction site present in the ssDNA template (e.g. SphI in pET29), so that the chosen endonuclease can cleave the local dsDNA template generated. This creates free 3'- and 5'-ends enabling exoIII and exonuclease to degrade the starting template, along with any partial assembly products and free oligonucleotides. A more detailed version of this approach is provided in the accompanying Supplementary Information (Supplementary Protocol 2).

For the TgoT experiment, a 5x master mix consisting 4 U/μl exoI, 2 μl exoIII, 1 μl SphI-HF and 50 μM oligonucleotide pET_SphICut was prepared, of which 4 μl were added to a 16 μl isothermal assembly reaction, giving final concentrations of 0.8 U/μl exoI, 0.4 U/μl exoIII, 0.2 U/μl SphI-HF and 10 μM pET_SphICut. Reactions were carried out for 40 min at 37°C, followed by incubation for 20 min at 80°C to inactivate endo- and exonucleases.

PCR amplification of assembled DNA

PCR amplification was carried out directly from the assembly reactions, or after its purification (whether exonuclease or beads). When carried out directly from assembly reactions, typically 1 μl assembly was used as template per 25 μl PCR. Primers outnest_1 and outnest_2, targeting sites on the boundary assembly oligonucleotides, were used to specifically amplify the assembled libraries (as their sequences were chosen to not be present in the original plasmid DNA template).

Multiple enzymes [MyTaq (Bioline), Q5 Hot Start DNA polymerase (NEB) and KOD Xtreme (EMD Millipore)] were successfully used for post-assembly PCR. In general, KOD Xtreme was found to be the most robust regarding successful PCR amplification. Reactions consisted 1× KOD Xtreme buffer, 0.3 μM outnest_1, 0.3 μM outnest_2, 0.4 mM each dNTP and 0.01 U/μl KOD Xtreme DNA polymerase. The cycling parameters used were 2 min 95°C, followed by 20–28 cycles of 15 s at 98°C, 15 s at 64°C and 2 min 30 s at 68°C, with a final polishing step of 2 min at 72°C. These conditions were typical for 2–3 kb PCR products, and extension time was adjusted as per manufacturer recommendations for shorter assemblies.

After 25 μl pilot PCRs, PCRs were scaled up (usually to 2 × 50 μl reactions) and gel purified using Monarch PCR and DNA cleanup kits (NEB). These were preferred as they allow low elution volumes and result in clean DNA (low A260 contamination). The purified products were digested with BsaI-HF to remove outnesting priming sites and generate suitable overhangs for cloning, and again purified to remove enzyme, salts and cleaved primers. DNA concentration was determined by spectrophotometry (SPECTROstar Nano, BMG Labtech, UK).

Illumina sequencing and data analysis

All deep sequencing was carried out on an Illumina MiSeq at the UCL Genomics Centre using a 150 cycle MiSeq Reagent Kit v3 (Illumina). Libraries were prepared by PCR using KOD Xtreme Hot Start DNA Polymerase (EMD Millipore) to append barcodes and sequencing primers. Libraries were gel purified on agarose gels stained with SYBR safe (Life Technologies), visualised using a blue light source and purified using Monarch DNA Gel Extraction kits (NEB). Libraries were then quantified using a Qubit Fluorometer (Thermo Fisher Scientific).

Codon frequencies were counted in R version 3.3.0 (27) using RStudio version 1.0.136 for Mac OS X after data processing. Raw FASTQ files were obtained from the MiSeq and quality-filtered, trimmed, barcode-split and converted to FASTA using the FASTX-Toolkit (http://hannonlab.cshl.edu/fastx). If necessary, Prinseq (http://prinseq.sourceforge.net/) or TAGCleaner (http://tagcleaner.sourceforge.net/) were used to omit reads below a minimum or above a maximum length. FASTA headers were removed, reading frames aligned and the DNA sequence split into triplet codons to facilitate reading into R. More detail on the commands used is available in the Supplementary Methods (Supplementary Protocol 3).

RESULTS

Principles of Darwin Assembly—fast, efficient, multi-site base-pair mutagenesis

The Darwin Assembly process can be split into three steps: the generation of single-stranded DNA template, isothermal assembly, and subsequent amplification for cloning (Figures 1 and 2). Each step was designed such that the output of one step becomes the input of the subsequent step without the need for purification – minimising sample loss and handling time. The result is a fast protocol that goes
from plasmid DNA to library within a day and it is compatible with automation.

Single-stranded template is generated from purified plasmid DNA by the coupled reaction of a nicking endonuclease and exonuclease III (exoIII). The endonuclease specifically cuts one strand of plasmid DNA and exoIII selectively and efficiently degrades the nicked DNA strand, leaving behind a circular single-stranded template for assembly. Nevertheless, we found it increased the robustness of the assemblies to position a nicking site shortly after the 3’-end of the assembly reaction in some instances (data not shown). Nicking endonuclease and exoIII are heat inactivated and the single-stranded template can be used directly in assembly, as all enzymes used have activity in a common buffer.

For assembly, primers (except for the 5’ boundary assembly oligonucleotide) are phosphorylated enzymatically and added to the assembly reaction in the desired molar ratios. Mutations are introduced by the oligonucleotides and primer excess ensures efficient binding to ssDNA template, contributing to the high efficiency of Darwin Assembly. Boundary assembly primers are used at lower concentrations than mutagenic primers (but still in excess of ssDNA template) to limit side reactions that could undermine a successful assembly. Primer concentration was not extensively optimised as the concentrations used proved robust for assembly and generated minimal background or unwanted assemblies.

Once oligonucleotides and ssDNA template are annealed, assembly is carried out with a non-strand displacing thermostable DNA polymerase (Q5) and a thermostable DNA ligase (Taq DNA ligase), akin to Gibson Assembly (25). The high reaction temperature and the choice of DNA ligase both contribute to the specificity of the assembly, limiting off-target annealing of the primers and limiting ligation to nicks on the nascent DNA strand (28), respectively.

Depending on the complexity of the assembly, no purification may be necessary, with assembly reactions used directly as PCR templates. Purification, whether based on nucleases or isolation of biotinylated DNA post-assembly, improves the quality of subsequent PCR amplification making it less likely to contain secondary amplification products (data not shown). We assume the main source of secondary PCR products is carryover of inner assembly oligonucleotides to the PCR, where they act as primers. Reducing the concentration used in the assembly reaction may remove the need for purification but at the possible risk of missing out some mutations.

PCR amplification post-assembly generates sufficient material for Golden Gate cloning (29), ensuring highly efficient and scarless cloning of the assembled fragment. Our deep sequencing results show that 98–100% of targeted codons are mutated as expected. No wild-type clones were detected by Sanger sequencing, even in the absence of any purification steps, and deep sequencing of biotin-purified assembled libraries showed wild-type contamination was always <0.25%, a level comparable with sequencing error rates.

In developing the methodology and to explore its potential applications, we have applied Darwin Assembly to different genes, including chloramphenicol acetyl transferase (CAT), *Saccharomyces cerevisiae* tryptophanyl tRNA-synthetase (ScWRS), *Thermococcus gorgonarius* DNA polymerase (TgoT), *Thermococcus kodakaraensis* DNA polymerase (KOD) and T7 RNA polymerase (T7RSS), which differ in length as well as composition. We demonstrate Darwin Assembly can be used to wholesale reassign codons in a gene of interest, create amino acid scanning and insertion/deletion (indel) libraries, as well as complex mutational libraries – where diversity can be highly customisable (e.g. non-redundant equimolar incorporation of triplet codons).

**Whole gene codon reassignment**

We chose chloramphenicol acetyl transferase (CAT) as a suitable initial target to develop and optimise Darwin Assembly, as it is a small (660 bp), well-characterized gene. Our initial proof of principle was to reassign all leucine codons in CAT to CTG. Six oligonucleotides were used to introduce the required 10 point mutations (in seven different codons spread over 512 nucleotides, Figure 3A). Assemblies were carried out using two boundary oligonucleotides and without any pre-PCR purification. Two assembly conditions were tested, with either 125× inner and 25× boundary oligonucleotide molar excess over plasmid, or 250× inner and 50× boundary oligonucleotide molar excess over plasmid.

Both assemblies were successful (Figure 3B) and 100% of the sequenced transformants (seven colonies in total: three transformants for the lower and four of the higher concentrations) had all of the desired mutations with no mistargeting observed. In a more rigorous test of the technology, we successfully introduced 38 point mutations (in 19 different codons spanning 2378 nucleotides) in KOD DNA polymerase in a single Darwin Assembly reaction (Supplementary Figure S2), using the refinements described in Materi-
als and Methods (shown in Figure 1) to remove the initial ssDNA template and unincorporated oligonucleotides.

Encouraged by the high efficiency of the initial assembly reactions, we decided to investigate if the method was sufficiently robust and efficient for the generation of libraries of increasing complexity: site saturation (all possible mutations at a single position), partial saturation (mutually exclusive mutations targeting different sites), and indel libraries (deleting or inserting whole codons).

We tested the introduction of diversity by making a small library in Saccharomyces cerevisiae tryptophanyl tRNA synthetase, targeting positions important in substrate specificity (Thr107, Pro254 and Cys255 (11)) with limited degeneracy (KKC, KKC and RYA, respectively). Assembly was successful and a small number of transformants were sequenced. All three sites were successfully targeted in the five sequenced transformants and no off-site mutations were observed (data not shown).

### Table 2. Deep sequencing of CAT gene alanine scanning library

<table>
<thead>
<tr>
<th>Residue</th>
<th>Wild-type codon</th>
<th>Frequency</th>
<th>Mutated codon</th>
<th>Frequency</th>
<th>Other codons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glu 100</td>
<td>GAA</td>
<td>79 205 (99.87%)</td>
<td>n.a.</td>
<td>n.a.</td>
<td>103 (0.13%)</td>
</tr>
<tr>
<td>Thr 101</td>
<td>ACG</td>
<td>68 947 (86.94%)</td>
<td>GCT</td>
<td>9925 (12.51%)</td>
<td>436 (0.51%)</td>
</tr>
<tr>
<td>Phe 102</td>
<td>TTT</td>
<td>59 947 (75.59%)</td>
<td>GCA</td>
<td>18 663 (12.51%)</td>
<td>648 (0.82%)</td>
</tr>
<tr>
<td>Ser 103</td>
<td>TCA</td>
<td>58 395 (73.63%)</td>
<td>GCA</td>
<td>20 592 (25.96%)</td>
<td>309 (0.39%)</td>
</tr>
<tr>
<td>Ser 104</td>
<td>GCA</td>
<td>66 541 (83.90%)</td>
<td>GCA</td>
<td>12 433 (15.68%)</td>
<td>306 (0.39%)</td>
</tr>
<tr>
<td>Leu 105</td>
<td>CTC</td>
<td>135 (0.17%)</td>
<td>CTG</td>
<td>79 027 (99.68%)</td>
<td>146 (0.18%)</td>
</tr>
<tr>
<td>Trp 106</td>
<td>TGG</td>
<td>79 047 (99.67%)</td>
<td>n.a.</td>
<td>n.a.</td>
<td>260 (0.33%)</td>
</tr>
</tbody>
</table>

Near complete mutation of Leu105 confirms the efficiency of the approach with alanine represented in all target sites with frequencies between 12.51% and 25.96% (expected frequency: 20%). Sites not targeted by mutagenesis (Glu100 and Trp106)―and where mutations are not applicable (n.a.)―show the underlying error rate of sequencing (0.13–0.33%) is comparable with the error obtained in the engineered positions (0.18–0.82%).

The control Leu105 (CTC→CTG) mutation was present in 99.7% of the samples, a frequency comparable to sites that had not been targeted for mutagenesis (see Table 2).

Crucially, an expected proportion of reads (16 739 or 21.1%) were wild-type (containing only the Leu105 recoding). This demonstrates that even though there may be a degree of incorporation bias, with alanine incorporation ranging from Thr101 (12.5%) to Ser103 (26.0%), these are unlikely to have been caused by the mismatches alone.

### Generation of deletion and insertion (indel) libraries

Although insertions and deletions are known to play crucial roles in protein function (e.g. antibodies), there are few tools available to explore length as a parameter in protein engineering (BioRxiv: https://doi.org/10.1101/127829) (33). We therefore decided to investigate if Darwin Assembly could be used as a tool to explore indels for protein engineering.

We generated two libraries in CAT: one exploring single and double deletions and one exploring single and double insertions (Supplementary Figure S3). As with the scanning libraries, the indel libraries were assembled using an equimolar mixture of inner oligonucleotides designed against a common binding site. In addition, similar controls were included, using a wild-type oligonucleotide and the Leu105 (CTC→CTG) controls. The deletion library was assembled with four inner oligonucleotides, maintaining the wild-type sequence, deleting Phe102 or Ser104, or deleting both sites. The insertion library was generated with three oligonucleotides, maintaining the wild-type sequence, or introducing one or two degenerate codons (NNS) between Phe102 and Ser103.

Assemblies of both libraries were successful and several thousand transformants were obtained for each library. Libraries were pooled and sequenced obtaining approximately 2.1 × 10⁵ reads for analysis. As both libraries contained a wild-type assembly control and were pooled for sequencing, analysis of assembly biases cannot easily be determined.

Nevertheless, there is a clear drop in assembly efficiency between single (16.5% of reads ΔPhe102, 5% ΔSer104) and double deletions (0.39% for ΔPhe102ΔSer104), which we expect reflects the drop in melting temperature of the inner oligonucleotides used in assembly. For the insertion library, it is less clear whether a bias is observed. Importantly, all possible codons encoded by the degenerate inner oligonucleotides were observed for both single and double insertion populations (Supplementary Tables S2–4).
Figure 4. Darwin assembled T7 RNA polymerase library. Sequencing of three separate regions of the assembled gene were used to determine the assembled diversity in the targeted codons (A). Number of reads obtained for each segment are shown as well as the range of each reaction (shown in blue). Frequency of incorporation of the 32 possible codons (NNS) is shown as a box plot, with overrepresented outliers explicitly labelled. Wild-type contamination was determined where possible (N.D.: not determined). Incorporation trends were detected in all positions and best correlated to the number of mismatches they introduced. A typical correlation is shown (B) but other positions varied (0.03–0.37), always indicating that the higher the number of mutations being concomitantly introduced (i.e. larger Hamming distance), the lower incorporation of the variant in the population. Incorporation trends were also analysed to determine systematic biases of incorporation. Incorporation frequencies were ranked at each position and rank order analysed (C)—the top three highest ranked (greens) and lowest ranked (oranges) codons are highlighted for clarity. Incorporation biases were detected but it was not clear whether that was the result of the very imbalanced sequence composition of the wild-type residues targeted. Outnest PCR of the T7 RNA polymerase library (expected product of 2840 bp) confirms that as little as 10 minutes of isothermal assembly were sufficient for successful amplification of a library (D). MW: 1 kb ladder (NEB). NT: no template PCR control.

Together, the insertion and deletion libraries confirm that Darwin Assembly can be used to investigate small changes in length in a target gene, but may require further optimisation to minimise assembly biases of deletion libraries. Deletions and insertions decrease the melting temperature of the mutagenic primers, with the ΔPhe102ΔSer104 oligonucleotide having a predicted melting temperature of nearly 15°C lower than the Leu105CTG controls (Supplementary Table S5). As assembly was carried out with both oligos present in a single reaction, kinetic and thermodynamic factors can account for the observed frequency variation. We expect that extending mutagenic primers that introduce insertions and deletions to normalise melting temperatures in the assembly reaction would result in a more even distribution across the different lengths.

Darwin Assembly enables the synthesis of custom complex libraries ideal for directed evolution

Having demonstrated the potential of Darwin Assembly on the CAT gene, we implemented assembly on longer targets: T7 RNA polymerase (2.7 kb) and TgoT DNA polymerase
(2.4 kb). These are more typical of enzymes targeted for engineering and represented a greater challenge.

As expected, assemblies longer than 1.2 kb were initially less robust and invariably required considerable PCR optimisation (data not shown). We concluded the PCR problems were caused by carryover of unincorporated inner oligonucleotides and/or partial assembly products from the assembly reaction to PCR. Therefore, we investigated two strategies to remove template and unextended (or partially assembled) oligonucleotides: either purifying assembled molecules (Figure 1) or by selectively degrading template and unassembled products (Figure 2). Both approaches were successful and enabled assembly of libraries spanning more than 2 kb with T7RSS and TgoT respectively.

For T7RSS, we designed a library targeting residues previously implicated in switching promoter specificity (9): Arg96, Lys98, Glu207, Glu222, Asn748, Pro759. The 2724 bp assembly, targeting six codons, used five inner oligonucleotides introducing degeneracy (NNS) at each of the target positions (Figure 4A). Assembled libraries were amplified, cloned and transformed, with approximately 1 × 10⁷ transformants isolated.

Sequencing of the pooled transformants confirmed that at each position, all possible variants were introduced (Figure 4, Supplementary Table S6). Wild-type contamination, in codons where that could be detected, was always <0.24% and comparable to sequencing errors detected in non-targeted positions. Incorporation biases were small and correlated primarily with the number of mismatches between template and target codon. Overall, there were also detectable biases over the whole assembly (Kruskal–Wallis ANOVA on ranked target codons, P = 0.001), but since the template residues were clustered in sequence space (5/6 were 67% purine or above), it is possible that the overall incorporation bias observed is a consequence of the individual mismatch biases, rather than a systematic bias.

A known limitation of targeting residues with NNS (or NNK) degenerate oligonucleotides is the bias they introduce in translation: oversampling residues that have multiple codons (e.g. Leu and Arg) and increasing the minimum library size required to adequately sample the sequence space introduced. Alternatives to NNS have been developed for site-directed mutagenesis, such as ‘small intelligent’ libraries that mix four oligonucleotides with different degrees of degeneracy (NDT, VMA, ATG and TGG) and in a suitable ratio (12:6:1:1, respectively) to generate equal coverage of all 20 amino acids (34). Such libraries are impossible when targeting multiple sites using simple iPCR. Nevertheless, as Darwin Assembly targets distal sites using independent inner oligonucleotides, it is well suited to generating multi-site intelligent libraries – where nucleic acid variation matches protein variation, maximising library quality.

We therefore designed the TgoT library to target eight residues implicated in template recognition and DNA duplex affinity modulation (Tyr384, Val389, Ala490, Tyr493, Phe587, Glu664, Gly711 and Asn735 (10,35)), with a mixture of limited or with small intelligent degeneracies (Figure 5).

The assembly spanned 2389 bp and targeted the eight residues with six inner oligonucleotides and a non-mutagenic ‘0’ boundary. The resulting library was successfully amplified, cloned and transformed with an estimated 2.25 × 10⁸ transformants generated. Pooled transformants were sequenced, confirming that the designed diversity was introduced in all targeted residues (Figure 5). As with the T7 library, a positional bias is detectable and correlates weakly to the number of mismatches between template and target codons. Overall, there is a detectable bias in the final assembly beyond a potential systematic error in the mixing of the oligonucleotides (Kruskal–Wallis ANOVA on ranked target NDT codons, P = 0.005), but it is less clear whether the bias is a consequence of the individual template/target codon mismatches—e.g. systematic underrepresentation of GGT even at residue 711 (GCC). (Figure 5B, Supplementary Tables S6–S8).

Overall, both T7RSS and TgoT libraries do effectively sample the introduced diversity, given transformants isolated and the measured per site diversity. Both libraries have the expected diversity at multiple distal sites, have minimal contamination of wild-type sequences and are free of assembly errors, despite the size of the genes. Moreover, Darwin Assembly enabled generation of those libraries within a single working day and at a lower cost than current potential commercial suppliers.

**DISCUSSION**

Library quality is a major contributing factor to directed evolution (36). It impacts the minimum library size that has to be generated to adequately sample a given sequence landscape (16,37) and it impacts the choice of selection strategy (depending on the relative rarity of a ‘successful’ variant and gain of function over wild-type). Generating high-quality libraries that target multiple distal sites in a gene is challenging, with commercial library synthesis and modular assembly of the library being available only recently.

Darwin Assembly is a highly efficient method for simultaneous mutagenesis of multiple sites. It is capable of introducing point mutations, bespoke diversity, insertions and/or deletions in complete or partial replacement of wild-type. It is fast, requires no specialised strains or chemical steps and libraries are easily generated in a single working day (from plasmid miniprep to transformation) using commercially available enzymes. We also observed very low levels of wild-type sequence contamination (peaking in TgoT E664 at 1.95% but generally <0.5%). Overall, this makes Darwin Assembly uniquely suited for generating high quality libraries for directed evolution experiments (36,38).

Darwin Assembly is amenable to automation, as most steps are enzymatic in compatible buffers and can readily be programmed. We cloned our libraries using type IIS (Golden Gate) restriction cloning, but the freedom to design outstems means any cloning strategy can be employed. Mutations are introduced by oligonucleotides and their design has few constraints – a binding site of at least 11 bases at each end of the ‘inner’ oligonucleotides. As a result, mutation type and density is limited primarily by the oligonucleotide synthetic constraints, which could, for instance, include unnatural bases (39,40). Similarly, while we only target coding regions of single genes, simultaneous mutation of multiple features (e.g. RBS and gene) or multiple genes
Figure 5. Darwin assembled TgoT DNA polymerase library. (A) Five separate sequencing reactions (range and reads shown in blue) were required to sample the diversity introduced across the eight target residues (shown in red along the TgoT gene). Mutations included focused degeneracies (e.g. YWC used against Y384) or ‘small intelligent’ (S-int) diversity (NDT, VMA, ATG and TGG oligonucleotides mixed in a 12:6:1:1 ratio). Resulting incorporation is shown in box plots with outliers explicitly labelled. Wild-type contamination was determined from positions where diversity excluded those sequences (N.A.: not applicable). As with the T7 RNA polymerase library, incorporation trends and biases were analysed to identify any biases in assembly. Ranked incorporation frequencies are shown for the residues targeted with ‘small intelligent’ diversity, and the top three highest (greens) and lowest (oranges) ranked codons (based on a straight sum of ranks) are highlighted (B). Outnest PCR of the TgoT DNA polymerase library (expected product of 2501 bp) showing that the final PCR can be carried out with either A- (MyTaq) or B-family (Q5) polymerases (C). MW: 1 kb ladder (NEB). NT: no template PCR control.

(e.g. tRNA and aminoacyl tRNA synthetase) on the same plasmid could be targeted with no protocol modifications, providing the boundary oligonucleotides are sited appropriately.

Post assembly cleanup was not necessary for short fragments, making a one-pot assembly feasible. Longer assemblies greatly benefited from the removal of wild-type template and unincorporated oligonucleotides and both methods we present here were successful. Post-assembly amplification was marginally better for T7RSS, suggesting that purification of assembled fragments may be superior to degradation of template and unused oligonucleotides. We demonstrate that assemblies spanning as much as 2.7 kb are possible and we postulate that maximum assembly length would be determined by the success of post-assembly amplification. In principle, using boundary oligonucleotides, it may be possible to replace post-assembly PCR with rolling circle amplification, which would enable assemblies spanning regions longer than 30 kb. We have successfully introduced
38 point mutations covering 19 codons in a single reaction and see no reason why this is the limit of the technology.

SUPPLEMENTARY DATA
Supplementary Data are available at NAR Online.

ACKNOWLEDGMENTS

The authors thank Jared Ellefson and Andy Ellington (University of Texas) for the kind gift of pRST.11B.AS3.4 (encoding ScWRS) and the T7RSS gene. We also thank John Chaput (University of California, Irvine) for the kind gift of the pGDR11-KOD plasmid harbouring the KOD DNA polymerase gene.

FUNDING

Biotechnology and Biosciences Research Council [BB/N01023X/1, BB/N010221/1]; European Research Council [ERC-2013-StG project 336936 (HNAepisome)]. Funding for open access charge: BBSRC [BB/N01023X/1]. Conflict of interest statement. None declared.

REFERENCES