Arturo Casini

Advanced DNA assembly strategies and standards for synthetic biology

Submitted for the Degree of Doctor of Philosophy

Supervised by Dr. Geoff Baldwin and Dr. Tom Ellis

Imperial College London
Department of Life Sciences
Centre for Synthetic Biology and Innovation

September 23, 2014

Abstract

DNA assembly is a fundamental enabling technology for synthetic biology, yet it is also extremely unreliable, expensive and time-consuming. The process usually requires a significant part of total time and effort that can be dedicated to a project, reducing the resources available for the rest of the research, and is also frequently subject to unexpected problems, introducing an undesirable element of unpredictability that might compromise an entire project. This thesis describes the development of three DNA assembly tools that aim to facilitate and speed up synthetic biology research: “MODAL” is a fast and easy to use assembly strategy that brings the advantages of standardisation and modularity to the latest-generation long overlap-based DNA assembly techniques. “Linker” is a software tool that generates DNA sequences specifically optimised to act as high-efficiency homology regions in long overlap-based DNA assembly reactions. Finally with “BASIC” we propose a new DNA assembly standard that incorporates the advances of MODAL and Linker and brings an additional series of improvements in an original assembly workflow. BASIC aims first of all to make DNA assembly significantly more reliable by addressing and/or removing all the unpredictability elements. It also maintains the speed, ease of use and flexibility of MODAL while achieving the same or better efficiency than the best currently available DNA assembly techniques and standards.

Table of contents

Index of Figures

Index of Tables

Declaration

I herewith certify that all the material in this thesis is my own work, except for quotations from published and unpublished sources which are clearly indicated and acknowledged as such. The source of any picture, diagram or other figure that is not my own work is also indicated.

Acknowledgments

I would like to start by thanking my supervisors, Geoff Baldwin and Tom Ellis, for the guidance and the support during all these years. I know I have been wrong many times and I have not always admitted it or even realised it sometimes. I sincerely thank you for your help and your patience: you have been invaluable for both my professional and my personal development. I would also like to thank everyone in the Baldwin and Ellis groups and at CSynBI in general. It's been great working with you, again both professionally and personally (and I apologise for the very loud and very weird music). I would especially like to thank Chris Hirst who essentially taught me from scratch how to work in a lab during my MRes year, giving me the confidence to embark on a PhD project.

Thanks my friends, especially Bre, Marco, Christian, Francesco, Simone, Jon and everyone else who kept me sane during these years in London. I'm actually not so sure you succeeded, but you've been amazing nonetheless. Thanks to my family, who supported me in every possible way for all my life and believed that I was actually accomplishing something despite the myriad of distractions I continuously gave in to.

And finally thanks to Caterina.

Glossary

CPEC: Circular Polymerase Extension Cloning, a DNA assembly method[1].

GC%: also called GC content, is the percentage of G and C bases over the total number of bases in a given DNA fragment.

Gibson assembly: a DNA assembly method which took the name of its author Daniel Gibson[2]. Always refers to the isothermal reaction described in the paper.

LCR: Ligase Cycling Reaction, a DNA assembly method[3].

MCS: Multiple Cloning Site, a region sometimes present in plasmids that contains a large number of unique restriction sites. It is used to insert DNA fragments in the plasmid using restriction/ligation cloning methods.

PCR: Polymerase Chain Reaction.

Scar: a sequence that is left between two DNA fragments after they are joined with each other. The DNA assembly standard or method used to join the two fragments defines the scar sequence (see **Figure 18**).

Scarless: a DNA assembly standard or method that is able to join to DNA fragments without leaving any scar sequence in between (see **Figure 18**).

SLIC: Sequence and Ligation Independent Cloning, a DNA assembly method[4].

Sticky end: a short overhang (usually three or four bp) on the flank of a double stranded DNA fragment. Typically produced via restriction digestion.

$\mathbf{T_m}$: melting temperature of two complementary DNA strands.

TU: Transcription Unit, a region of DNA that is transcribed under the control of a single transcriptional promoter.

Type IIs restriction enzyme: while the most commonly used restriction enzymes have palindromic recognition sequences and cut in their middle, type IIs restriction enzymes recognise non-palindromic sequences and cut a few base pairs outside them.

UTR: Un-Translated Region, a region of DNA that is transcribed but not translated.

1. Introduction

1.1. Synthetic biology and DNA assembly

Synthetic biology is a field of research born around the turn of the millennium that focuses on the design and construction of new biological components such as genes, gene networks, or whole genomes, or the redesign of existing ones. Synthetic biology builds on the advances in molecular biology, cell biology, systems biology and genetic engineering but aims to frame this knowledge within an engineering-inspired approach: it aims to fully understand and characterise these new or redesigned core biological components so that they can be modelled, tuned and assembled into larger integrated systems that perform specific tasks[5].

The development of molecular cloning and PCR in the 1970s and 1980s led to impressive biotechnology applications, such as the production of human insulin in *E. coli*[6] and insect resistant cotton plants[7], but these required years of work, mainly based on trial-and-error and *ad hoc* solutions. The two main obstacles were, and still are, that building novel genetically modified systems is difficult and time consuming, and it is very hard to predict whether they will work as expected. Synthetic biologists believe that an engineering-based approach, employing specifically designed (or redesigned) biological parts that are functionally well understood and easy to be assembled with each other can be extremely helpful. The final goal is to be able to design novel biological systems using the engineering development cycle paradigm (**Figure 1**): initially a set of specifications is defined that describes exactly how the system should behave. The second step is to use these specifications to produce a fully detailed "blueprint" of the system: for example if the system is a bacterial plasmid containing a genetic network, the design will specify what genetic parts need to be included and the full DNA sequence of the plasmid. The third step is to test this design *in silico*, using mathematical models, in order to have a first validation of the design before moving on to more time-consuming and expensive tests. If the system is predicted to work it is then physically assembled, and lastly it is tested experimentally to verify whether it actually complies with the

specifications that were initially set. If it does not, or a problem is encountered at any other step of the cycle, the specifications are updated, a new design is produced and the cycle starts again until one of the prototypes is completely satisfactory.

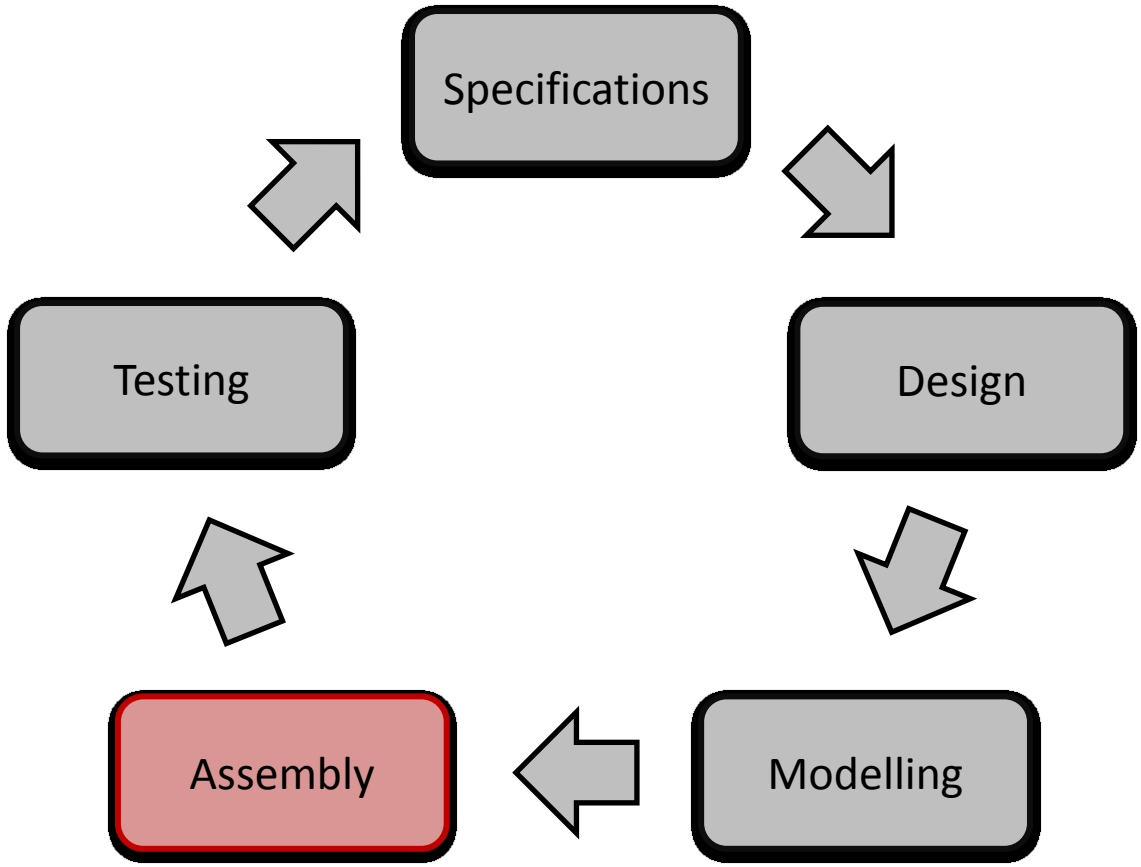


Figure 1: the engineering design cycle, used to develop new technologies and products. Bottlenecks at any step reduce the efficacy of the whole process, and the assembly step is particularly difficult in synthetic biology.

This approach has proven to be very effective as early synthetic biologists, following the push for more engineering in biology, successfully designed and built gene networks that mimicked electrical circuits such as toggle switches[8] and oscillators[9]. Later research in synthetic biology continued along the electrical engineering parallel with the development of complex gene circuits able to detect the edge between and illuminated and a dark area[10], to form an LCD-like clock made of oscillating "biopixels"[11] and many others[12]. Synthetic biology researchers achieved important successes in other areas as well[13], notably metabolic engineering, with the rational design of complex biosynthesis

pathways for polyketides and non-ribosomal peptides[14] and the production of an engineered yeast strain capable of producing commercially viable amounts of artemisinin[15]. A number of biotechnology companies have also been founded, which leverage synthetic biology techniques and tools to produce commercial products, such as Amyris[16], Gingko Bioworks[17] and Synthetic Genomics[18].

The advancement of the field requires the simultaneous development of our ability to perform all the steps in the engineering cycle, in order to be able to iterate over it as quickly as possible, avoiding bottleneck effects, but the assembly step has always been very problematic: DNA assembly is unpredictable, expensive and time consuming[19,20]. It requires a high level of craftsmanship and constantly entails finding *ad hoc* solutions to ever new and unexpected problems, which is clearly unsuitable for a development process such as the one that synthetic biology aspires to. Throughout the years numerous new techniques have been developed and today our ability to build DNA molecules and insert them in living organisms has significantly improved: building a bacterial plasmid, which in the 1970s used to be an endeavour worthy of a multi-author high-profile publication[21], today is routinely performed by undergraduate students[22]. Technology is pushing forward and new frontiers are opening up, such as genome-wide editing[23,24] and *de novo* genome construction[25], but most synthetic biology projects still involve plasmid-sized constructs (10^3 – 10^4 bp), and there still is not a completely satisfactory method to deal with assembly at this scale.

The parallel with engineering disciplines can provide some helpful insight for solving this problem: in electrical engineering, for example, all basic components are designed according to standard formats that allow them to be easily assembled with each other in a modular fashion. DNA assembly techniques only define reactions that join DNA molecules but, while it is important to keep improving the current methods and developing new ones, the history of engineering teaches us that these cannot be exploited to their full potential without framing them in carefully designed standards. The synthetic biology community soon realised that developing standard formats and

workflows for modular DNA assembly can make the process simpler, faster and more reliable, besides promoting exchange of material and knowledge between different laboratories[26]. The next chapters will review the current state of the art of DNA assembly, examining the most successful techniques and standards available to synthetic biologists today.

1.2. DNA assembly methods

1.2.1. Restriction & ligation methods

Gene cloning using type II restriction enzymes and DNA ligases has been employed for 40 years[21] in molecular biology and genetic engineering. It relies on a "cut & paste" approach, as shown in **Figure 2**: the DNA fragments to be joined are initially cut using a restriction enzyme digestion to generate appropriate sticky ends, then they are mixed together in a solution where the compatible sticky ends can anneal to each other and be covalently joined by a DNA ligase.

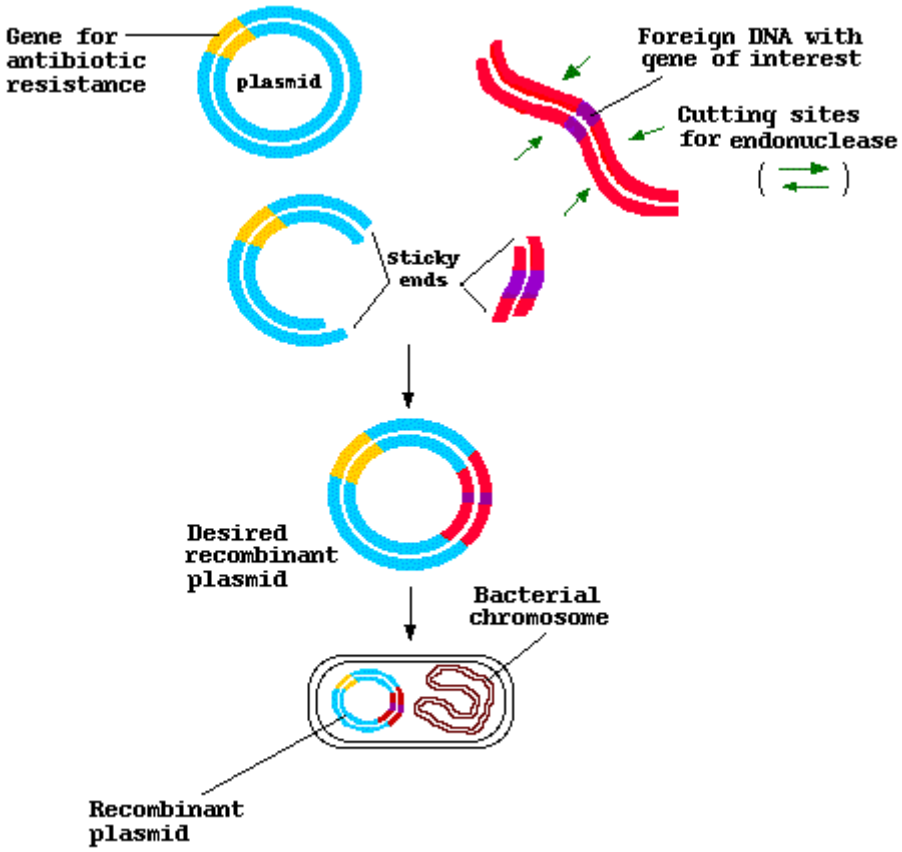


Figure 2: restriction & ligation-based molecular cloning. The plasmid (cyan) and the insert (magenta) are separately digested with a restriction enzyme to generate sticky ends. The two fragments are then mixed in a ligation reaction, where the matching sticky ends guide their assembly in the correct orientation. The new hybrid plasmid is then introduced in bacterial cells, where it is maintained and replicated alongside the bacterial chromosome.

One of the biggest issues with this method is that the parts being assembled cannot contain the recognition sites that are used during the assembly process, and the likelihood for them to occur is actually quite high: type II restriction enzymes typically recognise 6 bp sequences, which are found about once every 4096 bp on average (4^6). Researchers attempted to overcome this limitation by exploiting methylation systems: pairwise selection assembly (PSA)[27] uses a CpG methylase to methylate and thus "deactivate" any undesired restriction sites present on DNA parts, while making sure that the enzyme cannot access the sites required for the assembly process, so that they remain active.

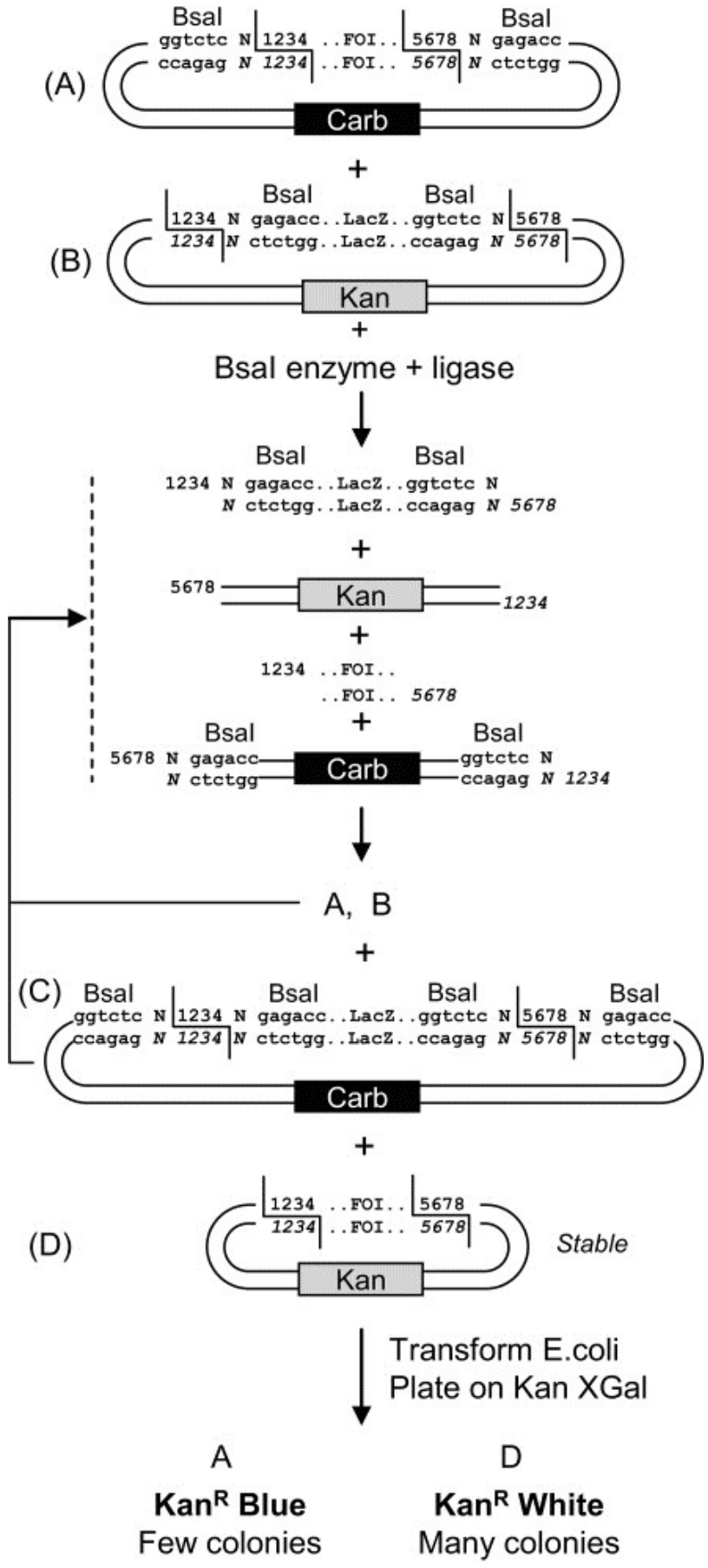
The MASTER Ligation method[28] instead takes the opposite approach and employs the MspJI enzyme, which only cuts methylated recognition sites. These can be added to the DNA fragments using methylated oligonucleotides either by PCR or ligation. Both PSA and MASTER employ type IIs restriction enzymes (or similar ones such as MspJI) that cut a few base pairs away from their recognition sites: this confers a high level of flexibility by giving researchers the freedom to choose the sequence of the sticky ends generated, and is also beneficial for the efficiency of the reaction. Researchers have found that the sequence of these 4 bp regions can have a significant impact on the number of colonies obtained and on the accuracy of assembly[29–31].

An important feature of PSA and MASTER ligation is that these techniques have the ability to use assembled constructs as starting points for new assembly reactions in a hierarchical fashion. This is particularly important for PSA which can only assemble two parts at a time, and would otherwise be unable to build complex constructs. MASTER Ligation can perform multi-fragment assembly, but this comes at a high cost for the accuracy of the reaction.

Type IIs enzymes are also at the core of one of the most important advancements in DNA assembly, the development of Golden Gate assembly[32]. In this technique DNA fragments from one or more entry vectors are assembled into a destination vector using a simultaneous digestion and ligation reaction. As shown in **Figure 3** the system is designed so that all undesired plasmids can be

selected against via antibiotic or LacZ-mediated colour selection, and the colonies containing the desired plasmid are easily identified. This way all plasmids and enzymes involved can be simultaneously mixed in a one-pot reaction, greatly simplifying the preparation. It is important to note that this design generates a reaction that is spontaneously driven to completion, since for as long as it is incubated any undesired plasmid gets digested and the fragments have a chance to be irreversibly ligated to form the correct construct. For this reason Golden Gate achieves an incredibly high efficiency, unparalleled by any other restriction/digestion-based technique, and it has been shown to be able to assemble as many as 15 fragments in parallel[30] and constructs as big as 33 kb[33].

Figure 3 (next page): the Golden Gate assembly schematic[32]. A is the entry vector, a plasmid containing the DNA Fragment Of Interest (FOI) to be cloned, flanked by two inward-facing BsaI sites. B is the destination vector that contains a colour-selection gene (LacZ) flanked by outward-facing BsaI sites that will be replaced with the FOI. The two carry different antibiotic selection markers (carbenicillin and kanamycin). Both plasmids are mixed together in a simultaneous restriction/ligation reaction, where the type IIs restriction enzyme BsaI separates the FOI and LacZ from the respective backbones. The fragments can ligate back to reconstitute plasmids A and B, which are then cut again, or cross-ligate to generate plasmids C and D. Plasmid C can also be cut again, but plasmid D, the desired construct that contains the FOI and the destination backbone, is the only one that does not contain any BsaI recognition sites and cannot be cut again. This drives the reaction towards completion, *i.e.* producing plasmid D. Once the reaction mix is used to transform *E. coli* cells under kanamycin selection, plasmids A and C are not viable, any remaining plasmids B will look blue on the plate, and all the non-blue colonies will contain the desired plasmid D.


(A)
BsaI
ggtctc N 1234 ..FOI.. 5678 N gagacc
ccagag N 1234 ..FOI.. 5678 N ctctgg
BsaI
Carb
+
(B)
BsaI
1234 N gagacc..LacZ..ggtctc N 5678
1234 N ctctgg..LacZ..ccagag N 5678
BsaI
Kan
+
BsaI enzyme + ligase
BsaI
BsaI
1234 N gagacc..LacZ..ggtctc N
N ctctgg..LacZ..ccagag N 5678
+
5678
Kan
1234
+
1234 ..FOI..
..FOI.. 5678
+
BsaI
BsaI
5678 N gagacc
N ctctgg
Carb
ggtctc N
ccagag N 1234
A, B
+
(C)
BsaI
BsaI
BsaI
BsaI
ggtctc N 1234 N gagacc..LacZ..ggtctc N 5678 N gagacc
ccagag N 1234 N ctctgg..LacZ..ccagag N 5678 N ctctgg
Carb
+
(D)
1234 ..FOI.. 5678
1234 ..FOI.. 5678
Kan
Stable
Transform E.coli
Plate on Kan XGal
A
KanR Blue
Few colonies
D
KanR White
Many colonies

1.2.2. Long overlap methods

Long overlap methods differ from restriction & ligation methods because the homology regions between the DNA fragments being joined are usually around 20-50 bp long, much more than the classic 4 bp sticky ends generated by restriction enzymes. There is a huge variety of methods that rely on long homology regions, both *in vitro* and *in vivo*, and they differ mainly in the mechanism by which they render these regions single stranded, so that they are free to anneal to the homology regions of the fragments they are meant to be joined with. Starting with the *in vitro* methods, one of the simplest yet most successful ones is CPEC which is derived from overlap extension PCR[34], an old technique developed to fuse two to four DNA fragments into linear constructs[35].

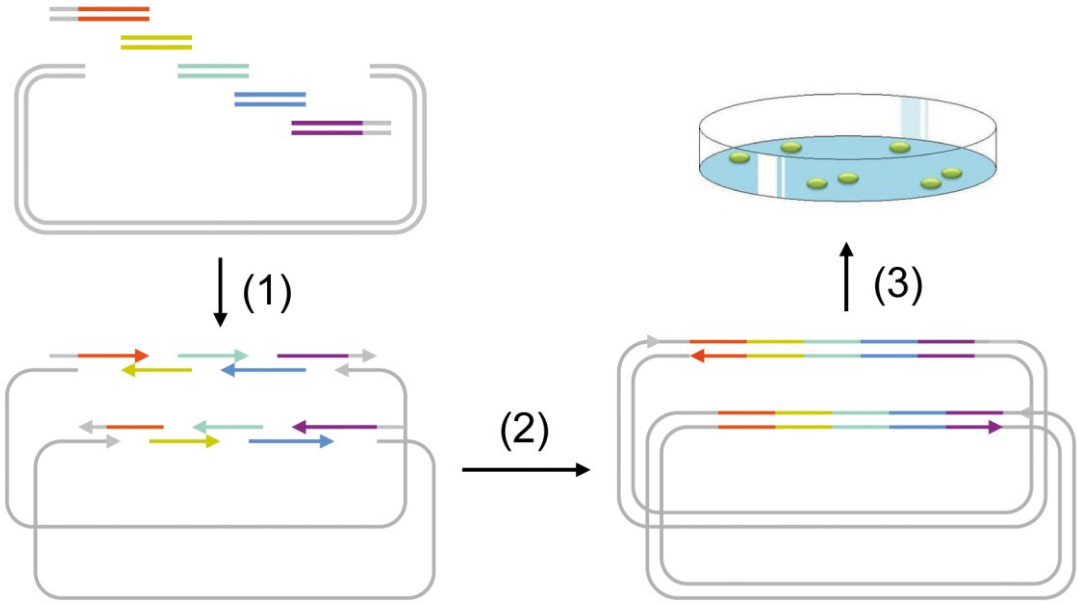


Figure 4: CPEC cloning schematic[1]. The fragments to be cloned (coloured) are mixed in one pot with the destination backbone (grey) in a PCR-like reaction. The fragments' sequences overlap so that they can essentially act as PCR primers on each other. The result is a fully assembled construct where the original fragments are separated by nicks that are repaired *in vivo* after cell transformation.

CPEC is essentially set up as a high-fidelity PCR, with the difference that instead of template or primers it contains a number of DNA fragments to be joined together to form a plasmid (**Figure 4**). The sequences of these fragments need to overlap with each other at the extremities by enough

base pairs to have a T_m of about 60-70°C (typically about 20 bp), to ensure efficient and specific annealing. The reaction mix undergoes temperature cycling like a PCR, and once all DNA is denatured these homology regions essentially act as primers on the neighbouring fragments. The result is that at every cycle new nicked circular molecules are formed, that will be automatically repaired *in vivo* once transformed in the host. The authors recommend the technique for constructs up to 20 kb in size and made of up to 4 parts. It has also been shown that the efficiency and specificity of OE-PCR, and thus of CPEC as well, can be improved with the use of specifically designed GC-rich homology regions[36].

While the use PCR to join DNA fragments is advantageous for its simplicity and efficiency, it also has a few notable drawbacks: it has a chance of introducing mutations and it has problems with very long or very GC-rich fragments, with secondary structures and with repeated motifs. All these issues might require *ad hoc* troubleshooting, labour-intensive gel extraction of the correct fragment or might even make the reaction impossible. For this reason researchers developed methods that make only the homology regions single stranded, instead of the whole fragment, so that DNA fragments can be joined without the need for DNA polymerisation.

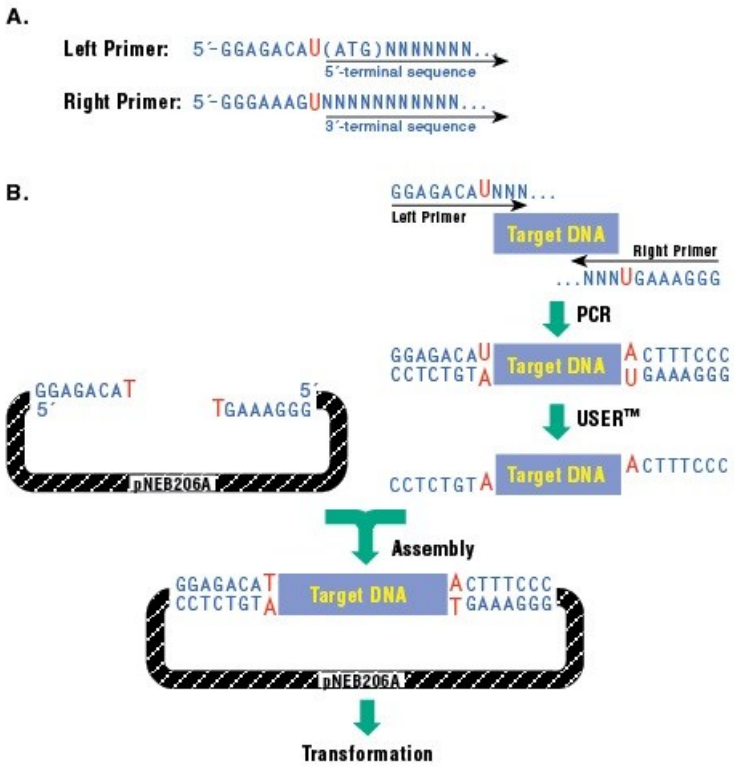


Figure 5: USER cloning schematic[37]. (A) Design of the primers to prepare DNA fragments for USER assembly: in this example a 8 bp overlap sequence is used, with an uracil residue (red) located between the overlap region and the priming region of the oligonucleotide. (B) After PCR amplification with these primers, the target DNA is flanked by double stranded overlap regions that contain the uracil residue on the 5'-3' strand. This DNA fragment is then mixed with the USER enzymes that excise the uracils and nick the 5'-3' strand in those locations. The 8 bp fragments are thus released leaving 8 bp 3'-5' overhangs on the flanks of the target DNA. A destination backbone prepared similarly is also included in the USER reaction so that as the overhangs are generated the two can anneal, generating the final construct, which can be used for bacterial transformation.

A method that gained some traction is USER cloning[38] (**Figure 5**) where the DNA parts are prepared for assembly with a PCR amplification that adds on each flank a deoxyuridin residue (dU) followed by a 6-10 bp homology region. The parts are then simply mixed together in a solution containing uracil DNA glycosylase and endonuclease VIII: the first removes the dU residue from the DNA backbone, and the second cuts it at the abasic site. This way the 6-10 bp fragment is spontaneously released from the DNA part making the homology region single stranded, so that the DNA parts can anneal to each other to form a nicked construct that can be used for bacterial

transformation. One of the main drawbacks of this method is that there are very few DNA polymerases able to correctly amplify DNA molecules containing dU residues, such as Taq, which are not as good as the latest polymerases in terms of fidelity, speed and specificity. Recently a new enzyme has been engineered, based on Pfu, that has all the advantages of the latest DNA polymerases while also being compatible with dU modifications[39].

A similar method was also developed, called Cross-Lapping *In Vitro* Assembly (CLIVA)[40], which employs phosphorothioate chemistry to cause breaks in the DNA backbone leaving the homology regions single stranded. The reaction is completely enzyme-free, as phosphorothioate modified nucleotides are spontaneously cleaved when exposed to iodine in an ethanolic solution, and multiple ones can be used to obtain longer (36-38 bp) single stranded regions. Additionally this modification can be placed on any of the four bases (whereas USER cloning requires specifically the presence of dU/dA pairs) and is well tolerated by all DNA polymerases. The authors demonstrated the successful assembly of a 22 kb 6-part plasmid, but only <10% of the colonies resulted to be correct: both USER and CLIVA work best with three or four-part assemblies.

The methods just described, beside requiring expensive modified oligonucleotides and exotic reactions, also necessarily require PCR amplifications in order to attach the homology regions to the DNA fragments, and are thus subject to all the limitations and drawbacks of PCR, as mentioned earlier for CPEC. The development of engineered restriction enzymes that cut only one strand of the DNA duplex gave researchers a new tool that has been used to generate long overhangs for various applications[41], including DNA assembly[42] (**Figure 6**).

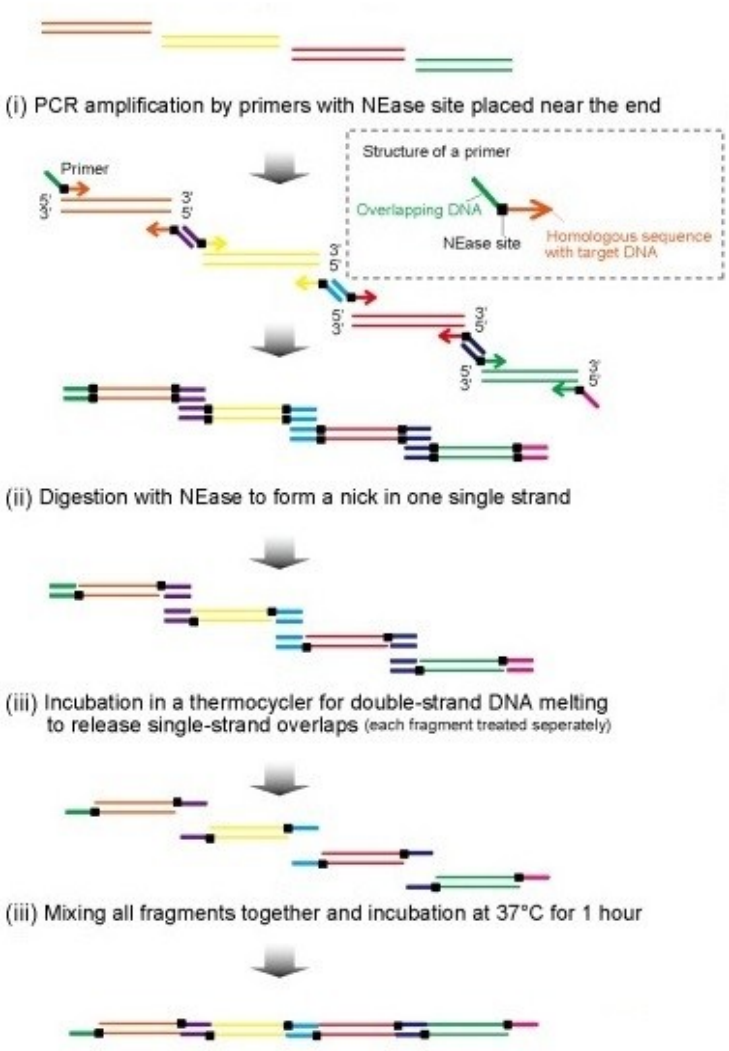


Figure 6: nicking enzymes-based DNA assembly[42]. The DNA fragments are prepared for assembly with a PCR amplification that adds the overlap regions and the nicking enzyme recognition sites (see inset). The fragments are then digested and the short single stranded fragments are released using high temperatures, generating single stranded overhangs. Finally the fragments are mixed in a single assembly reaction where they anneal to each other according to their overlap region, forming the final construct. If this final construct is a circular plasmid it can then be used for cell transformation.

The DNA fragments to be assembled are prepared by adding the nicking enzyme sites and the homology regions to their flanks, which can be still be done through PCR amplification but also through other methods. The prepared parts are then digested with the nicking enzymes and incubated at a high temperature to release the small fragments and expose the single stranded

homology regions. Finally they are mixed together and incubated with T4 DNA ligase to be joined with each other, forming the desired final construct. The authors attempted a six-part assembly using 15 bp homology regions, with some success: they obtained "several" (*sic*) colonies, but they had to use a semi-hierarchical approach where pairs of consecutive parts were incubated separately before mixing the pre-assembled pairs together to complete the assembly reaction. The mix was then transformed, and all three colonies that were checked proved to be correct.

Another method, developed by Schmid-Burgk *et al.* [43] as an improvement of a previous technique called SLIC[4], exploits the 3'-5' exonuclease activity of T4 DNA polymerase to generate single stranded regions of a defined length (**Figure 7**). In SLIC the enzyme is used in absence of dNTPs, and it "chews back" the 3'-5' strand at both ends of any DNA fragment generating single stranded regions of undefined length. In the version developed by Schmid-Burgk *et al.* the process is stopped at a defined position by designing the ends so that they do not contain one of the four nucleotides (*i.e.* A), and by placing the first A where the digestion should stop. If dATP is included in the reaction mix T4 DNA polymerase will digest the 3'-5' strands until it finds the first A in the sequence, and it will stall there, generating an overhang of defined length. Three to five consecutive "stop" bases are required to ensure complete stalling of T4 DNA polymerase. It was shown that this method could successfully assemble up to four DNA fragments in parallel using 20 bp homology regions, and it is amenable for hierarchical assembly by simply releasing the intermediate constructs from the backbone with a type IIs restriction digestion.

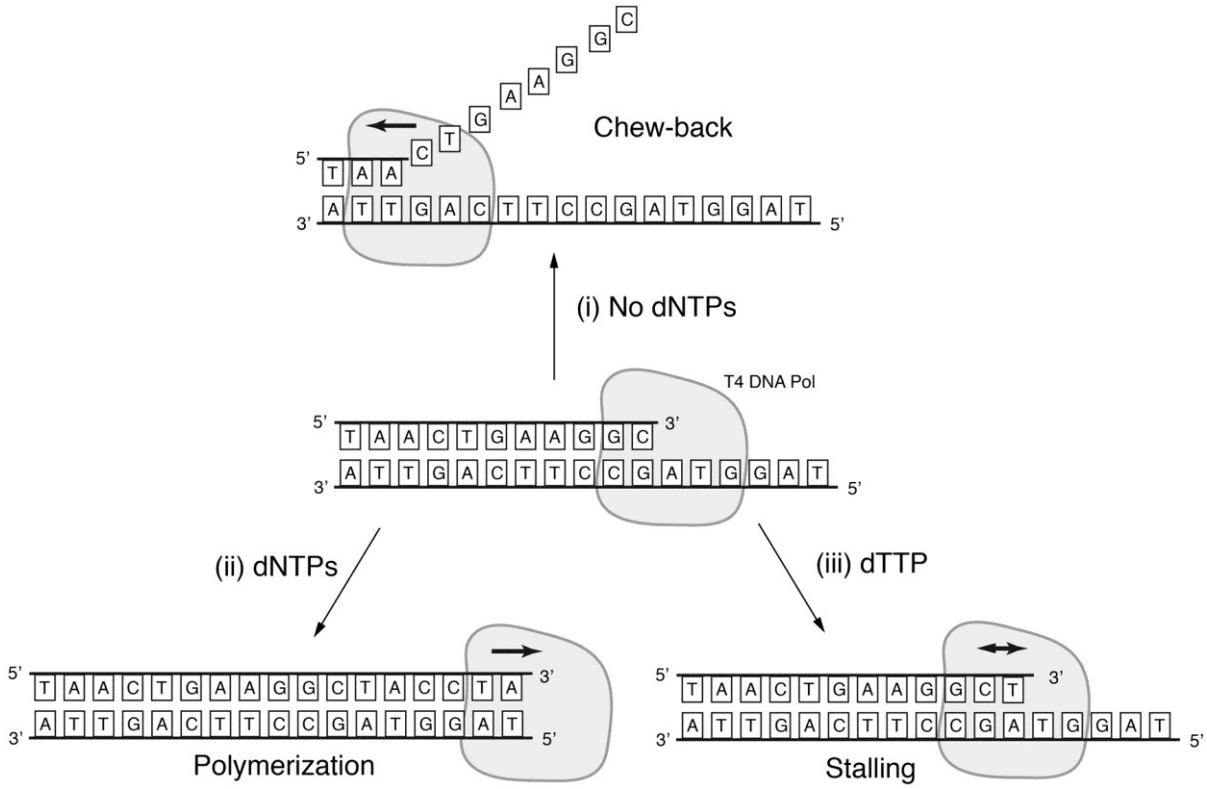


Figure 7: T4 DNA polymerase-based cloning[43]. The figure shows the three possible behaviours of T4 DNA polymerase on the end of a DNA fragment. (i) In absence of dNTPs the 3'-5' exonuclease activity of the enzyme progressively removes all the nucleotides from that strand. This mechanism is employed in SLIC. (ii) In presence of dNTPs the enzyme leaves a blunt end unchanged or fills in an end with a 5' overhang generating a blunt end. (iii) In presence of a single dNTP (dTTP in the example) the enzyme chews back the 5'-3' strand until it finds the corresponding residue (T in the example), where it then stalls, generating a 5' overhang. This is used by Schmid-Burgk *et al.* to guide the annealing and assembly of multiple DNA fragments.

Stopping T4 DNA polymerase chew-back puts a few constraints on the sequence of the homology region, which can only contain three of the four oligonucleotides. This prevents scarless assembly and makes sequence optimisation more difficult. Alternatively it is possible to use all four nucleotides and the reaction is stopped by simply adding any dNTP after a certain incubation time[4], but this creates a population of molecules digested by different amounts: those that are not digested enough do not expose the full homology regions and cannot be assembled, while those that are overdigested will be able to anneal correctly to their target, but the final construct will contain gaps instead of just nicks, which has been shown to be deleterious for the efficiency of

assembly[42]. The method was nonetheless able to reliably assemble five-part plasmids although with small (<500 bp) inserts, while a ten-part plasmid could only be assembled with an accuracy of about 20%. Additionally, this method requires longer (40 bp) homology regions to work best, compared to the methods to create overhangs of a defined length (4-20 bp).

One of the most important breakthroughs in *in vitro* long-overlap based DNA assembly happened when Gibson *et al.*[2] devised a method to keep the complete sequence independence of SLIC while obviating the problem of generating gapped constructs and actually significantly increasing assembly efficiency compared. The method essentially is very similar to SLIC, but with the addition of an *in vitro* DNA repair system: it uses T5 exonuclease to generate the overhangs, Phusion DNA polymerase to fill in the gaps and Taq DNA ligase to seal the nicks (**Figure 8**).

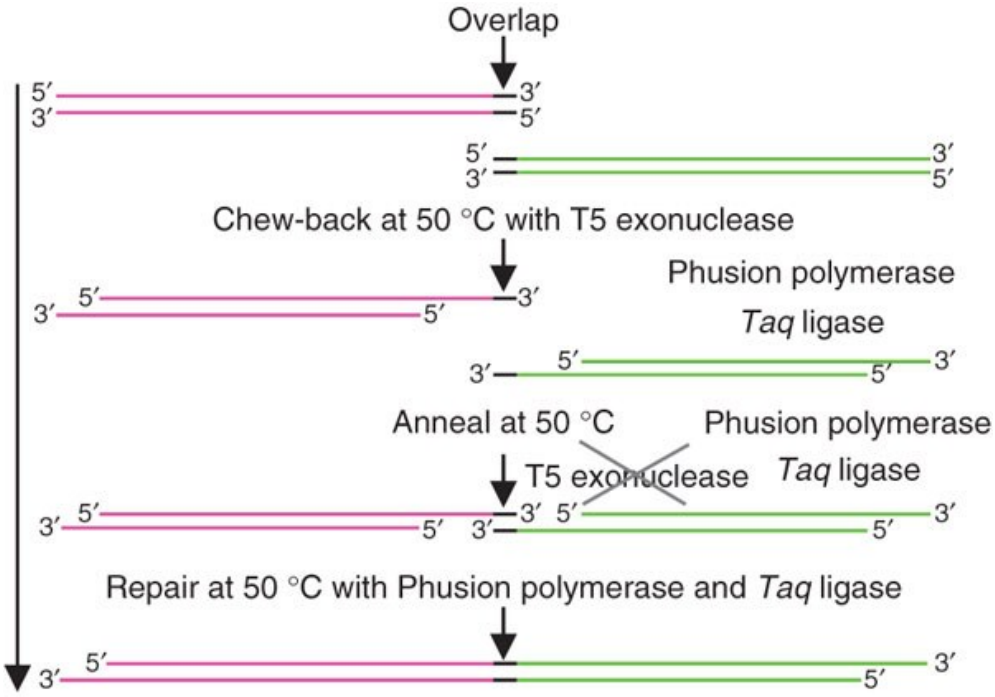


Figure 8: mechanism of action of Gibson isothermal assembly[44]. DNA fragments with overlapping regions (black) are mixed in a reaction together with three enzymes, at a fixed temperature of 50°C. T5 DNA exonuclease chews back the 5' ends, making the overlap regions single stranded and available of assembly. Because of their polarity these overhangs cannot be filled in by the DNA polymerase. Once these anneal to their complementary DNA fragment, they act as primers so that Phusion DNA polymerase can fill in the gap left by T5 polymerase. Finally Taq DNA ligase seals the nicks left by the polymerase, generating an intact construct.

The method is also extremely simple as all three enzymes are mixed simultaneously with all the parts to be assembled (which have to be equipped with the appropriate homology regions), and the solution is incubated at a fixed temperature of 50°C for one hour after which it is transformed directly. All enzymes act simultaneously during the isothermal incubation and T5 exonuclease, the only non-thermostable enzyme, is also slowly heat-inactivated so that during the last part of the incubation time only the DNA repair enzymes are active, to ensure the integrity of the products.

Similarly to SLIC, Gibson isothermal assembly requires 40 bp homology regions to work best, even though it has been shown to work with much longer ones as well (450 bp) by increasing the amount of T5 exonuclease in the mix. The authors showed that the technique can be used to assemble constructs as big as 583 kb from four fragments and in another publication it was used to assemble a 16.3 kb construct starting from 60 bp oligonucleotides, proving that it works at different scales of assembly: eight single stranded 60 bp oligonucleotides were assembled into 284 bp double stranded fragments, which were then assembled five by five into 1.2 kb fragments, and again five by five into 5.6 kb fragments. Finally three of these were joined to build the complete 16.3 kb construct[45].

There is one more approach to long overlap-based *in vitro* DNA assembly, which employs DNA "bridges" to join fragments. These bridges carry homology regions for both of the parts they are meant to join, and they have been implemented in various fashions, using nicking endonucleases[46] and Gibson isothermal assembly[47], but the version that uses Ligase Cycling Reaction (LCR), pioneered by Pachuk *et al.*[48] and refined by de Kok *et al.*[3], is particularly interesting because it achieves very high efficiency and accuracy.

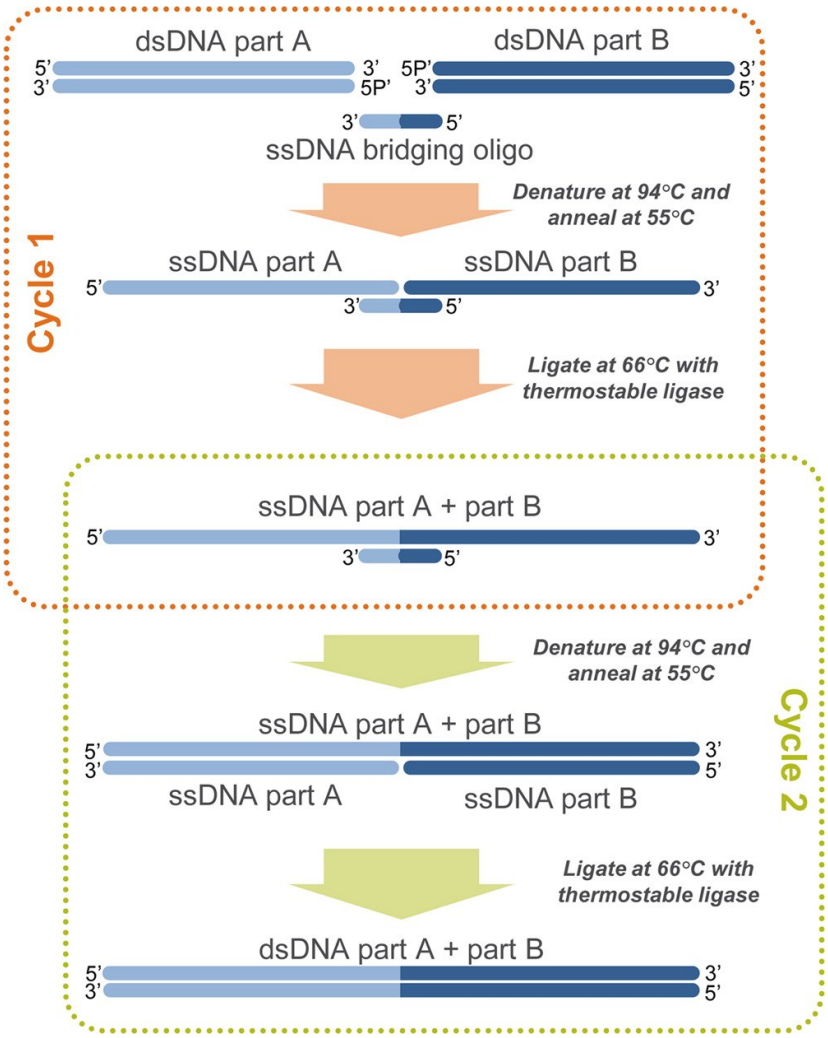


Figure 9: mechanism of action of the Ligase Cycling Reaction[3]: in the example two DNA fragments to be assembled are mixed in a reaction that contains a thermostable ligase and bridging oligonucleotides that overlap with both fragments. The reaction is incubated using thermal cycles similar to those of a PCR. During the first cycle the two denatured fragments anneal to the oligonucleotides, so that only a nick is left between them, that is sealed by the ligase. From the second cycle both the bridging oligonucleotides and the previously ligated fragments can act as template for more fragments to anneal to them and become available for ligation. The same process can be applied to multiple DNA fragments without any changes using multiple bridging oligonucleotides.

LCR (**Figure 9**) employs single stranded bridge oligonucleotides, constituted by two homology regions of varying length (13-40 bp) depending on the desired T_m, that bring together the two fragments to be joined by annealing to both of them. In order to do so temperature cycling is used: a high-temperature denaturation step separates the two strands of the fragments, an annealing step brings the temperature down to a point where the bridge oligonucleotides can anneal to the melted

strands, and finally during ligation step the reaction mix is brought to the optimal temperature for thermostable ligase activity, so that the two single stranded DNA parts brought together by the bridge oligonucleotides can be ligated. During the following cycle these ligated fragments are melted again and can then act as template to bring other unligated fragments together, driving the reaction forward. The authors showed that it is able to assemble up to 20 parts simultaneously and constructs up to 20 kb with very high accuracy, which is unparalleled by any other *in vitro* DNA assembly technique.

Lastly, there are a variety of *in vivo* long overlap-based DNA assembly methods that exploit the natural recombination capabilities of different organisms, including *Bacillus subtilis*, *Escherichia coli* and certain plants[49], but the most widely used host is certainly *Saccharomyces cerevisiae*. The assembly protocol itself is extremely simple: once the parts to be assembled are equipped with the necessary homology regions (usually around 40 bp), they are transformed in yeast cells where they are spontaneously joined together by the natural recombination machinery. The main complexity comes from making the yeast cells competent, for which there are two commonly used protocols, a quick one[50] and a more complex but reportedly more efficient one[51].

There are notable disadvantages to yeast recombination, such as the fact that colony growth requires two to four days (instead of just one for *E. coli*) and that plasmid isolation is much less efficient, but the technique is remarkably efficient and accurate and works at any scale. It has been used to assemble ~1 kb genes from 38 oligonucleotides[52], ~20 kb pathways from 9 gene-sized fragments[53] and the entire *M. genitalium* genome from 25 fragments of about 24 kb each[54]. It has also been shown to be extremely accurate[3] and to tolerate homology regions as short as 20 bp[52] and as long as 62 kb[55]. Yeast recombination has been widely used for a variety of purposes and a number of improvements have been devised, such as placing the origin of replication and the selectable marker on two separate fragments when assembling plasmids to reduce background colony growth[56], and adding more origins of replication to stably maintain large GC rich constructs[55].

1.2.3. Recombinase-mediated methods

Site-specific recombinases are enzymes that catalyse the insertion, excision, inversion or exchange of DNA fragments, both *in vivo* and *in vitro,* guided by specific recognition sites (**Figure 10**). The most commonly used types are the Cre and Flp tyrosine recombinases[57], and the lambda and phiC31 serine recombinase[58,59], which have been exploited for a number of purposes including library cloning, genome editing and DNA assembly.

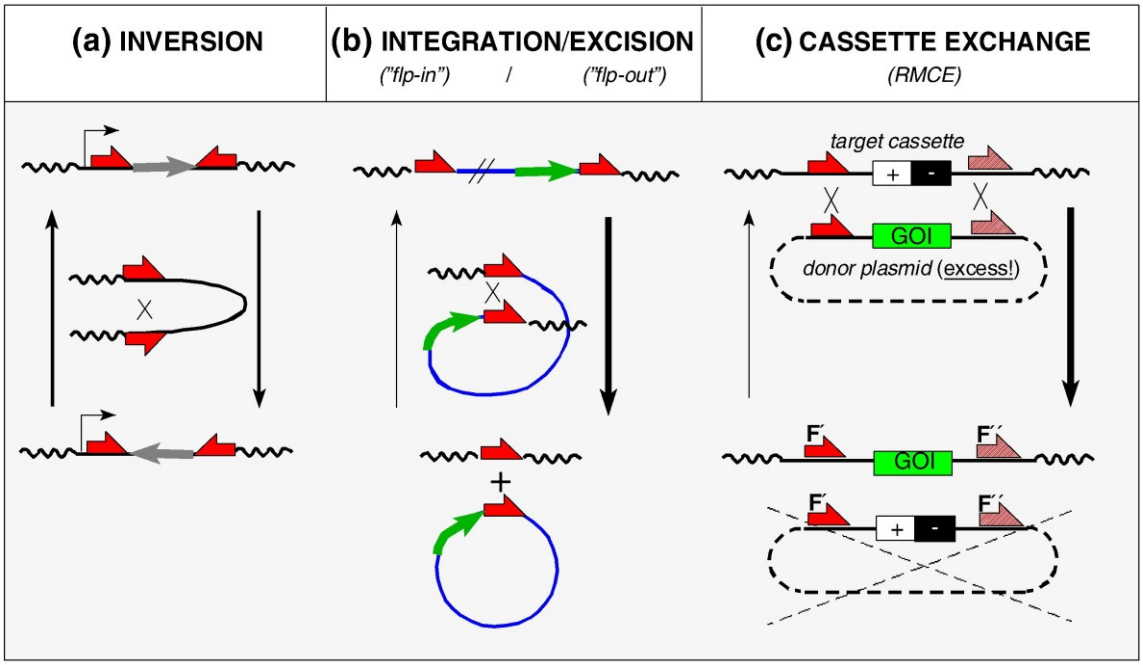


Figure 10: an overview of the reactions that can be catalysed by recombinase enzymes on DNA substrates. The red arrows represent the recombinase recognition sites and their orientation. The black arrows represent the fact the, in general, recombination reactions can go in both directions. (a) Two recombination sites that face each other catalyse the inversion of the orientation of the DNA fragment between them (grey). (b) Two recombination sites facing the same direction catalyse the excision of the region between them (blue and green). Likewise a plasmid containing a recombination site can be integrated in DNA region containing the same recombination site. (c) An exchange of DNA can happen between two DNA regions containing two different recombination sites (red and pink). In the example a donor plasmid containing a Gene Of Interest (GOI) exchanges a cassette with a chromosomal region. The excess of donor plasmid is meant to drive the reaction in the desired direction. From Turan *et al.*[57]

The substantial difference from the previously presented methods is that while those are guided by base pair homology recognition, here the process is mediated by recombinase proteins. One of the main advantages of recombinase-based methods is that they allow new types of DNA manipulation that are impossible with other DNA assembly methods, such as excising and removing a fragment that is not needed anymore from a previously assembled construct[60], or to exchange fragments between two plasmids. This can be used for example to exchange a lethal gene with a gene of interest to remove background colony formation and achieve near-100% cloning efficiency[58], or to "recycle" previously assembled constructs by moving part of them from a plasmid to another[59]. On the other hand the number of sequences that recombinases can recognise, and thus the number of fragments that can be assembled simultaneously, is very limited. For this reason there has been a lot of research aimed at expanding the number of orthogonal recombination sites available[59,61–63], and the effort has been particularly successful with the phiC31 recombinase: six orthogonal sites have been discovered, allowing the parallel assembly of up to six fragments, although with low accuracy (<20% of the colonies obtained contained the expected construct)[59]. It is also important to note that recombination sites are quite large inverted repeats (25-46 bp) and remain in the constructs as "scars" after assembly, which can be problematic because their sequence cannot be modified: this makes it very difficult to troubleshoot possible unwanted interactions with the neighbouring sequences.

1.3. DNA assembly standards

1.3.1. Restriction & ligation standards

Restriction and ligation-based DNA assembly has been very widely adopted and the synthetic biology community has developed a number of improvements for it, most notably with the development of the BioBrick standard[64].

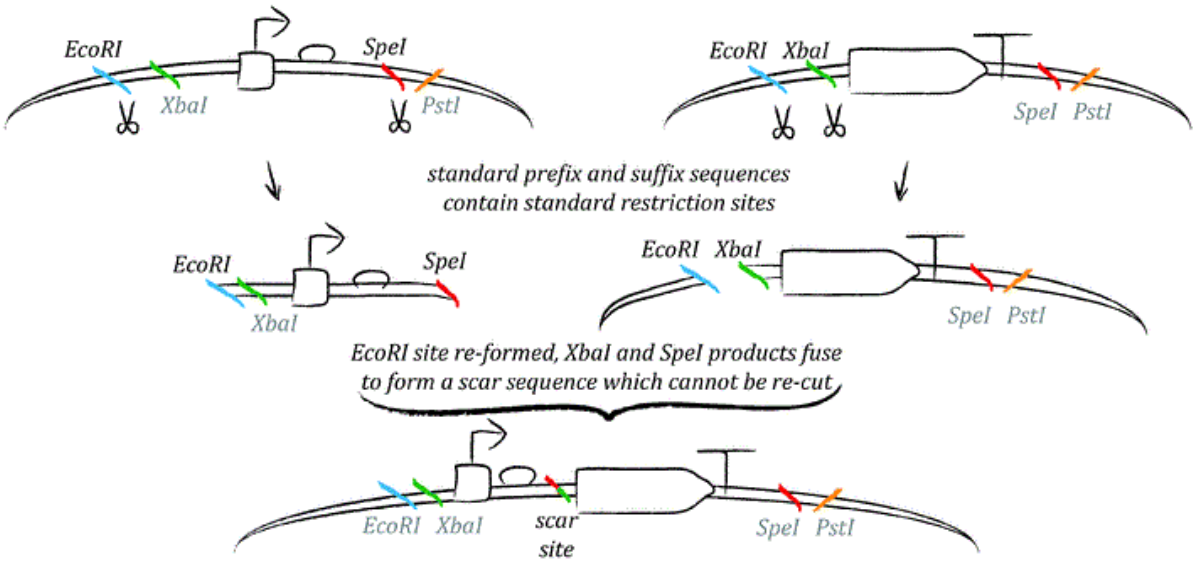


Figure 11: a schematic of BioBrick assembly[64]. On the upper left is shown part of the plasmid that contains the insert to be cloned in the plasmid on the upper right of the figure, both equipped with the standard set of restriction sites of the BioBrick standard (blue, gree, red and orange marks). The plasmids are separately digested with the appropriate enzymes to generate matching sticky ends (SpeI and XbaI generate compatible overhangs). The fragments of interest are purified via gel extraction, and mixed in a ligation reaction to build the desired construct. The new plasmid contains again the standard set of restriction sites, and a scar site is left in between. Image from Ellis *et al.*[65]

This was the first DNA assembly standard developed for synthetic biology, and it has been widely adopted. It essentially consists of the definition of a physical format for DNA fragments, called "BioBricks", where the functional sequence is flanked by standard prefix and suffix regions containing certain restriction sites (**Figure 11**). The assembly process is idempotent, which means that when two BioBricks are joined together they form a new BioBrick flanked by the same prefix and suffix regions so that it can be reused indefinitely in new rounds of assembly. A scar sequence,

which does not contain any restriction sites, is left in the middle of the new molecule between the two parent BioBricks so that they cannot be separated again.

The standard also defines a format for the plasmids to be used to carry the BioBrick and to receive the assembled constructs, and the protocols to be used for the assembly reactions. This strategy allows researchers to run sequential rounds of idempotent assembly to join small BioBricks into larger constructs. A number of improvements were successively made: Shetty *et al.*[66] devised a strategy to join three BioBricks together at the same time instead of just two; Xu *et al.*[67] developed "ePathBrick", a set of BioBrick-compatible modified Duet vectors that provide a simple way of assembling different regulatory elements in a combinatorial fashion; Norville *et al.*[68] designed a "BioScaffold" part that can be used to remove scar sequences or to introduce new parts inside an existing BioBrick, even if they are not compatible with the BioBrick standard (*i.e.* they contain a forbidden restriction site).

New standards that rely on similar mechanisms but use different sets of restriction enzymes were also developed, such as the Standard European Vector Architecture (SEVA)[69] which focuses less on the assembly process itself and more on defining a highly flexible and modular plasmid structure that allows for post-assembly part swapping and is compatible for a broad range of hosts. Litcofsky *et al.*[70] also have developed a plug-and-play plasmid system that is capable of post-assembly modifications, such as the replacement of parts or the insertion of new ones, using a large array of unique restriction sites located in the MCS. Another standard, proposed as an evolution of the BioBrick standard, is BglBricks[71], which utilises more efficient enzymes and leaves a protein fusion-friendly scar between joined parts. Leguia *et al.*[72] improved it by defining the "2ab" assembly strategy which exploits methylation mechanisms and double antibiotic selection to achieve a very high (>96%) success rate, while eliminating the need for gel extraction purification, which is a major hurdle in this type of methods, being very labour-intensive and unreliable.

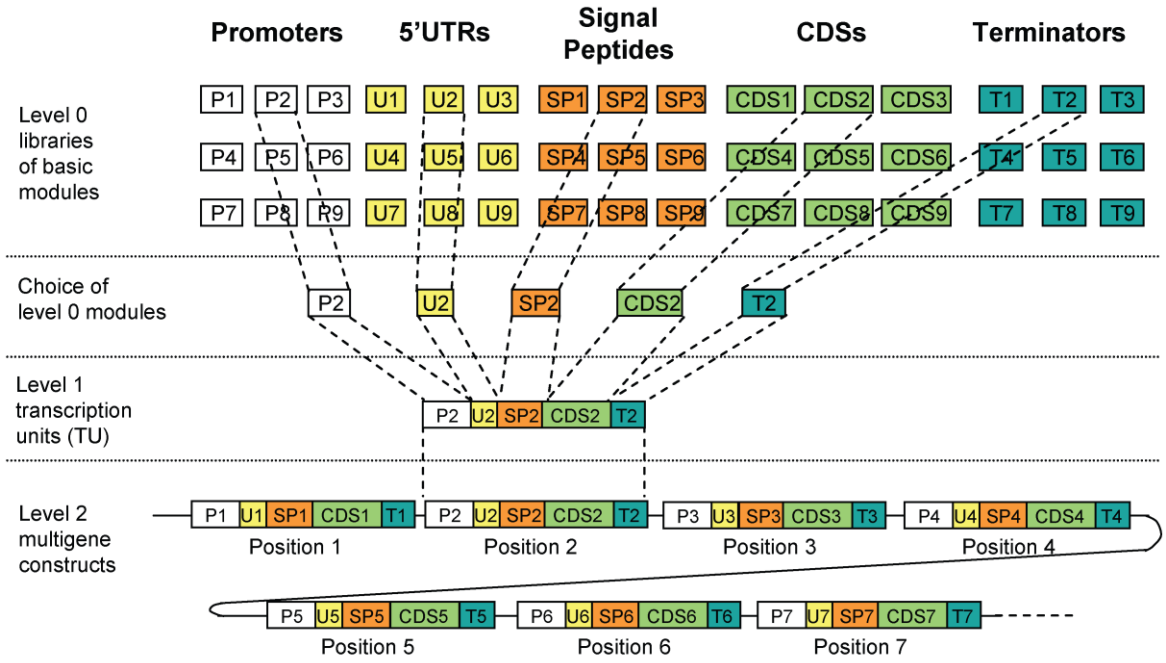


Figure 12: the MoClo standard[33]. The diagram shows how level 0 sub-gene modules are combined to obtain level 1 transcription units and finally level 2 multigene constructs. The position of each level 0 modules in level 1 TUs, and the position of each level 1 TU in level 2 multi gene constructs are defined by the fixed 4 bp overhangs that flank them.

The more advanced restriction and ligation-based assembly technique, Golden Gate, has been employed in a variety of applications[29,30,73], but in order to fully exploit its power a number of standardised modular assembly frameworks have been proposed: MoClo[33] adopts a two-tiers approach with different processes required to go from sub-gene parts to genes, and from genes to pathways, all of which use Golden Gate assembly (**Figure 12**). This is advantageous because gene-level parts tend to be composed by the same fixed sequence of sub-gene parts, such as promoter, open reading frame and terminator (or slightly different configurations depending on the host and the requirements of the project). MoClo takes advantage of this and defines a standard set of 4 bp sticky ends for each category of sub-gene parts, so that they can readily be assembled with each other into a gene unit. The gene-to-pathway level instead must give researchers the freedom to decide the number of parts to be assembled and their order, and MoClo solves this problem by employing a very large array of destination plasmids where genes are assembled into, which define the position of these genes in the final construct. This system allows the parallel assembly of up to 8

fragments simultaneously: the backbone of the plasmid, an end-linker part required for plasmid circularisation and up to six genes. The genes can also be replaced by previously assembled multi-gene fragments to perform multiple rounds of hierarchical assembly. In order to guarantee assembly accuracy and facilitate screening MoClo employs two different type IIs restriction enzymes, three antibiotic selection markers and two visual selection markers.

Another standard framework for modular assembly based on the Golden Gate technique, called Golden Braid[74], has been proposed to provide a simpler alternative to MoClo, while keeping the ability to assemble large pathways from sub-gene parts (**Figure 13**). Golden Braid is very similar to MoClo in its two-tiers structure, in using the Golden Gate method for every assembly reaction and in the way genes are assembled from sub-gene parts using a fixed structure, but in order to limit complexity at the second tier it adopts a very different approach. It uses only four different plasmids divided in two groups, alpha and omega, so that for example any two genes separately cloned into two alpha plasmids con be combined together to form an omega plasmid that contains both genes. It is then possible to combine this omega plasmid with its partner omega plasmid, containing one or more genes, to form a new alpha plasmid containing all their genes, and the cycle starts again. Compared to MoClo this approach is slower, since it only allows pairwise assembly, but it is slightly simpler, if not conceptually at least in the number of plasmids required. An updated version of this assembly strategy, named Golden Braid 2.0[75], was created in collaboration with the MoClo developers bringing a series of improvements: it refines the choice of fixed junctions for the sub-gene to gene assembly in order for these scars to be as “benign” as possible, it simplifies the design of the entry vectors containing the gene-level fragments so that a smaller number of them is required, and facilitates the performance of non-standard assemblies.

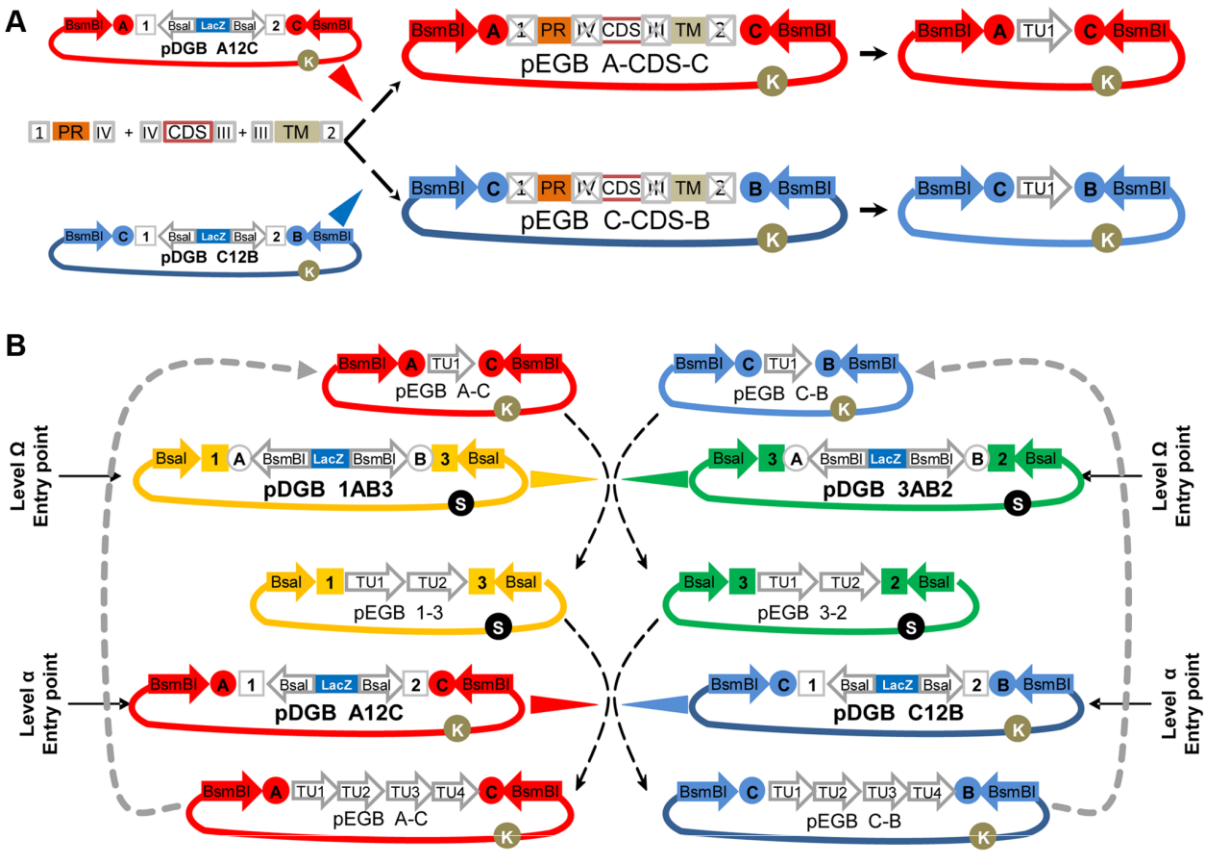
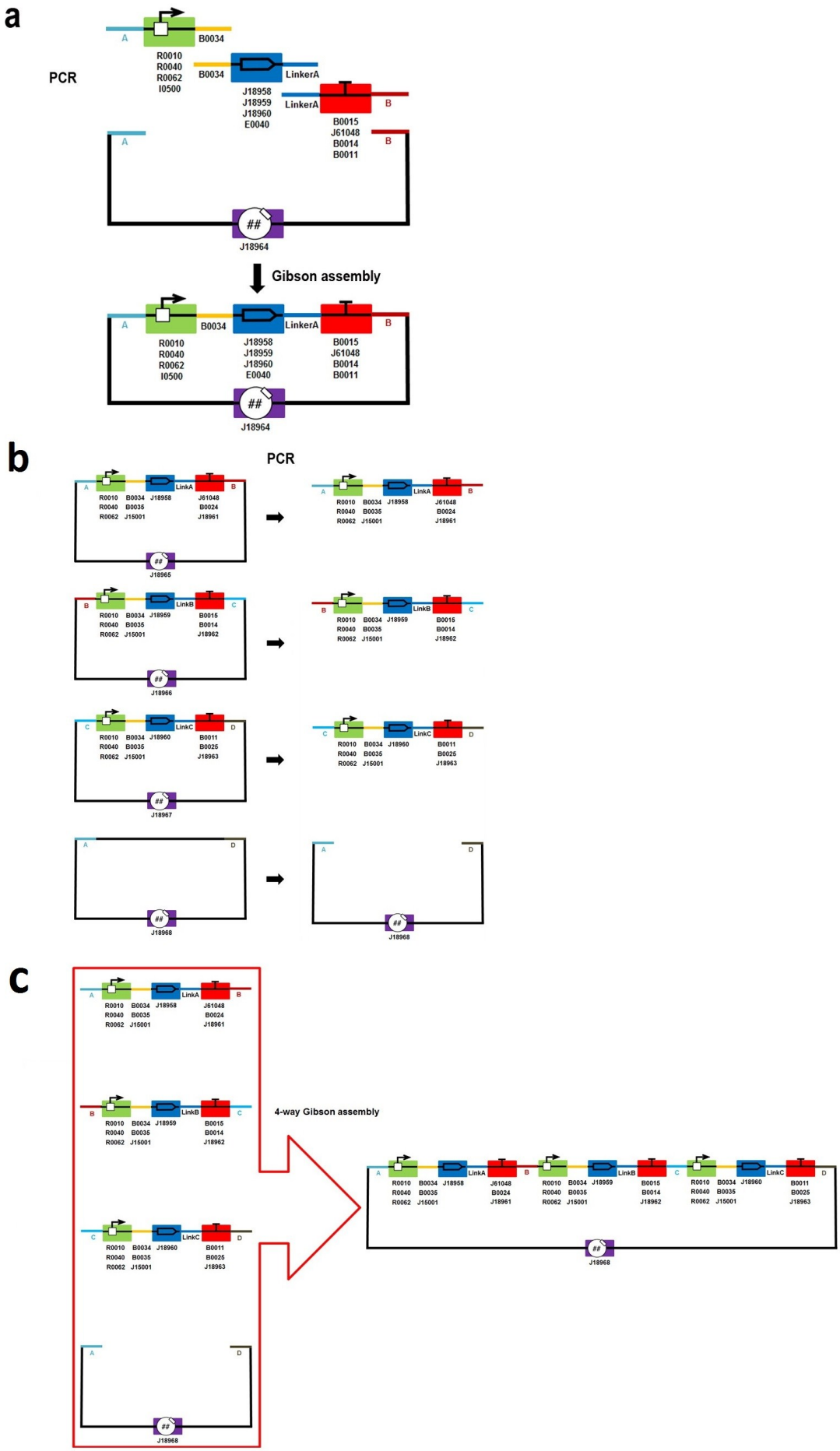


Figure 13: the mechanism of Golden Braid 1.0 standard[74]. (A) Standard parts such as promoters (PR), coding sequences (CDS) and terminators (TM) are flanked by fixed BsaI cleavage sites (represented as Arabic and Latin boxed numbers). They are assembled into transcription units (TU) using level α plasmids (pDGB A12C or pDGB C12B). This causes the BsaI recognition sites to disappear and the resulting boundary is not cleavable anymore (represented as a crossed box). The newly assembled transcriptional unit (TU1, now represented for simplification as an arrow) is flanked by BsmBI sites (represented as encircled capital letters). (B) Two TUs assembled in complementary α plasmids can be re-used as entry vectors (pEGB) for a subsequent level Ω binary assembly, as long as they share a BsmBI sticky end (marked as encircled C). Similarly, constructs assembled using opposite Ω plasmids can be re-used as entry vectors for a subsequent level α binary assembly, provided that they share a BsaI sticky end (marked as squared 3). Level α and level Ω can be alternated indefinitely to generate larger constructs, as shown by the grey arrows closing the double loop. Encircled K and S represent kanamycin resistance and spectinomycin resistance genes respectively.

1.3.2. Long overlap standards

There have been various attempts at modularising and standardising long overlap-based DNA assembly techniques, most of which were designed to use *E. coli* as host. Sleight *et al*.[76] designed a simple strategy that aims to conserve full compatibility with the BioBrick standard while replacing the restriction and ligation steps with a more efficient InFusion reaction (a commercial kit similar to SLIC). The system uses appropriate homology regions that generate constructs absolutely identical as if they had been built with the classic BioBrick assembly protocol, including prefix, suffix and scars. The subsequent version of their strategy[77] (**Figure 14**) uses Gibson isothermal assembly instead of InFusion and abandons the BioBrick format for a tiered system: sub-gene parts are assembled into genes using fixed homology regions added by PCR, so that parts of the same type (*e.g.* promoters, ORFs, terminators, *etc*.) are always flanked by the same sequences and can be interchanged freely. The genes are also built so that they are flanked by other homology regions that define their position in the final construct. The plasmids where the genes are assembled are then isolated, the gene regions amplified and finally assembled in the desired construct.

Figure 14 (next page): Sleight *et al.*[77] assembly diagram. (a) A PCR amplification is used to attach the appropriate homology regions to promoters (green), coding sequences (blue), terminators (red) and plasmid backbones (purple). These are then assembled using Gibson assembly to generate a plasmid containing the desired transcription unit. The homology regions at the flanks of the transcription units (A, B, C, D, *etc.*) define its position in the final multigene construct, the one between the promoter and the coding sequence encodes the RBS while the one called LinkerA, LinkerB, *etc.* simply joins the coding sequence and the terminator. (b) The previously assembled plasmids are used as template for a PCR that amplifies the transcription units together with the positional homology regions. A plasmid backbone with the appropriate homology regions is also prepared. (c) The transcription units are assembled together with a plasmid backbone using Gibson assembly.


a
PCR
A
B0034
R0010
R0040
R0062
I0500
B0034
LinkerA
J18958
J18959
J18960
E0040
LinkerA
B
B0015
J61048
B0014
B0011
A
B
##
J18964
Gibson assembly
A
B0034
LinkerA
B
R0010
R0040
R0062
I0500
J18958
J18959
J18960
E0040
B0015
J61048
B0014
B0011
##
J18964
b
PCR
A
LinkA
B
R0010
B0034
J18958
J61048
R0040
B0035
B0024
R0062
J15001
J18961
##
J18965
B
LinkB
C
R0010
B0034
J18959
B0015
R0040
B0035
B0014
R0062
J15001
J18962
##
J18966
C
LinkC
D
R0010
B0034
J18960
B0011
R0040
B0035
B0025
R0062
J15001
J18963
##
J18967
A
D
##
J18968
c
A
LinkA
B
R0010
B0034
J18958
J61048
R0040
B0035
B0024
R0062
J15001
J18961
B
LinkB
C
R0010
B0034
J18959
B0015
R0040
B0035
B0014
R0062
J15001
J18962
C
LinkC
D
R0010
B0034
J18960
B0011
R0040
B0035
B0025
R0062
J15001
J18963
A
D
##
J18968
4-way Gibson assembly
A
LinkA
B
R0010
B0034
J18958
J61048
R0040
B0035
B0024
R0062
J15001
J18961
B
LinkB
C
R0010
B0034
J18959
B0015
R0040
B0035
B0014
R0062
J15001
J18962
C
LinkC
D
R0010
B0034
J18960
B0011
R0040
B0035
B0025
R0062
J15001
J18963
##
J18968

Guye *et al.*[78] developed a similar tiered strategy (**Figure 15**) that aims to avoid two issues: the requirement for forbidden restriction sites and the chance of PCR-introduced mutations. In order to do so the assembly process starts from sequence-verified sub-gene parts cloned into Gateway vectors (a recombinase-based commercial assembly kit, see **Chapter 1.2.3**) that can be assembled in fixed positions to build gene-level parts using a recombination reaction. This process also places the genes in plasmids that carry the homology regions that define their position in the final construct, wherefrom they can be released with an I-SceI digestion (a homing endonuclease that recognises a 18 bp sequence, which has a very low likelihood to be encountered by random chance). Once released, the parts can be assembled into the final construct, together with a backbone (carrier vector) and an "adaptor", which joins the last part with the backbone and carries a second selection marker. Thanks to this double selection system it is possible to proceed to the final Gibson assembly reaction without removing the backbones of the original plasmids. The backbone can also contain homology regions and I-SceI sites again to generate an idempotent plasmid ready for further rounds of hierarchical assembly.

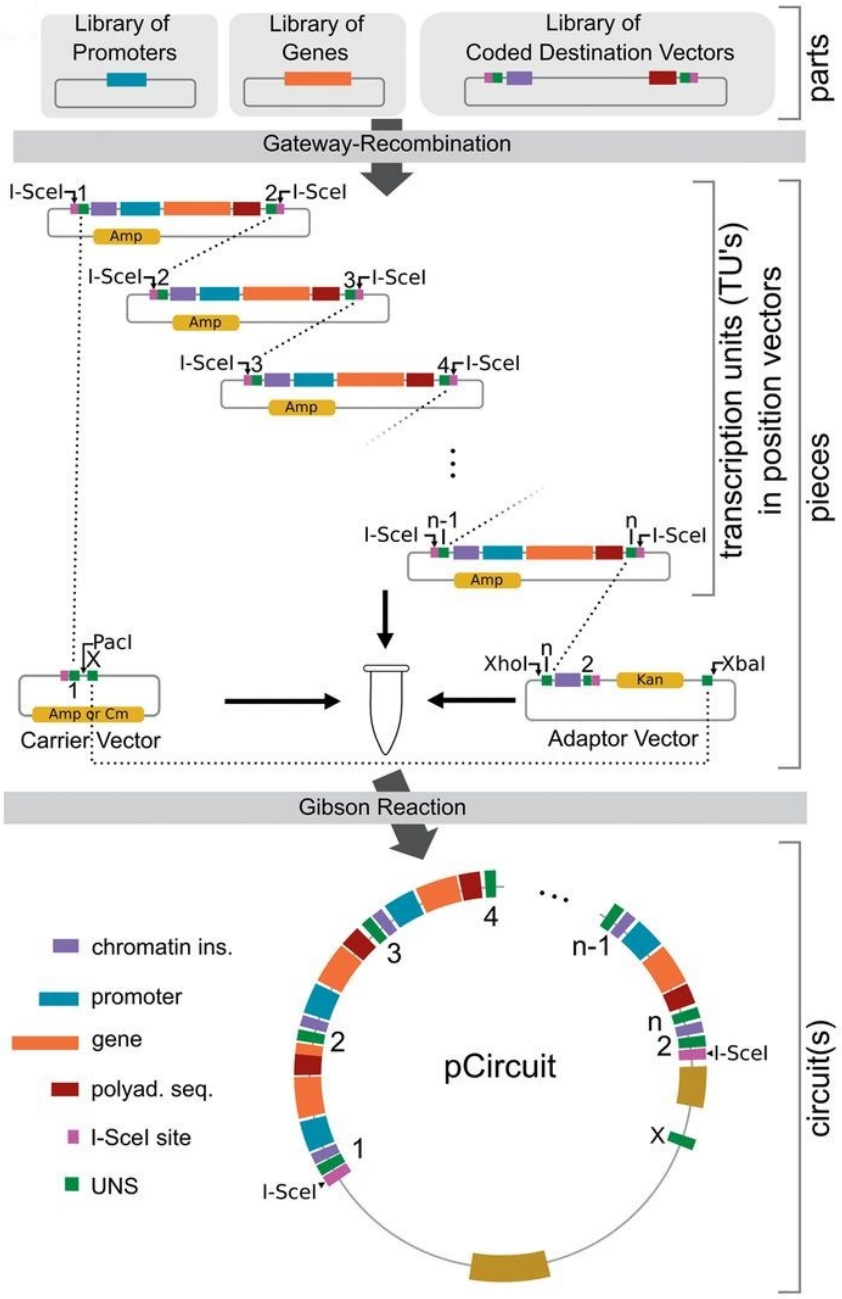


Figure 15: Guye *et al.*[78] assembly diagram. Parts such as promoters, genes and destination vectors containing a chromatine insulator and a polyadenilation sequence are assembled in fixed positions using Gateway recombination, forming complete transcription units. The destination vectors also contain computationally designed homology regions (UNS) that define the position of the transcription units in the final construct. The transcription units are then released from the plasmids together with their UNS using an I-SceI digestion, and are finally mixed in a Gibson isothermal reaction to build the final multigene circuit.

The strategy devised by Guye *et al.* is very powerful but also quite complex, requiring two assembly methods and a large number of different plasmids. Torella *et al.*[79] (**Figure 16**) adopted a simpler single-tier assembly strategy and developed highly optimised homology regions that can also act as insulator between expression cassettes. Initially the parts are cloned in plasmids that carry the appropriate homology regions using BioBrick or BglBrick cloning. These plasmids also carry a set of restriction sites that are used to release the part together with the homology regions, so that it can be purified via gel extraction and assembled with the other purified parts using a Gibson isothermal reaction. The last part of the construct is digested with different enzymes so that it is released together with a special terminal homology region, designed be joined with the backbone.

Modular assembly strategies have been developed for yeast recombination assembly too, but they are in general much simpler: DNA assembler[53] uses a tiered approach where OE-PCR is used to build expression cassettes designed to overlap with the other neighbouring cassettes by 40 bp. These fragments are then purified by gel extraction and transformed in yeast together with a plasmid backbone or a chromosomal integration helper fragment, where they are assembled thanks to the 40 bp overlap. The same authors successively expanded this work by defining a strategy to rapidly build plasmids for the expression of pathways in uncommon hosts[80]. These plasmids are designed to include backbone fragments for *S. cerevisiae*, *E. coli* and the desired expression host so that they can be assembled in yeast, isolated, transformed in *E. coli* for amplification, and finally isolated again in large quantities for transformation in the final host. Kuijpers *et al.*[56] found that this kind of assembly strategy in yeast can be improved by placing the origin of replication and the selection marker on two separate fragments to be assembled in non-consecutive position, which greatly reduces background colony formation. They also advocate the use of computationally designed 60 bp homology regions, which they obtained from an *S. cerevisiae* genome bar-coding project[81], to minimise undesired recombination events with other homology regions, with the internal sequences of the parts or with the genome.

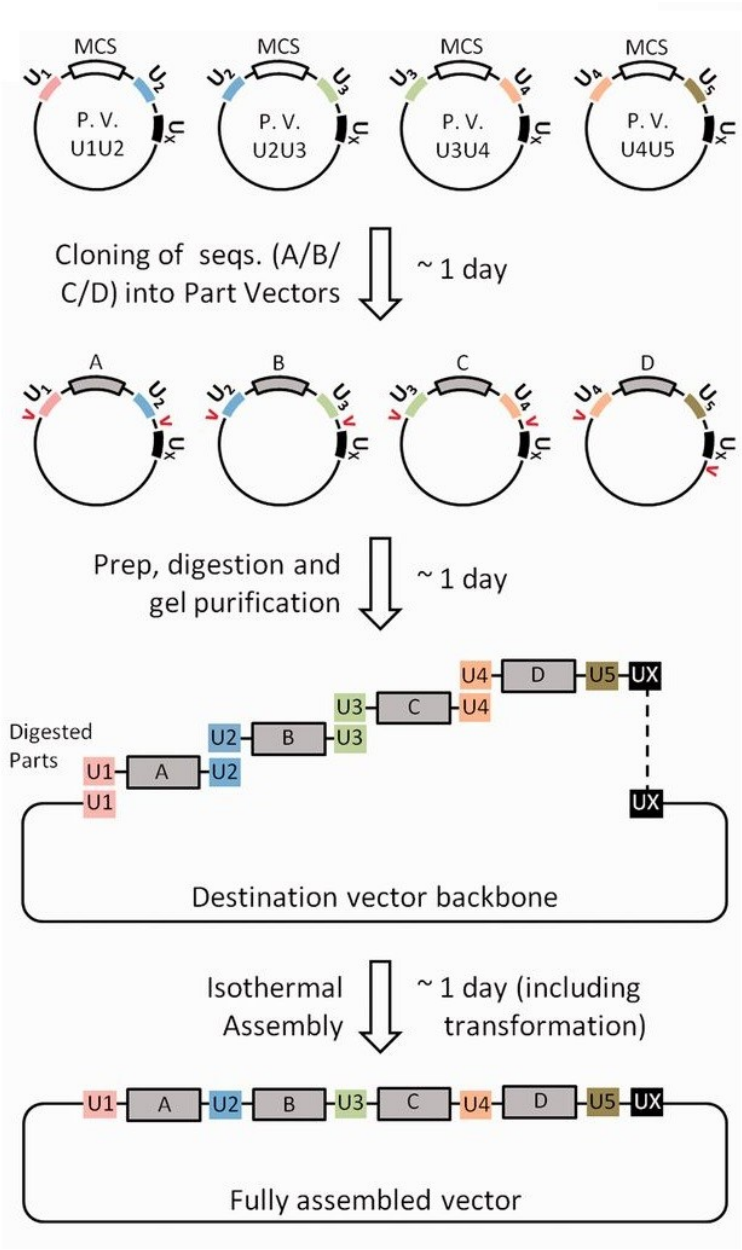


Figure 16: Torella *et al.*[79] assembly diagram. Example of a five-piece assembly composed of four part vectors (P.V.) and a destination vector. DNA parts are cloned in the part vectors using the BioBrick/BglBrick compatible multiple cloning site (MCS). Part vectors also contain the homology regions (U1, U2, …,Un, UX) that define the position of the parts in the final construct. These are digested to release the part flanked by the homology regions (Un), except for the last, which is digested with different enzymes to release it with the terminal homology regions (UX). Finally the parts flanked by the homology regions are purified via gel extraction and mixed in a Gibson isothermal assembly reaction to obtain the final construct.

1.3.3. Recombinase-mediated standards

Recombinase-based DNA assembly has been employed in a few modular DNA assembly strategies as well: Moriarity *et al*.[60] developed RecWay, a process specifically aimed at generating multigene cassettes ready for transposon-mediated integration in mammalian chromosomes. Even though the strategy does not require any forbidden restriction sites or PCR amplification steps, it is quite complex and slow, as it uses three different recombination systems: Gateway (a commercial kit that exploits *in vitro* lambda recombination), *in vitro* Cre and *in vivo* Flp. It also requires a digestion/ligation step and a large number of different pre-prepared plasmids. The whole process takes seven days and can only assemble up to six parts, even though the modularity of the system and its reliability can save some time over an *ad hoc* approach.

Some of the creators of RecWay were involved in the development of HomeRun[82], a simpler but still restriction enzyme and PCR-free DNA assembly strategy (**Figure 17**). Similarly to RecWay, the DNA parts are stored in Gateway plasmids, from which they can be assembled four by four in a new destination plasmid using an *in vitro* recombination reaction. These destination plasmids are equipped with homing endonuclease recognition sites that can be used to release the four parts as a single module, which can then be cloned in the pHR plasmid that carries a homologous homing endonuclease site. This is achieved by digesting both plasmids separately and isolating by gel extraction the module and the pHR fragments so that they can finally be mixed together and joined using T4 DNA ligase. The pHR plasmid contains four different homing endonuclease recognition sites, so this step can be repeated four times to clone four modules, equivalent to sixteen total parts. The developers anticipate that new homing endonucleases will soon become available, significantly expanding the power of this strategy.

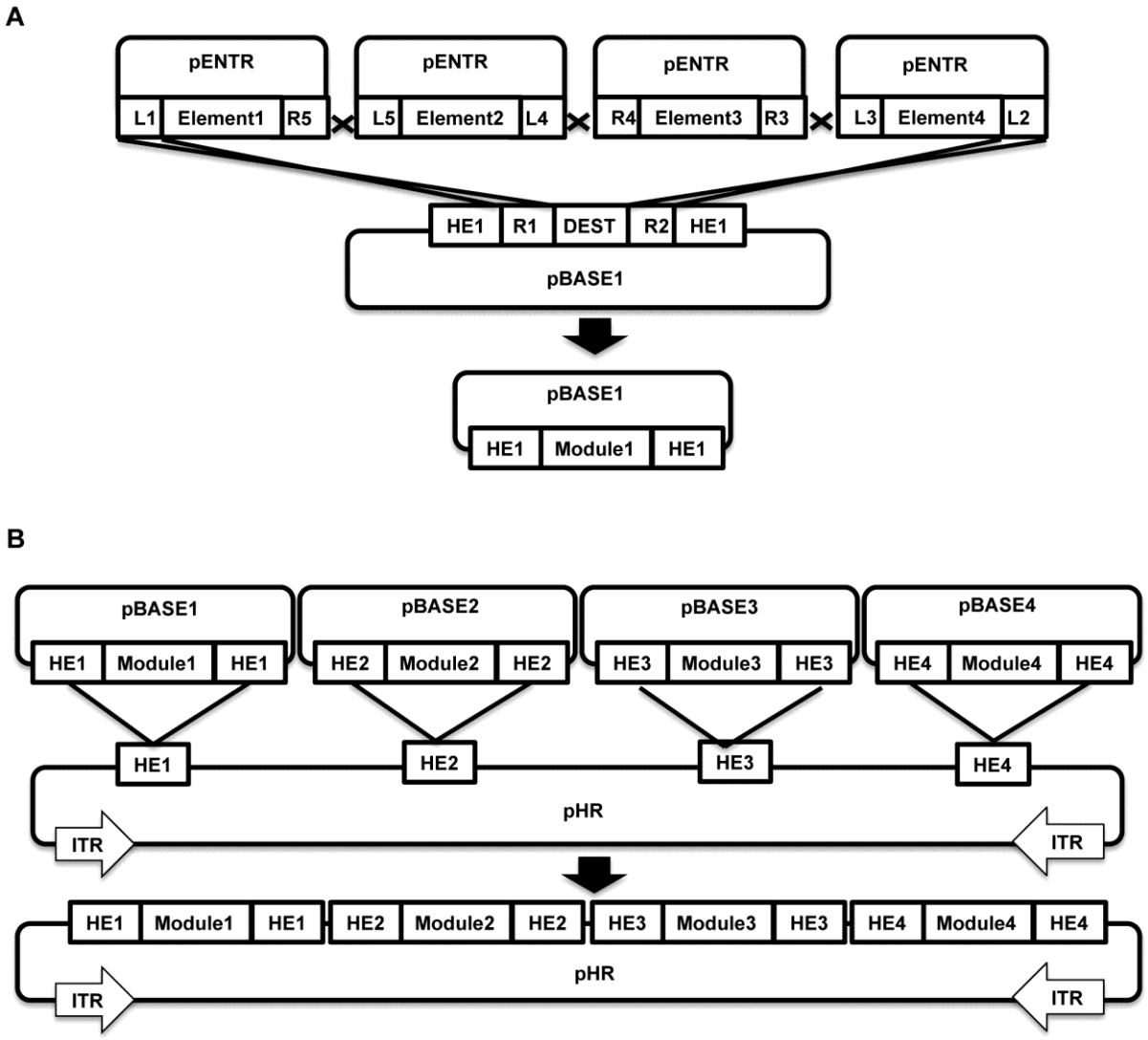


Figure 17: schematic of the HomeRun standard[82]. (A) Assembly of a functional module in pBASE vectors from sub-gene elements in pENTR vectors. Up to 4 elements in pENTR vectors can be assembled simultaneously in a pBASE vector. (B) Assembly of multi-modular construct in pHR assembly vectors from modules in pBASE shuttle vectors. Up to 4 modules in the pBASE vectors can be sequentially assembled into the pHR vector. L(number) and R(number) represent the attL(number) and attR(number) recombination sites respectively. HE(number) represent different homing endonuclease sites, DEST is the destination cassette to be replaced with an assembled module and ITR are inverted terminal repeat sequences for mammalian chromosomal integration.

1.3.4. Software tools

The adoption of standard modular frameworks surely simplifies the process of assembling DNA constructs, but manually planning all the experimental steps can still be quite difficult, especially for a very complex project. This can result in experimental plans that are more time-consuming, expensive and error-prone than they could be. In order to address this Appleton *et al.*[83] developed "Raven", an algorithm-driven software tool for DNA assembly planning that support the BioBrick and MoClo standards. It also supports PCR-based scarless cloning using Golden Gate, Gibson, CPEC and SLIC assembly, and by similarity it can likely be extended to other long overlap-based assembly methods such as yeast *in vivo* recombination. The user provides the software with the sequences of the initial parts, defines the desired final construct and chooses an assembly standard, and Raven produces an optimised experimental plan. This includes the PCR steps necessary to prepare parts for assembly, the relative oligonucleotides, and the assembly steps required to go from the parts to the final construct. The optimisation aims to minimise the PCR steps and the cloning steps by finding homology regions and intermediate constructs can be re-used as many times as possible. The developers demonstrate the usefulness of Raven by calculating optimised assembly plans for a number of well-known published constructs, and comparing them with the plans used by the original authors. The results showed that Raven's plans were an improvement over the original ones in almost every case. Additionally if a problem is encountered during the assembly process it is possible to take it into account when recalculating the experimental plan so that it does not appear again.

It is also worth mentioning that there is a similar but less advanced software tool, called J5[84], that focuses more on optimising the cost-effectiveness of the assembly process. For every assembly step J5 finds the best portion of the parts to be used as homology regions, and checks if any of the fragments to be assembled is small enough to be incorporated as a tail in a PCR primer instead, and also if it would rather be cheaper to have the construct synthesised by a company. In addition to using a less refined optimisation algorithm, this tool does not support any assembly standard, only

scarless assembly (with Golden Gate, Gibson, CPEC or SLIC). All considered, using a tool like J5 or Raven for projects that necessarily require a scarless approach will prove extremely helpful and will help regaining some of the advantages of using a modular DNA assembly standard, like for example the ability of reusing certain parts or intermediates.

1.5. Homology region design rules

Most DNA assembly methods rely on base pair recognition and annealing mechanics to join DNA molecules together, and the sequence of the regions involved has a strong influence on the efficiency and specificity of the process. This problem is encountered in many other molecular biology techniques, most notably PCR[85] and oligonucleotide microarrays[86,87]: it is widely understood and accepted that proper design of the oligonucleotides involved in these reactions is essential both for their efficiency (yield of the PCR amplification or signal strength of the microarray) and specificity (absence of undesired products in PCRs and of false positives in microarrays), and a number of software tools have been developed to help researchers generate optimised sequences[88,89]. This problem has not had the same attention in DNA assembly, for a number of reasons: early cloning methods used restriction enzymes that cut within their recognition sequence, generating sticky ends with a fixed sequence. Only later, when design practices were already consolidated, the development of new methods gave partial (e.g. USER) or complete (e.g. Gibson isothermal, Golden Gate) freedom of choice to the user. Additionally, the development of these sequence-independent methods sparked an interest in scarless assembly, which again removed the freedom to choose and optimise the sequence of the homology regions, since these had to exactly match the sequence of the parts being assembled (**Figure 18**).

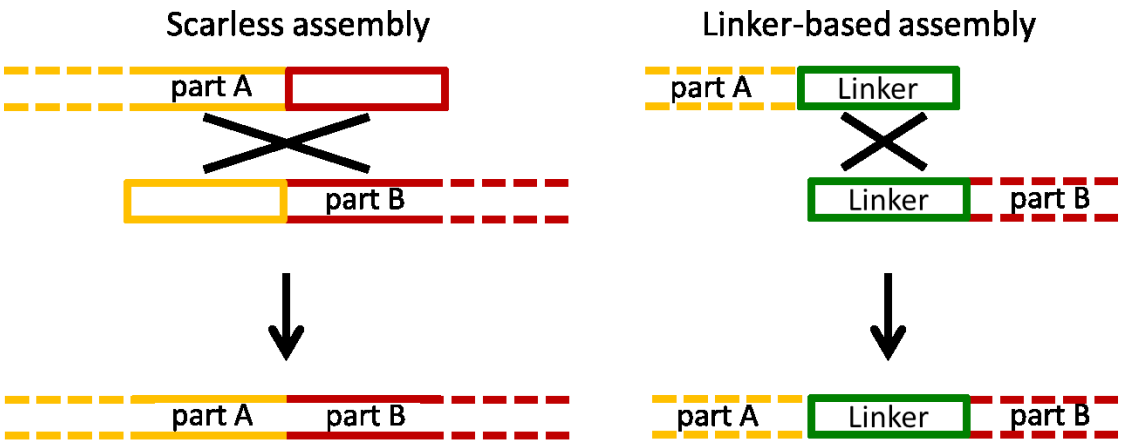


Figure 18: the diagram compares the assembly of two DNA parts (yellow and red) using a scarless approach and a linker-based approach. In the first case the homology region is generated by attaching a small portion of part B to part A and viceversa, usually via PCR, so that after assembly the two parts are joined directly to each other. In the second case a linker sequence (green) is attached to both parts and acts as homology region, so that after assembly the two parts are separated by it.

Nevertheless the idea of using specifically designed and optimised homology regions for DNA assembly has slowly gained popularity in synthetic biology, helped by the fact that these could be also used for a number of other things. The first is modularity: as shown in **Chapter 1.3** essentially every DNA assembly standard relies on standard homology regions that act as modular junctions in a lego-like fashion. The second advantage is that using externally added sequences solves the problem of mixing parts that are too similar to each other in the same assembly reaction: for example if three parts have the same promoter and terminator at their extremities, it is impossible to assemble them in a defined order using a scarless approach. Finally, if long enough, these regions can be useful beyond assembly by either acting as insulators[79] or by incorporating small functional regions such as RBSs, RNAse sites, *etc.*[77,90]

Type IIs restriction enzymes, which give the user the ability of choosing the sequence of the sticky ends they generate, have recently gained a lot of popularity, especially after the development of the highly efficient Golden Gate assembly method. This has brought some attention to how to design these sticky ends, and it has been shown that even though they are only 4 bp long their

performance can vary greatly: a study has found that GC content can account for performance differences as large as 30%[30] and it is not alone in recommending the use of GC-rich (50%-75%) sequences[31]. It has also been shown that, because of their structure, palindromic sticky ends are able to anneal to themselves, leading to non-specific ligation and reducing reaction yield up to 15-fold[91]. Palindromic sticky ends are produced by all traditional non-type IIs restriction enzymes, which is one of the reasons of the success of type IIs-based assembly methods, and many papers that employ Golden Gate assembly explicitly advise against using palindromic sticky ends[32,33]. When assembling many fragments simultaneously it is clearly essential to make sure these homology sequences are unique, but a study has found that even having 3 complementary bases out of 4 is enough to cause non-specific ligation[29]. A software tool, named NP-Sticky, has been developed to assist researchers in designing appropriate sticky end sequences for their ligations. It uses a thermodynamic model that is able to predict the yield and identity of all the possible products of a given ligation reaction, taking into account both correct and mismatched annealing. NP-Sticky can optimise reaction conditions to maximise yield of the desired product, to minimise yield of unspecific ligations, *etc.*[91]

The same holds true for methods that rely on long overlap regions: assembly reactions that rely on mechanisms similar to PCR clearly benefit from similar optimisation strategies: for example CPEC's authors explicitly state that it is important to design homology regions to have very high (60-70 °C) melting temperatures[1]. Cha-aim *et al*.[36] published an extensive study aimed at finding the best possible overlap sequences for OE-PCR, looking at different sequence lengths, the use of long G and/or C stretches to maximise annealing strength*, etc*.

PCR-like reactions are not the only ones to benefit from overlap optimisation: bridging oligonucleotides-based assembly method LCR was subject to a meticulous optimisation process, and the T_m of the bridging oligonucleotides was found to be one of the most influential parameters[3]. The most work in this direction has been done on Gibson isothermal assembly: both DNA assembly

strategies devised by Guye *et al*.[78] and Torella *et al*.[79] employ this technique, and both include a computational algorithm to generate optimised homology regions. Even though they did not perform any experimental analysis or comparison of the benefits of this optimisation, they both assume that these are significant, probably because of the previously mentioned parallel with other molecular biology techniques. Briefly, their optimisation processes are quite similar and aim at ensuring that the homology regions generated are within certain GC content and T_m ranges, that they can anneal to their target with high specificity, and that they do not contain undesired features such as secondary structures, certain restriction sites, biologically active motifs, *etc.* (discussed in full in **Chapter 2.1**).

1.6. Aims and motivations

DNA assembly is one of the fundamental enabling technologies for synthetic biology, and is currently also one of the main bottlenecks: it is time-consuming, expensive and requires a high level of craftsmanship. In a landscape that is continuously blooming with new methods and standards this work aims first of all to develop tools, standards, methods and practices for DNA assembly that are of general interest and can be applied across the field. The Linker software presented in Chapter 2 can help designing and evaluating homology regions for a wide range of reactions and the MODAL DNA assembly strategies provides a modular framework for some of the most popular DNA assembly methods. Finally, we propose our own novel method and standard, the Biopart Assembly Standard for Idempotent Cloning (BASIC). We believe that this is a significant contribution to the current landscape because it focuses on an aspect that has been up to now rather undervalued: reliability. The final goal is to make DNA assembly less dependent on circumstances and personal skill, and more like a mature technology that every synthetic biologist can benefit from.

2. Linker: a software tool for the computational design of DNA linker sequences

Aims:

- Developing a software tool that generates short DNA sequences specifically optimised to be used as linker regions for DNA assembly reactions.
- Laying the foundations for a more advanced version of the software, to be made publicly available for the synthetic biology community.

2.1. Introduction

The efficiency and specificity of DNA assembly reactions is known to be significantly influenced by the features of the sequences involved, as discussed in **Chapter 1.5**, but only very recently two publications, by Guye *et al.*[78] and Torella *et al.*[79], presented their own software tools to generate optimised homology regions for DNA assembly (also called "linkers"). While all long overlap-based DNA assembly techniques can benefit by linker optimisation, this is particularly true for Gibson isothermal assembly, which is employed by both publications. Gibson's method is very powerful but also exposed to various issues, as shown in **Figure 19**, due to the fact that it generates single stranded regions of undefined length. In addition to this, the reaction employs an *in vitro* DNA repair mechanism that can help in consolidating partially mismatched annealing, introducing mutations and allowing non-specific assembly.

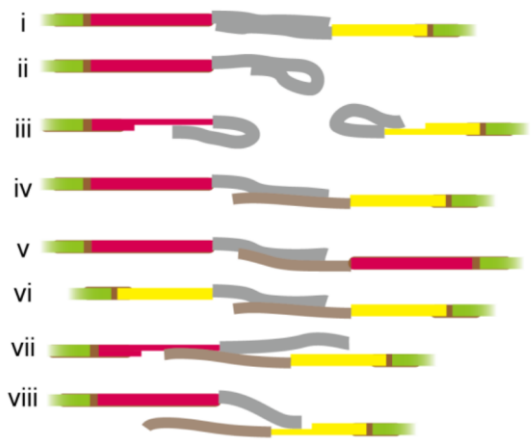


Figure 19: possible incorrect annealing modes in linker-based assembly using Gibson isothermal or other reactions that employ indefinite chew-back. In grey are the homology regions, yellow and red represent the portions of the parts to be assembled that might be made single stranded and thus be available for annealing, green are the rest of the DNA parts that remain double stranded. (i) Correct pairing between cognate linkers, (ii) duplex formation within linkers or (iii) between linkers and sequences in close proximity exposed during the assembly reaction, (iv) pairing between non-cognate linkers, (v, vi) pairing between different linkers on the same DNA part, (vii, viii) pairing between linkers and sequence in close proximity exposed during assembly reaction. From Guye *et al.*[78]

Both software tools mentioned above work similarly: they initially generate a pool of random sequences and then they refine it according to a list of rules. The algorithm developed by Guye *et al*. aims to minimise the probability of undesired annealing (cases ii to viii in **Figure 19**) while keeping a number of parameters within pre-set constraints: strong secondary structures and certain restriction sites must not be present, T_m must be between 65°C and 75°C, GC content between 40% and 80%. The terminal 7 bp on both sides of the homology regions are subject to stricter constraints for what concerns undesired annealing because they can act as seeding region for the misannealing of the rest. This is a very comprehensive rule-set for undesired annealing, but it lacks other important details. The GC content and T_m rules are very relaxed (maybe too much, as the data in **Chapter 3** suggests), and it does not include any rules to eliminate functional motifs which might influence the behaviour of the surrounding sequences.

The algorithm developed by Torella *et al*. is very similar, as it essentially tries to achieve the same goal, but uses a slightly different set of rules. Their criteria for undesired annealing are much less refined, as they only specify that sequences must not be able to anneal strongly to each other, but on the other hand the other the parameters for GC content are stricter: only values between 45% and 55% are allowed, no continuous stretches of AT-only or GC-only longer than 4 bp and there must be one or two Gs or Cs at the extremities of every homology region (to "seal" the annealing region, a common practice in PCR primer design). The T_m is not checked, possibly as it is essentially tied to the strictly controlled GC content. The algorithm also includes a number of criteria aimed at eliminating sequences that might have biological activity: it employs two external tools to check for the presence of bacterial promoters, ensures the absence of any start codons and performs a BLAST search of the generated sequences against the host's genome (*E. coli* MG1655), removing any strong matches. Finally, similarly to Guye *et al*.'s algorithm, it checks for strong secondary structures and undesired restriction sites.

The software tools described in this chapter were developed while this work was still in progress, as the need for such tools was evidently felt by many groups in the synthetic biology community at the same time. As explained in **Chapter 1.5** the use of computational tools to design optimised synthetic sequences for biochemical reactions such as PCRs and oligonucleotide microarrays has been a consolidated practice for years. This gave us, and most likely the other groups as well, the motivation to develop similar tools for DNA assembly too.

2.2. Results

2.2.1. Software overview

"Linker" is a software tool developed as a MATLAB script that randomly generates DNA sequences according to a set of user-defined rules. The user enters an initial string of characters of any desired length including only "A", "T", "C", "G" and "N" as characters. The software will randomly replace all the "N" characters in the string with "A", "T", "C" or "G", while leaving the others as they are. It will then analyse the generated sequence according to the defined rules and score it accordingly, with higher scores meaning worse sequences. It will then run a "mutation" cycle (**Figure 20**), in which it randomly picks one of the characters that was an "N" initially, replaces it with one of the four bases and scores the sequence again. If the score of this sequence is better than the previous one, it will proceed with another "mutation" cycle. If it is worse, it will run a number of "probing" sub-cycles: this was implemented in order to deal with situations in which a better sequence cannot be generated by replacing a single base only, but it needs multiple changes. For each "probing" sub-cycle the script will cumulatively change a base in the "bad" sequence, keep all these sequences and score them. It will then check if any of these variations of the "bad" sequence is better than the previous sequence. If so it will use that for the next "mutation" cycle, otherwise it will use the previous sequence again. Both the number of "mutation" cycles and of "probing" sub-cycles can be defined by the user, and the process stops when the software has run all the "mutation" cycles. The sequences can be of any length, but very short (about 10bp or less) and very long (about 100 bp or more) ones might not return any useful (low scoring) results: for short sequences it is more difficult for the software to find solutions that satisfy all the rules, while for long ones the processing time on a normal desktop computer might increase too much to be practical. The user can also select the probability for each base for appear by changing the three values that define the probability brackets: the script's "rand" function randomly generates a number between 0 and 1, and an "A" is selected if that is between 0 and the first of the three probability brackets values, "T" if it is between the first and the second value, "C" if it is between the second and the

third value and “G” if it is between the third and 1. The default values “0.25, 0.5, 0.75” confer equal probability to all four bases.

Parameter name	Content	Default value
Input sequence	A string of characters of any length (typically between 10 and 100) containing any of the following characters: A, C, G, T, N.	-
Other sequences	Accepts an optional list of DNA sequences of the same length of the input sequence.	-
Bases probability brackets	Accepts three values between 0 and 1 that determine the probability brackets used by the random base selector function.	0.25, 0.5, 0.75
Mutation cycles	Number of mutation cycles that are performed by the script.	500
Probing subcycles	Number of probing cycles that are performed by the script.	5
Target type	Accepts "1" and "2" as values. "1" means the script targets a GC% value, "2" a T_m value.	1
GC% target	A number between 0 and 1.	0.5
T_m target	Accepts any integer, but only values that make sense as T_m will result in useful sequences being generated	50
Forbidden sequences	Accepts an array of DNA sequences of any length.	See **Table 2**
SC threshold	Minimum number of consecutive annealing bases that gives an increase in the score for "Self" annealing.	6
SO threshold	Minimum percentage of overall annealing bases in the whole sequence that gives an increase in the score for "Self" annealing.	0.3
CC threshold	Minimum number of consecutive annealing bases that gives an increase in the score for "Complementary" annealing.	6
CO threshold	Minimum percentage of overall annealing bases in the whole sequence that gives an increase in the score for "Complementary" annealing.	0.3
OC threshold	Minimum number of consecutive annealing bases that gives an increase in the score for "Others" annealing.	6
OO threshold	Minimum percentage of overall annealing bases in the whole sequence that gives an increase in the score for "Others" annealing.	0.3
GC% target weight	Weight multiplier associated with the distance from the GC target value.	1
T_m target weight	Weight multiplier associated with the distance from the T_m target value.	1
Forbidden sequences weight	Weight multiplier associated with the presence of forbidden sequences.	11
SC threshold weight	Weight multiplier associated with consecutive annealing to "Self" beyond the threshold.	1
SO threshold weight	Weight multiplier associated with overall annealing to "Self" beyond the threshold.	1
CC threshold weight	Weight multiplier associated with the consecutive annealing to "Complementary" beyond the threshold.	1
CO threshold weight	Weight multiplier associated with overall annealing to "Complementary" beyond the threshold.	1
OC threshold weight	Weight multiplier associated with the consecutive annealing to "Others" beyond the threshold.	11
OO threshold weight	Weight multiplier associated with overall annealing to "Others" beyond the threshold.	11

Table 1: list of parameters that the user can customise within the Linker script. The default values shown are set to generate linker DNA sequences optimised for Gibson assembly in E.coli. Assuming an arbitrary threshold of 10 or less for the final score of a "good" sequence, these setting prioritise the absence of forbidden sequences and orthogonality to the sequences in the "Others" list.

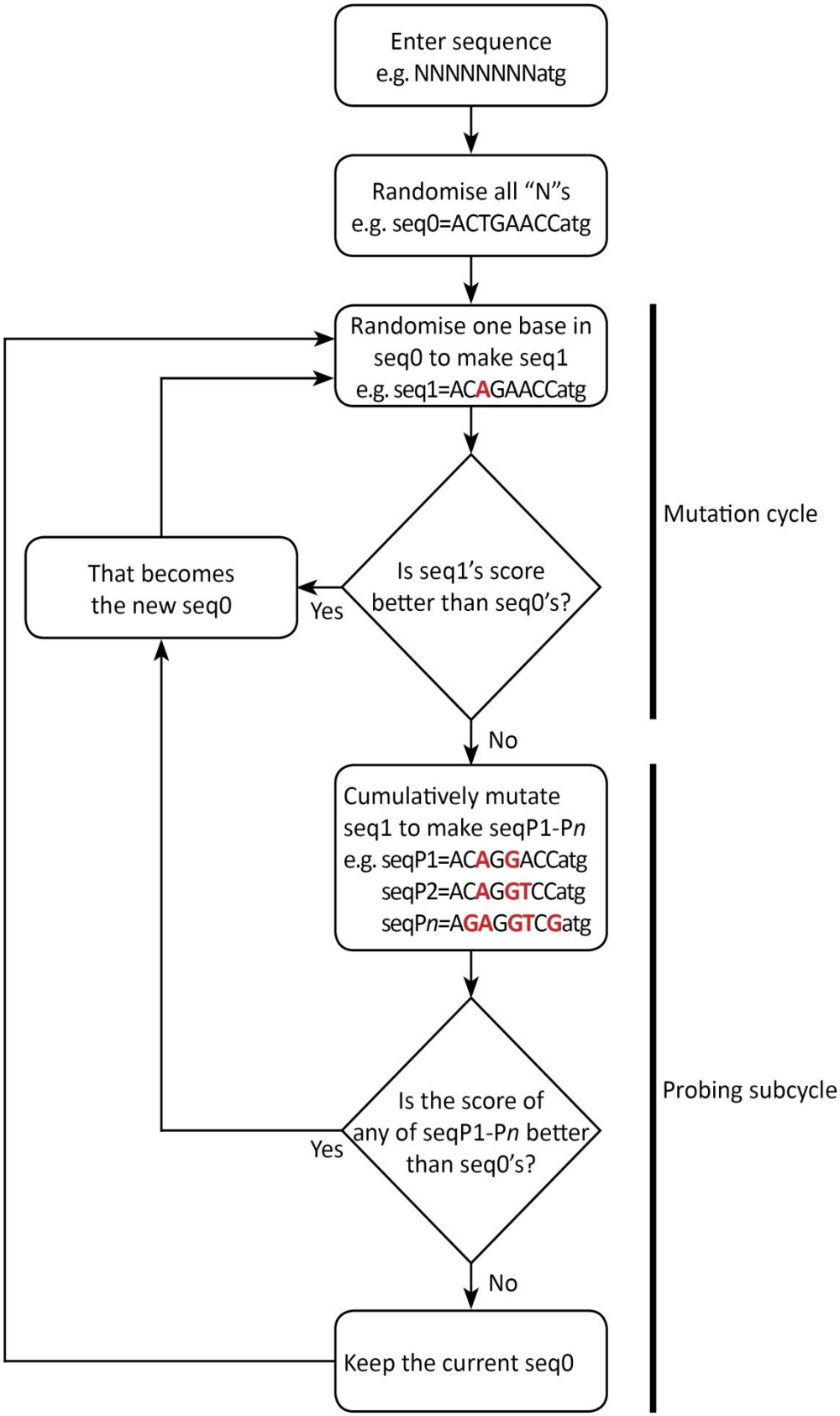

Enter sequence
e.g. NNNNNNNNatg
Randomise all "N"s
e.g. seq0=ACTGAACCatg
Randomise one base in seq0 to make seq1
e.g. seq1=ACAGAACCatg
Is seq1's score better than seq0's?
Yes
That becomes the new seq0
No
Mutation cycle
Cumulatively mutate seq1 to make seqP1-Pn
e.g. seqP1=ACAGGACCatg
seqP2=ACAGGTCCatg
seqPn=AGAGGTCGatg
Is the score of any of seqP1-Pn better than seq0's?
Yes
No
Keep the current seq0
Probing subcycle

Figure 20: the Linker script's cycling process: the user enters an input sequence where all the "N" characters will be randomised and optimised by the script, while the "a, c, t, g" characters will remain unchanged. Before the actual cycling starts all the "N" characters are randomised at the same time, and the resulting sequence is named seq0. The mutation cycle starts with the creation of a seq1 sequence, by randomly changing one of the bases of seq0 that was originally an "N" in the input sequence: if the score of this newly created seq1 is lower than that of seq0, seq1 becomes the new seq0 and another mutation cycle is performed. If the score is higher, then a probing subcycle starts: a number "*n*" of sequences are created by cumulatively randomising bases in seq1 (only those that were "N" in the input sequence). For example, seqP1 is seq1 with one randomised base, seqP2 is seqP1 with an additional randomised base, and so on. When all the seqPx sequences have been generated, the script checks if the score of any of them is lower than seq0's. If so, that sequence becomes the new seq0, and a new mutation cycle starts. If not, a new mutation cycle is performed again on the current seq0. This continues until all the mutation cycles defined by the user have been performed.

At the end of the process the script returns the sequence of the best linker generated, together with its score, GC content percentage and melting temperature. It also creates a plot (**Figure 21**) of the scores of the sequences at each mutation cycle (or the best out of the probing subcycle if it was run during that mutation cycle). This allows the user to have a quick visual report of how well the sequence generation process performed. **Figure 21** shows a typical successful case: the script found better sequences very fast at first and more slowly during the last cycles, until it stabilised around a value. The spikes are caused by unsuccessful probing cycles that only find "dead ends". The plot can also be useful for troubleshooting purposes: for example if the plot only shows constant values interrupted by spikes, right from the first cycles, it means that the script cannot find any good results and it might be beneficial to relax the rules. If the plot shows a marked downward trend that does not slow down during the latest cycles instead it might be beneficial to increase the number of cycles that are performed.

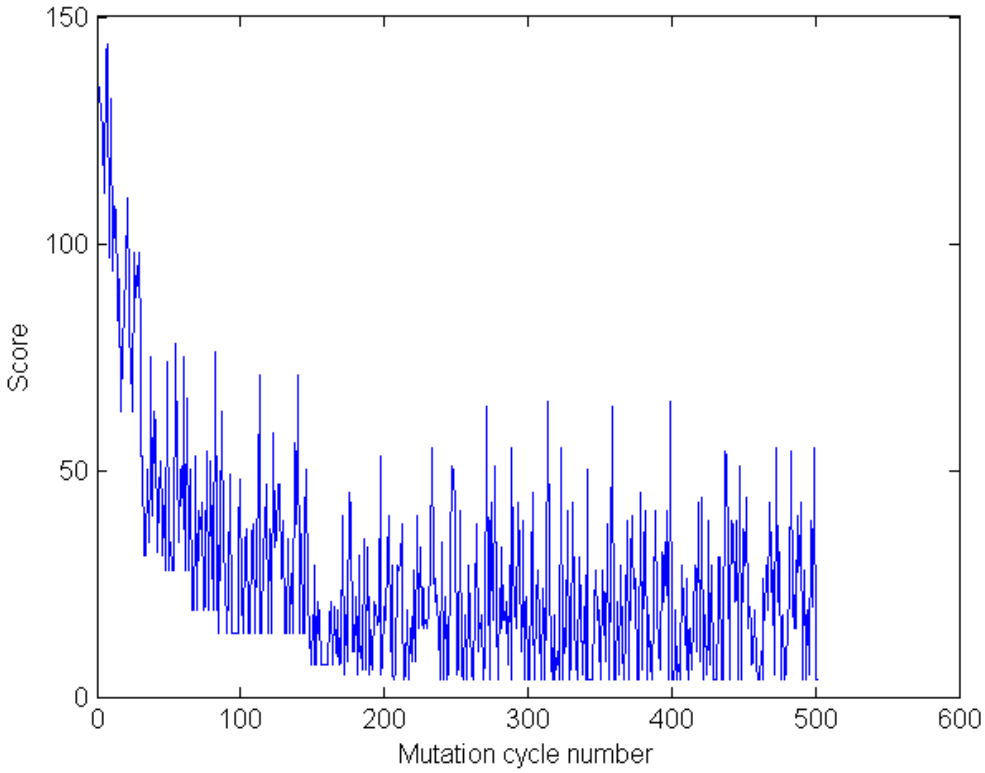


Figure 21: an example of the plot that is produced by the script at the end of the process. For each mutation cycle it shows the score of the best sequence it has generated, either from the mutation cycle itself or the best one from the probing cycle. The spikes appear when both the mutation cycle and the probing cycle fail to find a better sequence, and the best value from the probing cycle is shown, which can be significantly worse than seq0's.

2.2.2. Scoring algorithm

The scoring of the sequences is based on a set of rules. Each rule contains one or more parameters that the user can customise, and is associated with a "weight" value. This allows the user to tune how important that rule is for determining the overall score of the sequence.

GC content (GC%) and melting temperature (T_m): The user can select which of the two to consider when scoring the sequence and enter a target value for it (the T_m is calculated as in Santalucia[92]. The script calculates the difference between the user's target value (tVal) and the sequence's actual value (aVal), and multiplies that by the weight (**Table 1**):

$$target\ value\ score = \sqrt{(aVal - tVal)^2} * valWeight$$

Forbidden sequences: The user can enter a list of forbidden sequences whose presence will negatively impact the scoring of the sequence. The list below (**Table 2**) was used in this version of the software, which includes a few generic undesirable sequences for linkers to be used in *E. coli*. The script counts the number of times each forbidden sequence occurs in the generated sequence (nForb) and multiplies that by the user-defined weight.

$$forbidden\ sequences\ score = nForb * forbWeight$$

Sequences	Function
ACTAGT	SpeI recognition site
TCTAGA	XbaI recognition site
CTGCAG	PstI recognition site
GGATG	FokI recognition site
GCGGCCGC	NotI recognition site
GGTCTC	BsaI recognition site
AGGAGG, CCTCCT	Shine-Dalgarno sequence and its complementary sequence
ATG, CAT	Translation initiation site and its complementary sequence
CTAG, CTAA, CAAG	IS5 insertion sites
ATATAT, ACACAC, AGAGAG, TATATA, TCTCTC, TGTGTG, CACACA, CTCTCT, CGCGCG, GAGAGA, GTGTGT, GCGCGC	Short repeated sequences

Table 2: the default list of forbidden sequences in the Linker script.

Undesired annealing: The script checks if the generated sequence is able to anneal to a variety of undesired targets which fall under three categories. Category “Self” includes annealing of the linker oligonucleotide molecule to itself (causing secondary structures) or to other identical molecules present in the solution. These are undesirable because they can make the linkers unable to find their correct target during the DNA assembly reaction. Category “Complementary” analyses the ways in which a linker molecule can anneal to another molecule with a reverse-complementary sequence. They are meant to interact by completely and perfectly annealing to one another, but they may also be able to partially anneal in a “shifted” fashion, which could lower the efficiency of the DNA assembly reaction or lead to mutations in the final construct. Category “Others” is optional: the user can enter a list of DNA sequences of the same length of the one being generated. The script checks that the new sequence is orthogonal to the ones in the list, by not being able to anneal to them. The user can decide what constitutes excessive undesired annealing by tuning two thresholds (independently for each of the three categories): “Consecutive Annealing” sets the limit of consecutive bases that are able to anneal, while “Overall Annealing” is the percentage of bases that anneal to each other along the whole sequence, regardless of their position. These two different thresholds are meant to take into account the different ways in which two oligonucleotides can interact with each other, both with strong localised annealing (which within the same molecule can cause secondary structures) or with dispersed but frequent annealing which can keep the two molecules together even if there are mismatches all along.

The script performs these calculations using the same algorithm: it converts the generated sequence and all the other ones to strings of numbers (where A=1, C=2, G=3, T=4), and creates a Toeplitz matrix from the generated sequence. All the sequences that need to be checked against the generated sequence (“Self, “Complementary” and “Others”) are then converted to their complementary so that positions that anneal to each other are now represented by identical bases, and converted to reversed columns, to match the structure of the Toeplitz matrix. The “bsxfun” function is used to compare these columns to all the columns in the Toeplitz matrix, and the

identical bases are counted. This is essentially equivalent to sliding the sequences along each other, aligning them in every possible position to check how many bases can anneal in each case. The script then counts the number of alignment positions where a consecutive stretch of annealing bases that is longer than the "Consecutive Annealing" threshold appears, and the number of alignment positions where the overall number of annealing bases is beyond the "Overall Annealing" threshold. The script thus generates six values: "sc" ("Self", "Consecutive Annealing"), "so" ("Self", "Overall Annealing"), "cc" ("Complementary", "Consecutive Annealing"), "co" ("Complementary", "Overall Annealing"), "oc" ("Others", "Consecutive Annealing"), "oo" ("Others", "Overall Annealing"). These are finally multiplied by their specific weight and added up to calculate the "undesired annealing" score.

$$undesired\ annealing\ score = \\ = sc * scWeight + so * soWeight + cc * ccWeight + co * coWeight + oc \\ * ocWeight + oo * ooWeight$$

2.3. Discussion

2.3.1. The usefulness of the Linker script

The script can successfully find "useful" solutions when given the typical constraints required for linker regions, which corresponds to sequences with a score below 10. This threshold is arbitrary and depends mainly on the weights set by the user: our priorities for this study were the absence of forbidden sequences and the orthogonality towards the "Other" sequences, so we set the correspondent weights to 11: the presence of any of those would immediately set score of the whole sequence above our "usefulness" threshold. The running time is usually about a few minutes, depending on the strictness of the rules and the number of cycles. The plot that the script generates at the end of each run makes it very easy for the user to troubleshoot the process, to decide whether the constraints are well calibrated, if the number of cycles is sufficient, etc. The most common causes of problems are user-entered fixed bases in the input sequence or the inclusion of very short sequences in the forbidden sequences list, as these might make it very difficult for the script to find useful solutions. In order to confirm that the sequences generated by Linker were behaving as expected we tested some of them experimentally: we generated four orthogonal 45 bp long sequences, ordered them as oligonucleotides and used them as linkers in four different Gibson isothermal assembly reactions (see **Chapter 3.3.2**): all assembly reactions were successful and the constructs were fully functional. In addition to this the data (**Figure 30**) seems to suggest that the linkers have no influence on the expression of the genes around them, making them also viable as spacer regions between expression cassettes in *E. coli* plasmids.

2.3.2. The development of "R2oDNA Designer"

Given the success of the Linker script, we decided to take the ideas behind it forward and develop them into a more advanced software tool with graphical user interface (GUI) that is available via a website. We therefore developed "R2oDNA Designer", as published in "R2oDNA Designer: Computational Design of Biologically Neutral Synthetic DNA Sequences", by Casini *et al.*[93] (**Figure 22**). This software was largely coded by James MacDonald with development advice from the rest of the team. The author contributed to the design of the software and performed the experimental testing of the generated sequences (see **Chapter 3.3.1**).

The scope of R2oDNR2o Designer was expanded compared to the Linker script: the latter was developed exclusively to generate DNA sequences to be used as linkers for DNA assembly reactions, while the new software was designed to generate biologically neutral, non-functional DNA sequences in general, including not only DNA assembly linkers but also spacers to insulate functional DNA regions, barcode sequences, negative controls for functional regions, *etc*. It also has a "reverse mode", where the user can enter a defined DNA sequence, and the software processes it as it would any generated sequence, assigning it a score that can be used to evaluate how well it complies with the given rules.

R2oDNA Designer's usability is significantly improved compared to the Linker script: it is an easily accessible web-based software tool written in Java (available at http://www.r2odna.com/). The user can set all the parameters both manually through the GUI or by uploading a settings file, it generates any user-defined number of orthogonal sequences in a single run, the jobs are run on a cluster to increase execution speed, and the results are mailed to the user when execution is complete.

It also incorporates a number of improvements for what concerns the sequence generation process: it uses a powerful Monte Carlo Simulated Annealing (MCSA) algorithm to randomly generate sequences according to the rules, the Pairfold software to check for secondary structures[94], a network elimination algorithm to make sure all the generated sequences are orthogonal to each

other[87] and BLASTN to check the sequences against a number of user-defined targets, such as the host's genome, the BioBrick library *etc*. This helps ensuring that the sequences are as biologically neutral as possible.

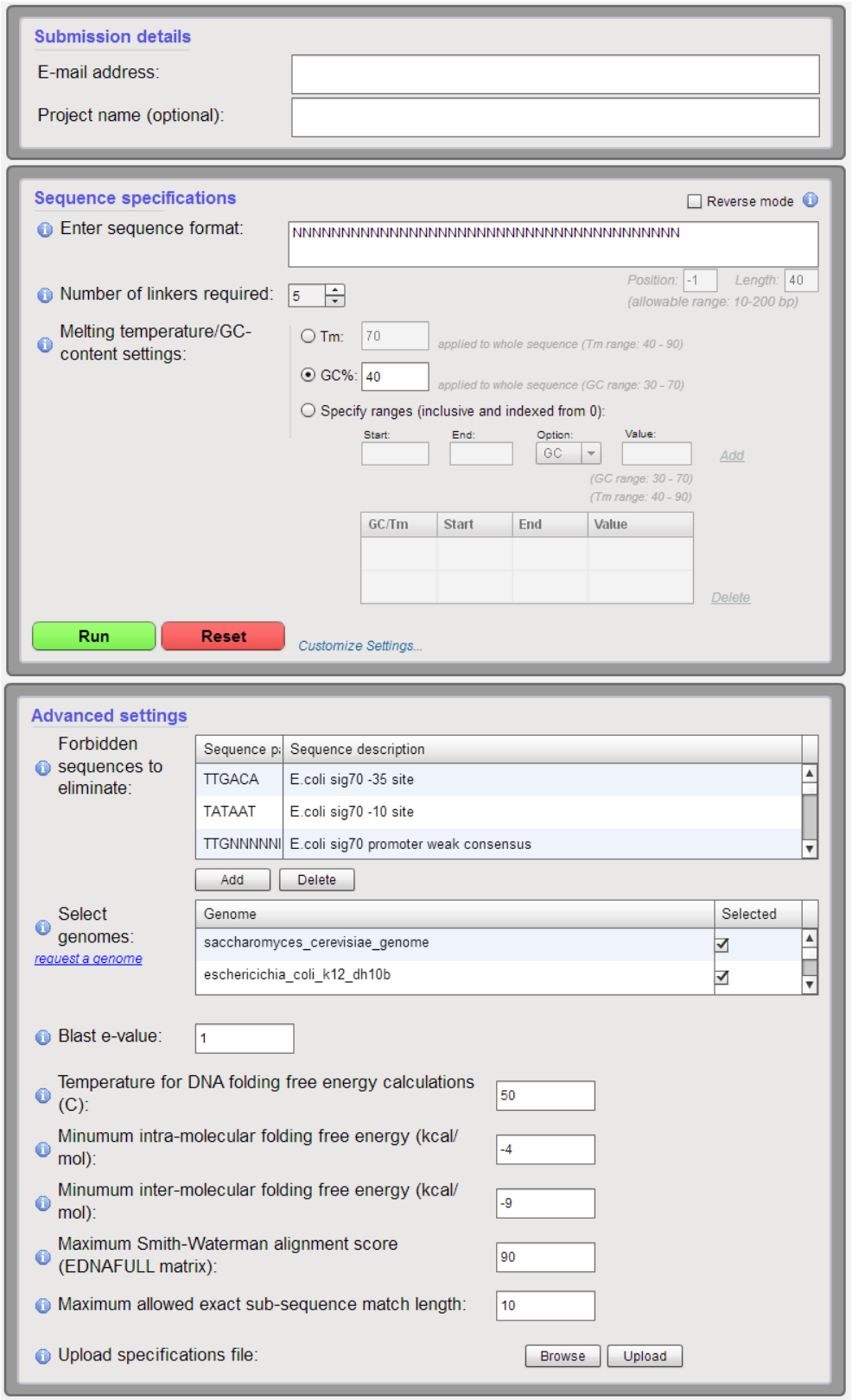


Figure 22: A screenshot of the R2oDNA Designer online software tool, available at http://www.r2odna.com/

2.3.3. Conclusion

We developed Linker, a MATLAB script that generates short DNA sequences and optimises them to comply with a number of customisable constraints. It is based on a cycling system where a randomly generated sequence is iteratively checked against the constraints and then mutated in order to improve it. The cycling system also includes a mechanism to escape local minima. We developed this software to generate sequences to be used as linker regions in DNA assembly reactions, and the script can successfully find solutions when set with the typical constraints for that purpose. A few of these sequences were tested experimentally and confirmed to work as expected. The Linker script was also the starting point for the development of R2oDNA Designer[93], a fully-fledged online software tool for the design of biologically neutral DNA sequences in general, including DNA assembly linkers, non-functional spacers, DNA barcodes, *etc*.

Optimising the sequence of the DNA molecules involved in biochemical reactions is known to be crucial for the efficiency and specificity of the reactions. One of the aims of this project is applying this design principle to DNA assembly reactions, and the development of a software tool to generate these optimised DNA sequences is the logical first step. The results presented in the following chapters of this work confirm this hypothesis and show that a great benefit can be obtained using optimised sequences generated with our software tools across a variety of DNA assembly reactions.

2.3.4. Future work

Further development of R2oDNA Designer should first of all aim to improve its usability and flexibility. The software's web page should include a list of the typical parameter settings to generate sequences for common uses, such as linker sequences for popular DNA assembly techniques (e.g. their optimal GC content and melting temperature, and their sensitivity to secondary structures), spacer sequences for commonly used host organisms (e.g. an expanded set of genomes to BLAST against and organism-specific forbidden sequences), *etc*. The website should also allow the user to enter an existing set of DNA sequences that the newly generated sequences should be orthogonal to, in order to be able to expand existing libraries of sequences. Secondly, the algorithm itself can be improved by adding stricter checks for undesired annealing involving the terminal regions of the linkers, which can act as "guides" for the misannealing of the whole linker, and by integrating it with external tools that can identify undesired functional motifs of various kinds that would compromise the neutrality of the linkers generated.

3. MODAL: a Modular Overlap-Directed Assembly with Linkers strategy

Aims:

- Developing a general strategy for long overlap-based DNA assembly techniques that standardises and modularises the DNA fragments being assembled and the experimental protocols involved. The strategy must also employ specifically designed linker sequences to guide the assembly of the DNA fragments.
- Testing MODAL on three of the most commonly used long overlap-based DNA assembly techniques: Gibson isothermal, CPEC and yeast recombination.
- Determining whether the use of linker sequences designed with the R2oDNA Designer software is beneficial for these long overlap-based DNA assembly reactions.
- Investigating whether the “scar” sequences left by the MODAL strategy between the DNA parts being assembled have an impact on their behaviour.
- Demonstrating the usefulness and flexibility of the MODAL strategy with a proof of principle experiment.

3.1. Introduction

The development of new DNA assembly techniques for synthetic biology has been accompanied by the development of new standards and strategies that provide a framework for modularity, hierarchical assembly, part insulation, *etc*. This applies particularly well to long overlap-based assembly, because the long homology regions used by these techniques, and the processes used to attach them to the DNA fragments to be assembled, are very similar or even identical for many of these methods.

The J5[84] and Raven[83] software tools were both explicitly developed to be compatible with various long overlap-based methods, and the similar cross-method compatibility can be assumed of the DNA assembly standards proposed by Torella *et al.*[79] and Guye *et al.*[78], described in **Chapter 1.3.2**. Even though only Gibson isothermal assembly is used, it is likely that their strategies are compatible with other long overlap-based methods. Unfortunately neither standard fully leverages the advantages of these methods, such as speed, flexibility and ease of use.

This is highlighted by the evolution that the work of Torella *et al*. underwent between its first publication and the successive adaptation for Nature Protocols[95]. Initially they proposed that linker regions should be attached to the DNA fragments by cloning them in plasmids equipped with a traditional BioBrick / BglBrick multiple cloning site flanked by the linker sequences. This has several disadvantages: the parts need to be free of certain restriction sites, this process requires a gel extraction step and two days for cloning and selection, the scar regions flanking the parts include not only the linkers but also remnants of the MCS and the terminal part even has an additional linker region. This has then been updated to include two other options to achieve this: PCR amplification of the part with primers carrying tails that encode the linker regions and total synthesis of the fragment with the linkers, which are much faster and easier to perform.

The assembly strategy proposed by Guye *et al*. takes a different approach, adopting a tiered system that uses two different assembly methods: Gateway recombination for the bits-to-genes

level and Gibson assembly for the genes-to-pathway levels (see **Chapter 1.3.2**). This has several advantages, such as being both PCR-free and restriction enzymes-free and allowing hierarchical assembly, but in turn this gives the assembly strategy a very rigid structure. *Ad hoc* solutions are required, for example, to assemble gene-level parts that are not composed by a promoter and a gene, and Gateway recombination leaves behind fixed scars that cannot be modified by the user in case they cause undesired contextual interactions. Additionally the system is quite complex and slow: it requires a transformation step for each level of assembly (to insert sub gene parts in the initial plasmids, to go from these to genes and to go from genes to pathways), and each of these requires an array of destination plasmids and helper plasmids, with various selection markers and to be digested by different enzymes.

As the fortune of the BioBrick standard faded because classic restriction/ligation methods were being replaced by the more efficient long overlap-based methods, many groups felt the need to apply the same useful principles of standardisation and modularity to these new techniques. Our approach to this problem focuses on leveraging the advantages of these methods, such as speed, simplicity and flexibility, without forcing sub-optimal fixes to the intrinsic downsides they inevitably have. As mentioned in **Chapter 1.2**, the works presented above were developed and published almost simultaneously to this work.

3.2. Results

3.2.1. The MODAL strategy

MODAL is an approach to designing cloning strategies which can be applied virtually to all long overlap-based DNA assembly techniques. It comprises three steps: the first one, “Step 0”, only needs to be performed once for each DNA part in order to “format” them with the MODAL physical standard, while the following two constitute the actual assembly process. We demonstrated the usefulness of the MODAL strategy by applying it to three of the most commonly used long overlap-based DNA assembly techniques: Gibson isothermal, CPEC and Yeast recombination. Step 0 and Step 1 are performed identically for all three techniques, while Step 2 is different for each of them.

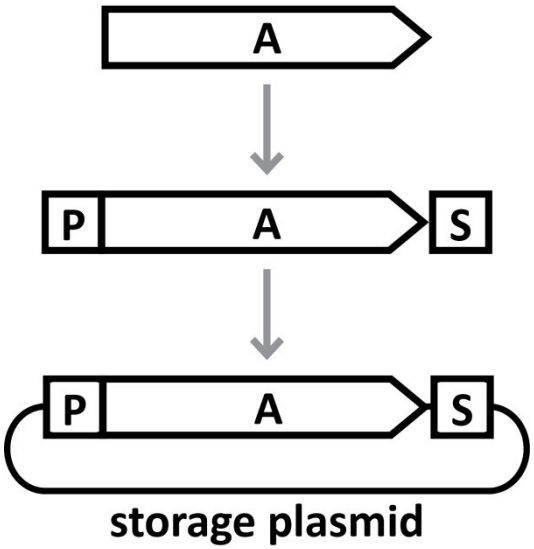


Figure 23: diagram of Step 0 of the MODAL strategy. Part A is made compliant with the standard format by amplifying it with a PCR that adds the prefix (P) and suffix (S) sequences and cloning it into a storage plasmid.

Prefix	5’-CAGCCTGCGGTCCGG-3’
Suffix	5’-CGGGCGTCCCAGCGA-3’

Table 3: the 15 bp long prefix and suffix sequences of MODAL.

Step 0: formatting. During Step 0 any desired DNA part is modified to comply with the required standard format, both to allow further processing within the MODAL strategy and to facilitate long-term storage. This is done by performing a PCR amplification on the desired DNA part using primers

that carry 5' tails which encode the prefix and the suffix. These are 15 bp sequences designed using the Linker software specifically to be used as PCR priming targets in Step 1 of the workflow (**Table 3**). They have a very high GC content (80%) in order to promote the specificity of the PCR in which they are involved: both because of their high melting temperature and because very few commonly used DNA parts will have such a high GC content, and thus risk being too similar to them. The part, which is now amplified and flanked by prefix and suffix, can be cloned in the pJET1.2 storage plasmid using the Thermo Scientific CloneJET PCR cloning kit. The pJET1.2 plasmid was chosen because it is extremely easy and quick to use, guarantees almost 100% cloning efficiency, and carries a well know, stable and high copy number origin of replication, PMB1. This allows long-term conservation of the part and allows easy preparation of large amount of DNA.

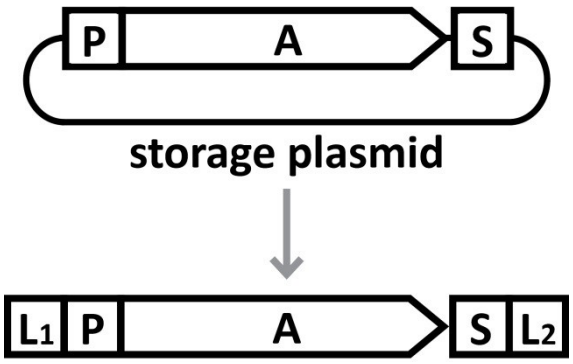


Figure 24: diagram of Step 1 of the MODAL strategy. The standard part A is amplified using the P/S sequences in order to attach the linker sequences (Ln) to its flanks. The PCR is followed by a column purification step.

Step 1: preparation. In Step 1 the parts are prepared for assembly, and it is at this stage that their relative position in the final construct is defined. A PCR amplification is run on the storage plasmid containing the desired part using universal primers: they can be used on any formatted part since they target the prefix and suffix sequences, and carry the linker sequences as 5' tails. The result is a DNA fragment constituted by the part flanked by the two linkers.

The linker sequences are 45 bp long (as recommended for the techniques listed above[44,53,96] and were generated using R2oDNA Designer. The settings used are listed in (**Table 1**), while the optimal

GC content varied depending on the assembly method used (**Figure 28**). This PCR amplification is run with conditions that promote extreme specificity at the expense of yield (see **Chapter 6.3.2**) since none of the assembly reactions in Step 2 require large amounts of DNA. These conditions include the use of the high-precision Phusion DNA polymerase, a 5% final concentration of DMSO, very little template (about 10 fmol) and the reaction mix is kept on ice until the PCR block is above 72 degsC (alternatively the Hot-Start version of the Phusion enzyme can be used). The annealing and extension steps are performed simultaneously at 72 degsC, and only 20 cycles are performed to avoid late cycle problems. This, together with the fact that the primers do not target the part itself but the prefix and suffix sequences, allows a single PCR protocol to work efficiently with a large variety of different DNA parts. Coincidentally these conditions also make the reactions very quick, lasting only about 20-30 minutes. Finally the part is purified using a DpnI digestion to destroy the template plasmid and a PCR purification kit to remove all small DNA fragments. The final result is a solution of water containing the DNA part flanked by prefix, suffix (unmodified) and linker sequences, as shown in **Figure 24**.

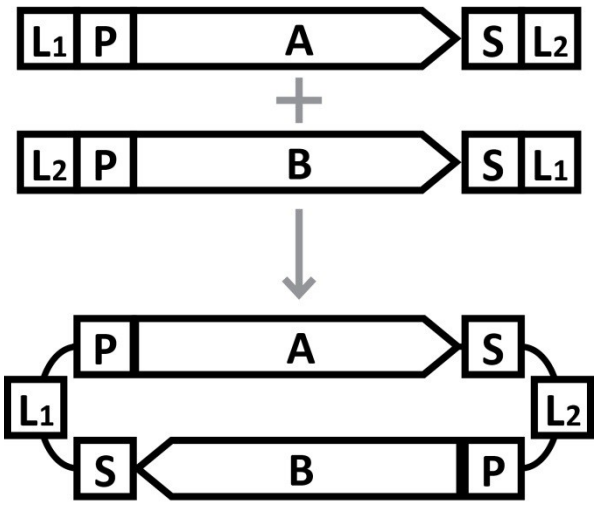


Figure 25: diagram of Step 2 of the MODAL strategy. In this example two parts are assembled, but the process is identical for any number. The two parts (A and B) to be assembled, now equipped with linkers and purified, are mixed in the chosen assembly reaction. The linker sequences guide the reaction, joining the parts in the desired order. The reaction mix can then be used for cell transformation.

Step 2: assembly. Step 2 is where all the parts prepared in Step 1 are assembled: they are mixed together under the appropriate reaction conditions, and the linker sequences guide the assembly into the final construct. There are two complementary versions for each linker sequence, called "forward" and "reverse", and they get fused to each other during the assembly reaction, so that it is possible to define exactly which parts are joined. All that is required to change the order or the orientation of the parts in the final construct is changing the linkers that get attached to the parts during Step 1. It is also possible to invert the orientation of a part by using special "inversion linkers" that attach the forward linker to the suffix of a part (rather than the prefix), and the reverse linker to the prefix (rather than the suffix). While the process of preparation and purification in Step 1 is identical for all three techniques considered in this study (and virtually for all long overlap-based DNA assembly techniques in general), this final assembly step is always different for each method (see **Chapter 6.2.5**). The common features, beside the fact that they can all use the same prepared DNA parts, is that all the parts that compose a construct are mixed in a single tube and assembled together simultaneously, and that the sequence of the final construct is identical regardless of what technique was employed.

3.2.2. Experimental test of MODAL and of the impact of linkers on assembly efficiency

The MODAL strategy was tested experimentally both to confirm it works as expected with all three DNA assembly techniques (Gibson isothermal, CPEC and yeast *in vivo* recombination) and to determine whether the use of optimised linker sequences is beneficial. The test was run by assembling test plasmids containing the same functional parts using the three techniques and six different linker sets in a few combinations. **Figure 26** shows the general structure of this plasmid, while the specific parts changed depending on the final recipient organism: *E. coli*-compatible parts were used for Gibson isothermal and CPEC assembly (P15A origin of replication, kanamycin resistance marker and bacterial constitutive GFP and RFP expression) while *S. cerevisiae*-compatible parts were used for yeast recombination assembly (2-μ origin of replication, uracil selection marker and yeast-optimised constitutive GFP and RFP expression). We chose to express fluorescent proteins on these plasmids in order to easily understand if the assembly was successful, without having to run diagnostic test or sequence them. A correctly assembled plasmid containing all four parts will produce yellow colonies. A plasmid missing one part will either not give colonies (if it lacks either the origin of replication or the selection marker) or give red or green colonies (if it lacks respectively the GFP or RFP gene). A plasmid missing two parts will either not give colonies (if it lack one or both of the essential parts) or give white colonies (if it lacks both fluorescence genes). A plasmid missing three parts will always be non-viable.

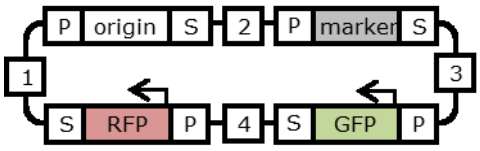


Figure 26: schematic of the plasmids that were built to test MODAL. Bacterial plasmids contained P15A origin of replication, kanamycin resistance marker and bacterial constitutive GFP and RFP expression, while yeast plasmids contained 2-μ origin of replication, uracil selection marker and yeast-optimised constitutive GFP and RFP expression. The numbers in the squares represent the linkers, 45 bp sequences generated with the R2oDNA Designer tool.

The six sets of linkers were designed as follows: the ones called “40% GC”, “50% GC” and “60% GC” were generated with R2oDNA Designer using the same optimised settings (**Table 1**) except for the different GC contents. The “Random” set was generated by the script using no rules or optimisation algorithms except that the GC content was set at 50%.

The four linkers in the “Functional” set encode short functional parts within their sequences: a promoter (BBa_K093000), a terminator (BBa_B1006), a peptide tag sequence (BBa_J32017) and an RNAse III site (BBa_I13536). These were selected using R2oDNA Designer’s ability to score existing sequences for suitability as use as linkers, which is described further in the relative paper[93]. We downloaded the list of DNA sequences of all BioBricks from the publically-available Registry of Biological Parts (www.partsregistry.org), we selected all sequences between 38 and 50 bp in length and deleted the rest. These short sequences were then assessed with R2oDNA Designer and we chose four of the highest scoring ones that we thought were representative of commonly used parts in synthetic biology. We converted these into 45 bp sequences either by trimming them or by randomly adding nucleotides at both ends, and they all have a GC content close to 50%, except for one which is about 25%.

Finally the “Scarless” set was designed differently from the rest of the sets: the DNA parts using this set have no prefix and suffix sequences, and instead of the 45 bp linkers they are flanked by 22 or 23 bp sequences that match the sequence of the parts that are next to them in the final construct. In this design each part still has a 45 bp homology region with the parts it needs to be assembled with (with GC contents evenly distributed from 44.4% to 60.7%), but the final construct will exclusively contain the sequences of the parts, without any “scars” between them.

In order to investigate the influence of the linkers’ features on the behaviour of long overlap-based DNA assembly reactions we used all six linker sets to assemble the test plasmid with Gibson isothermal reactions. In addition to this, in order to confirm that this strategy works with CPEC and yeast transformation as well, we used the 40% and 60% linkers to assemble the test plasmid with

both those techniques. This also allowed us to investigate which GC content is optimal for them. All transformation plates were imaged using a fluorescence scanner: **Figure 27** shows, as an example, a plate for each of the six sets of linkers from the Gibson isothermal transformations. The colonies appear yellow when they are correctly assembled and express both GFP and RFP, green or red when they are missing respectively the RFP or GFP gene and white if they miss both. **Figure 28** shows the total number of colonies obtained for each sample, and what percentage of them is correctly assembled (appears yellow in the scansions).

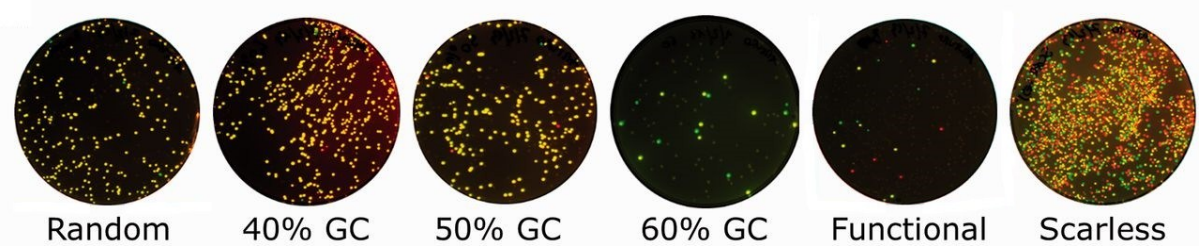


Figure 27: an example of the plates obtained from the MODAL efficiency test. Here DH10B *E. coli* cells were transformed with Gibson assembly reactions and grown on agar plates overnight and then scanned for green and red fluorescence the following day. This was repeated three times on different days, and a single set is shown here. Each plate represents an assembly reaction performed using one of six different linker sets: random, designed with 40% GC content, designed with 50% GC content, designed with 40% GC content, functional and scarless. Correctly assembled plasmids produce colonies that appear yellow due to simultaneous green and red fluorescence.

The results confirm that the MODAL strategy is perfectly viable with all three techniques: under optimal conditions (in brackets) over a thousand colonies were obtained with both Gibson isothermal (40% GC) and yeast transformation methods (40% GC) and about a hundred with the less powerful CPEC method (60% GC). In all cases accuracy of assembly was very high, the lowest being Gibson isothermal with about 80% correct colonies up to 100% correct colonies with yeast transformation. The data shows that GC content makes a significant difference both in terms of number of colonies and accuracy: Gibson isothermal reactions using 40% GC linkers have about ten times more colonies and are almost three times more accurate than those with 60% GC linkers. Similarly yeast transformations using 40% GC linkers produced about five times more colonies than

those using 60% GC linkers, and CPEC reactions using 60% GC linkers gave about nine times more colonies compared to those using 40% GC linkers.

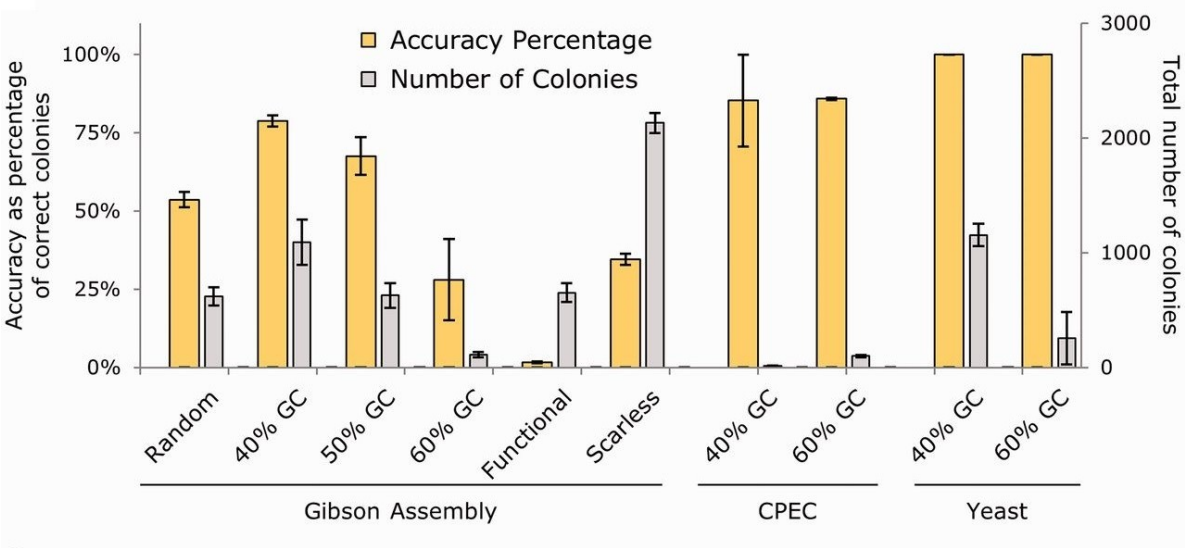


Figure 28: results of the MODAL efficiency test. The total number of colonies and the percentage of those containing correctly assembled plasmids (accuracy) were calculated from image analysis of each plate for DNA assemblies using different linker sequences and using the Gibson (n=3), CPEC (n=3) and yeast *in vivo* recombination (n=2) DNA assembly methods. Error bars indicate standard error. The 60% GC test was also repeated using a different set of linkers with the same GC%, obtaining essentially identical results.

This experiment also confirmed that the use of computationally optimised linker sequences is beneficial for Gibson isothermal DNA assembly reaction. The “Random” set gave a number of colonies similar to that of the “50% GC” set but with a significantly lower accuracy, even though they have an identical GC content. The “Functional” set also produced a number of colonies similar to the “50% GC” set, but the accuracy here is close to zero. The “Scarless” set finally gave an extremely high number of colonies but only about a third of the colonies were correct. Noticeably the incorrect colonies were mostly white for all sets except for “Scarless”, where there was a high number of white, green and especially red colonies.

3.2.3. Investigation the impact of linker regions on local gene expression

In the MODAL strategy all the parts that are assembled are separated by a 75 bp region containing the suffix-linker-prefix sequences, which essentially constitutes a "scar". We investigated whether the presence of this "scar" can affect gene expression in *E. coli*. Firstly we considered constructs where the "scar" is only located between expression cassettes, not inside them (**Figure 29**). This means that all the genes in the construct come as a single functionally independent part. We used the same four parts shown in the chapter above: two essential ones (P15A origin of replication and kanamycin resistance cassette) and two fluorescence genes (GFP and RFP), to easily identify correctly assembled constructs. We shuffled the parts around the construct in four different combinations, so that they would be separated by different linkers, and measured fluorescence emission for both GFP and RFP. This test was run before R2oDNA Designer was developed, so the linkers used here were generated using the Linker script with the standard settings reported in **Chapter 2**. The results in **Figure 30** seem to indicate that GFP and RFP expression are not affected by changing the linkers around them.

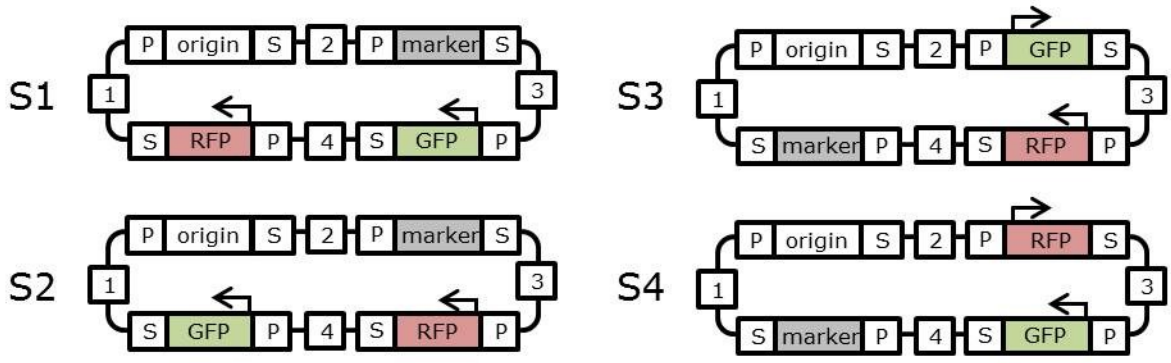


Figure 29: schematics of the plasmid variants (S1-S4) built to explore the context effects caused by linker sequences placed outside expression cassettes. All four contain identical parts and linkers, but arranged in different orders, so that they are flanked by different linkers. The linkers in this experiment are 45 bp sequences generated with the Linker script. Gibson assembly was used to build these plasmids.

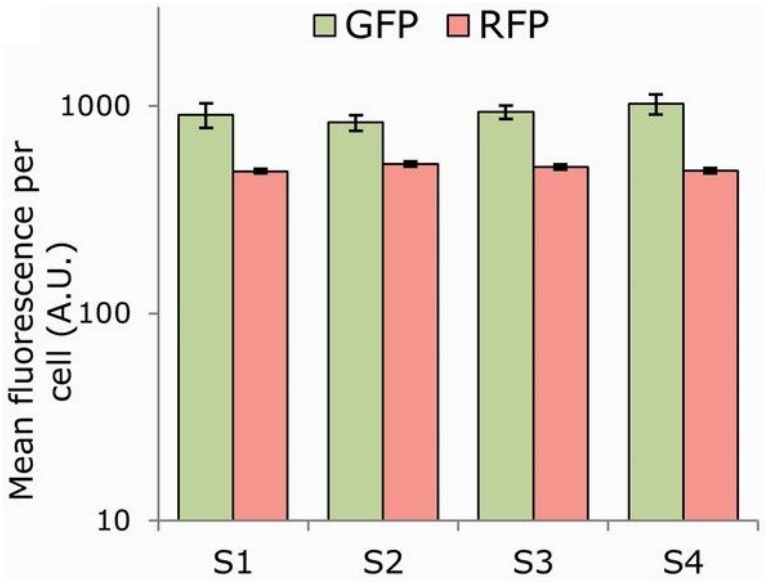


Figure 30: context effects caused by linker sequences placed outside expression cassettes. Plasmids S1-S4 contain GFP and RFP genes flanked by different linkers, but GFP and RFP expression per cell as measured by flow cytometry did not show significant variation. Mean fluorescence per cell was calculated from mean FL1 (GFP) and mean FL5 (RFP) measurements (n = 5). Error bars indicate standard error.

We then proceeded to test whether this hold true when the "scar" sequences are located inside expression cassettes. We designed a new set of plasmids made of five parts (**Figure 31**) where the essential parts are the same as before (P15A origin of replication and kanamycin resistance), but the GFP and RFP genes here are combined in a single operon. The operon is composed of three parts: a constitutive promoter, the GFP open reading frame and the RFP open reading frame. The two fluorescence ORFs appear in both orders in the various constructs, they are both preceded by an RBS and the last one in the operon is followed by a terminator. As shown in **Figure 31** linkers 1, 2 and 3 remain constant, while the linker between the promoter and the first ORF, and the one between the first and second ORF can change. **Figure 31** shows the GFP and RFP fluorescent emission of all the plasmids that were built. The code under each pair of bars represent the structure of the plasmid variant: the first number is the linker between the promoter and the first ORF, the letter that follows it specifies whether the first ORF is GFP (G) or RFP (R), the second number is the linker between the first and second ORF, and the last letter again represents the last ORF, either RFP or GFP. All the

linkers for this test were generated using R2oDNA designer with default settings. The results show that in this case gene expression is affected by the changing context: not only by what linkers are flanking each ORF, but also by the order in which they appear.

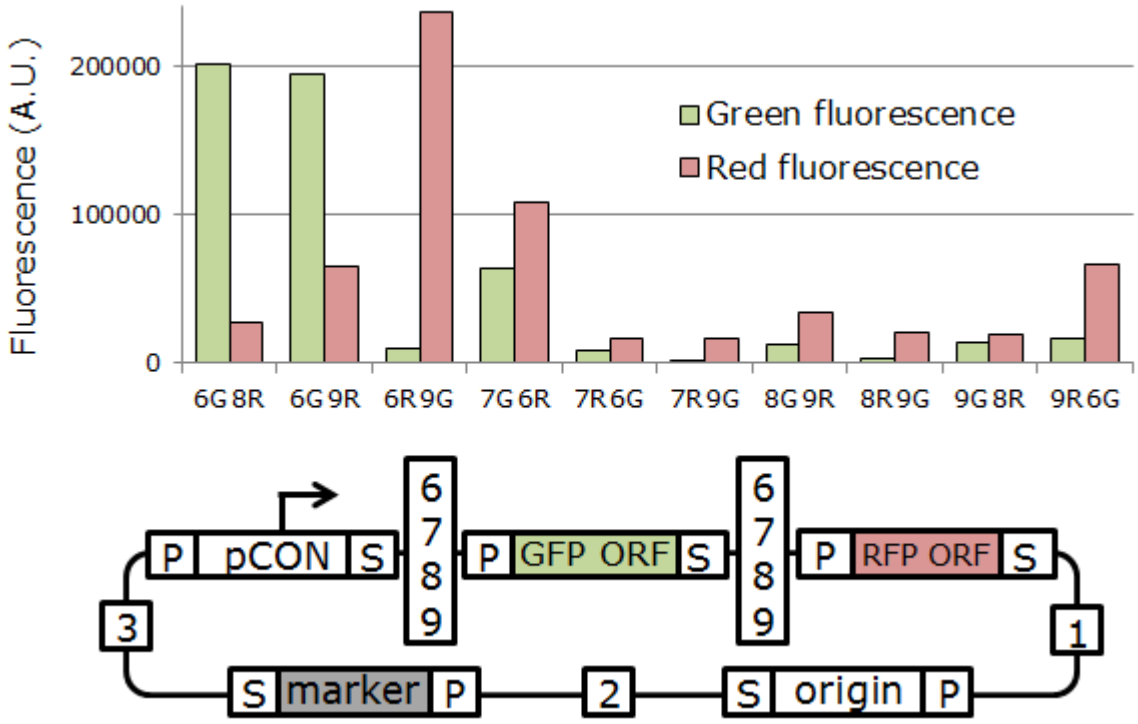


Figure 31: context effects caused by linker sequences placed inside expression cassettes. The diagram shows the general structure of the plasmids used for this test: they all contain the same parts, and the GFP and RFP genes are placed in an operon controlled by a single constitutive promoter (pCON, representing the pT7A1 promoter). Different variants were built, using both orders of the open reading frames (ORFs) in the operon (GFP-RFP and RFP-GFP) and four different linkers (6-9) between the promoter and the first ORF, and the first ORF and the second ORF. The linkers in this experiment are 45 bp sequences generated with the Linker script. Not all possible variants were built: the naming system in the plot above shows which of the linkers and what order in the operon was used for each tested variant (*e.g. "*6G 8R" represents a plasmid where GFP is placed first in the operon, and linker 6 is placed between the promoter and the GFP, while linker 8 was used between GFP and RFP). The bar chart shows the result of plate reader assay for GFP and RFP expression of the assembled plasmids (n=1): linkers placed inside an expression cassette have a significant influence on its behaviour.

3.2.4. Library generation with the MODAL strategy

The MODAL strategy can be adapted very easily to generate mutagenic libraries for any of the parts being assembled. While normally each part is prepared for assembly (Step 1) using a high-fidelity PCR amplification (black arrows in **Figure 32**), it is possible to replace it with a mutagenic PCR amplification (blue arrow) to generate a pool of mutated products which can then be assembled combinatorially in the final construct without any modification to the workflow. We decided to use the mutagenic PCR protocol developed by Zaccolo *et al*.[97] which uses the non-proofreading Taq polymerase enzyme in combination with two nucleotide analogues: dPTP and 8-oxodGTP. Their incorporation causes a variety of transition and transversion mutations but no deletions or insertions.

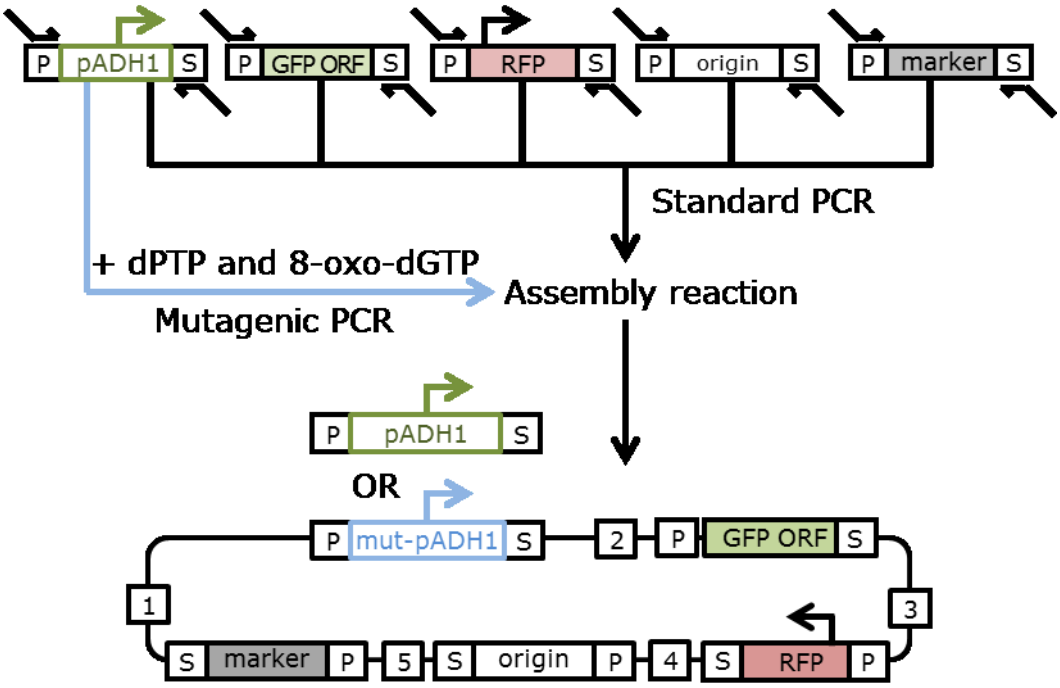


Figure 32: integration of mutagenesis within the MODAL strategy to create mutant libraries of one or more of the parts that are being assembled. Selected parts can be mutated as part of the standard assembly workflow by adopting a different protocol for the PCR in Step 1 (blue arrow). This mutagenic PCR[97] employs Taq DNA polymerase, dPTP and 8-oxo-dGTP to incorporate a high percentage of sequence errors in the amplicons. The yeast plasmid shown was built both using the standard workflow for all parts, and using the mutagenic workflow for the promoter part (pADH1) that controls the GFP part. Yeast *in vivo* recombination was used to join the parts.

We tested this variation of the MODAL strategy using yeast recombination assembly and the "40% GC" linker set on a five parts plasmid for *S. cerevisiae* similar to the one shown in **Chapter 3.2.2**: it contains a 2-µ origin of replication, a uracil selection marker, a yeast-optimised constitutive RFP expression cassette and a yeast-optimised GFP expression cassette which was split in two parts. For this test we separated the pADH1 promoter from the rest of the gene so that we could run the mutagenic PCR protocol on the promoter part alone, in order to obtain different GFP expression levels, while leaving the GFP open reading frame and the rest of the plasmid untouched.

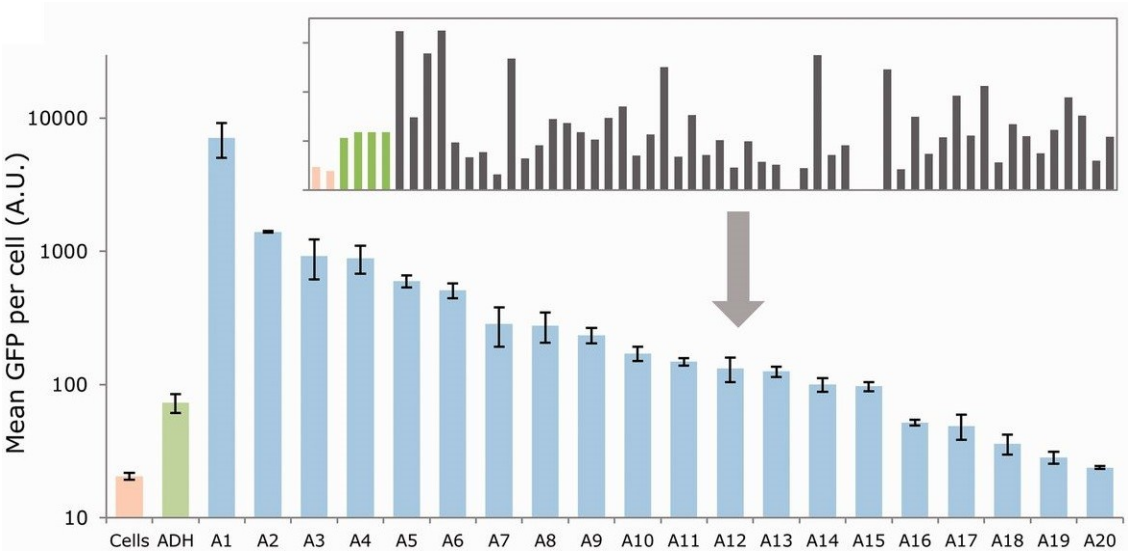


Figure 33: results of the selective mutagenesis experiment. Initially 52 colonies were randomly picked and analysed by flow cytometry for mean GFP expression (inset). Successively 20 samples were chosen to build a graded library of ADH1 promoters (A1-A20), covering a 3 orders of magnitude expression range,, above and below the output provided by the unmutated ADH1 promoter (ADH, green). The red bar represents control colonies with no GFP expression, and the error bars indicate standard error (n = 3).

We assembled this plasmid both using the mutagenic protocol on the pADH1 promoter and using the normal protocol, as a control. As the inset in **Figure 33** shows, we picked 52 colonies from the mutated protocol (grey bars) and 4 control colonies (green bars). We used dual colour flow cytometry to measure GFP and RFP expression: RFP expression was used to normalise for copy number variation, which is a very prominent effect in this kind of yeast plasmids (see **Chapter 6.4.1**),

while GFP expression was used to investigate the mutations in the promoter. The results showed that indeed the control colonies, containing a non-mutated promoter, kept a constant expression level, while those containing mutated promoters had significant differences. We then proceeded to select a library of 20 mutated ADH1 promoters that gave a wide variety of GFP expression levels, repeated the flow cytometry measurements in triplicate to have a solid characterisation of their behaviour (**Figure 33**). We also sequenced all 20 promoters in the library (**Figure 34**) and determined that our protocol gave a mutation rate of about 10%.

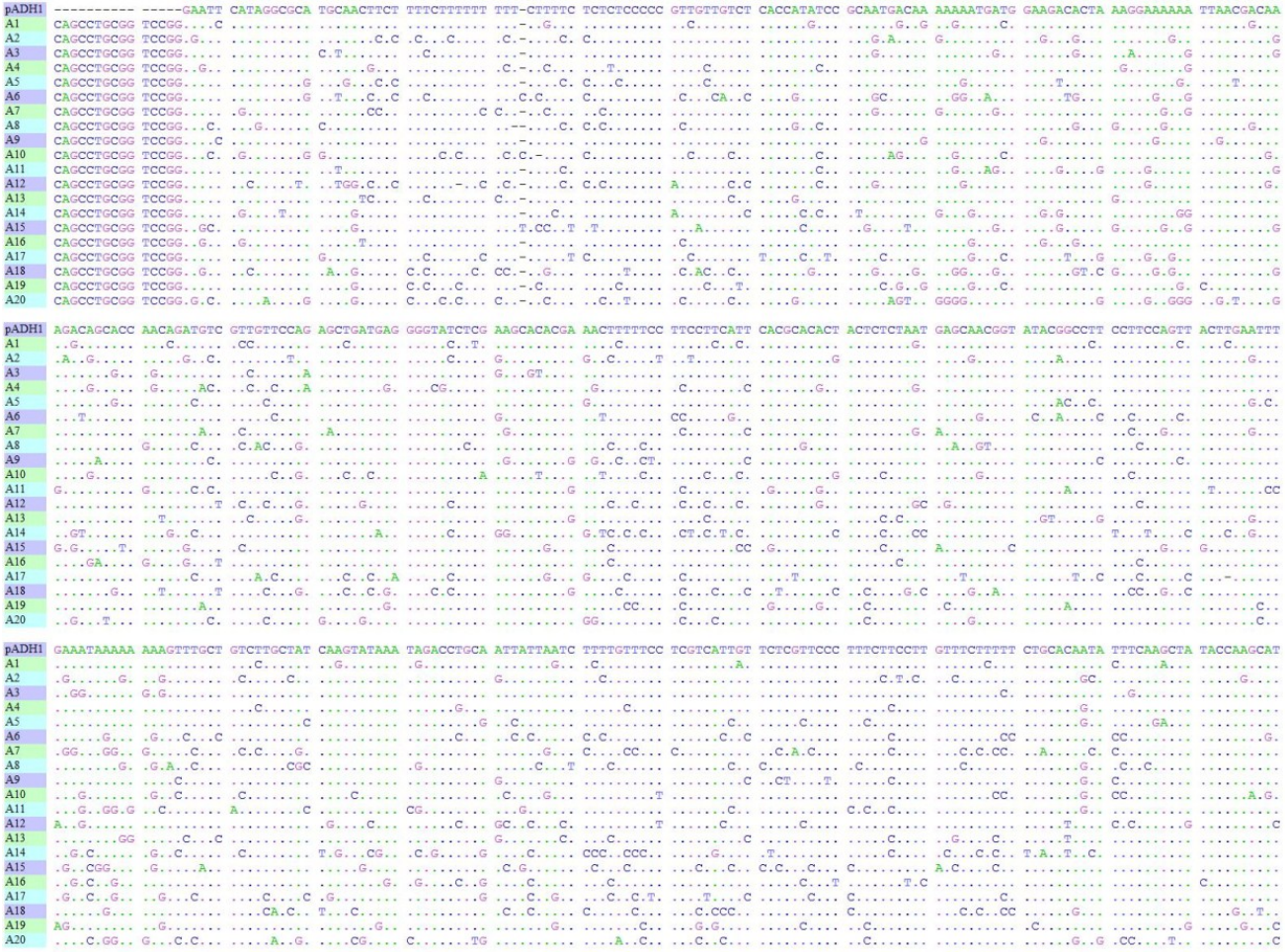


Figure 34: sequences of the original (pADH1) and mutated (A1-A20) promoters from the selective mutagenesis experiment. The first and last 15 bp sequences are the MODAL prefix and suffix. The sequences in between are represented as dots when they are unchanged compared to the original.

3.3. Discussion

3.3.1. Linker design has a significant impact on long overlap-based DNA assembly reactions

We validated our linker-based approach by assembling the usual four-parts test plasmid using Gibson isothermal assembly and six different sets of overlap sequences. Three of these were linkers designed and optimised using our R2oDNA Designer software, with three different GC contents (40%, 50% and 60%), and three were control sets, which represented common scenarios routinely used or encountered by researchers in the field. The goal of the test was twofold: on one hand verifying that our linkers are an improvement compared to previous approaches, on the other hand exploring to what extent GC content influences the DNA assembly reaction. The most basic control was the "Random" set, which was meant to compare our optimised sequences to non-optimised ones with the same GC content to determine whether the optimisation process was actually useful or not. Our results showed that our optimised "50% GC" set improved accuracy by about 25% compared to the "Random" set, and thus that our optimisation process is actually beneficial to Gibson isothermal reactions.

The "Functional" set instead was designed to be illustrative of the problems of an approach suggested previously where short functional parts are encoded within the sequences of the overlap regions[77]. This design can also be representative of certain problems that might arise with scarless approaches: it is often unavoidable in scarless designs to use functional sequences as overlap regions. Our design is biased because we actively chose sequences that R2oDNA Designer marked as "bad" for assembly reactions, but it is nonetheless plausible: even though this is not a scenario that will occur every time a functional sequence is used as overlap regions, it is still something that could occur and thus worth considering when designing cloning strategies. The results showed that one needs to be extremely careful when using a sequence that encodes for a function as overlap region: one or more of the sequences that were included in the "Functional" set had an extremely negative

impact on the reaction, essentially completely preventing it from working (accuracy was the lowest of the whole test, at 2%).

The last control set, “Scarless”, was meant to investigate the possible problems one might encounter when assembling the usual test plasmid with a scarless approach. In this case a problem arose because two of the parts in the plasmid, the GFP and RFP genes, contain the same promoter and terminator sequences, located respectively at the very beginning and very end of the part. Because of the scarless design these terminal regions are part of the overlap regions and thus create an orthogonality problem: the overlap region attached to the kanamycin resistance part matches both the GFP part (as intended) and the RFP part (which is incorrect), and similarly the origin of replication part matches both the RFP part (intended) and the GFP part (incorrect). Only half of the total 45 bp overlap between these parts is actually wrong though, because it is composed of 23 bp on one of the partners and 22 bp on the other, and the GFP and the RFP parts carry unambiguous homology regions that only match their correct partners. Unfortunately even only half-ambiguous homology regions are enough to facilitate the assembly of three-parts plasmids that exclude one of the two fluorescence genes, and since plasmids with fewer parts also have an intrinsically higher likelihood of being assembled this resulted in an extremely high number of colonies containing three-parts plasmids during the experiment (colonies were red or green instead of yellow on the transformation plates).

The results show that using the scarless approach an extremely high number of colonies is obtained, which suggests that three parts plasmids with a partial mismatch in the overlap regions might still be more likely to be assembled than perfectly matched four-parts plasmids. The low accuracy might be due either to the lack of optimisation in the overlap regions or to the reduced availability of parts for four-parts assembly because they were being preferentially used for three-parts assemblies. It is important to note that orthogonality issues can arise not only with scarless approaches but with any design that does not take homology region orthogonality into account

appropriately. The R2oDNA Designer software employs a network elimination algorithm to ensure the orthogonality of all sequences it generates within a single run.

The other three sets of linkers, “40% GC”, “50% GC” and “60% GC” were used to investigate the impact of the linkers’ GC content on the behaviour of Gibson isothermal reaction. The results show that GC content is an extremely important parameter: both number of colonies and accuracy progressively improve by a significant amount descending from 60% through 50% and to 40% GC content. It is important to note that while the “50% GC” and “40% GC” set both gave good results, better than the “Random” set, the “60% GC” set gave quite bad results, worse than using non-optimised linkers. In order to confirm this we repeated the experiment using a different set of linkers with 60% GC content and obtained the same result, demonstrating that the low efficiency was not a specific feature of one or more of the sequences used as a linker.

We also tested the “40% GC” and “60% GC” sets with two more long overlap-based DNA assembly techniques: CPEC and yeast transformation. The goal was both to confirm that the MODAL strategy works with these methods too, and to investigate whether the optimal GC content is the same of all three techniques or if it changes. These techniques were chosen not only because they are among the most commonly used DNA assembly reactions, but also because their mechanisms are extremely different from each other. Gibson isothermal reaction utilises a three-enzyme *in vitro* DNA repair mechanism[44], CPEC is essentially a PCR amplification where the DNA fragments to be assembled prime each other[1], and yeast transformation assembly[50] is an *in vivo* method that takes advantage of *S. cerevisiae*’s natural DNA repair machinery. This would help us understand if the advantages brought on by our approach are universal to long overlap-based DNA assembly methods or specific to a particular technique.

The results confirmed that our approach works with all three techniques but showed that their optimal GC content varies: Gibson isothermal and yeast transformation both obtained the best results with the “40% GC” set, while CPEC favoured the “60% GC” set. We attribute this difference to

the temperatures involved in each reaction, since the first two run at lower temperatures (Gibson isothermal at 50°C and yeast recombination uses a 42°C heat shock and 30°C cell growth) while CPEC uses thermal cycles that alternate between 72°C and 98°C. At CPEC's high temperatures all base pairing, regardless of GC content, can be melted and any misfolding or misannealing of single-stranded DNA is prevented. High GC content then acts to improve the accuracy of the linkers coming together. During Gibson isothermal assembly and yeast recombination instead the lower temperature might not be sufficient for the DNA overlaps to melt very efficiently: a high GC content is thus more likely to lead to the overlap sequences being caught in thermodynamic traps that inhibit the search process for the correct partner and this might prevent mismatched linker pairings from being resolved. On the other hand with lower GC content linkers the thermodynamic barrier to sampling different pairings is reduced, thus facilitating the search process.

The experiment presented here comes with certain limitations that is worth considering: the "Random" set control indicates that GC content is not the only important factor in determining the accuracy of a Gibson isothermal reaction, but cannot point to any other specific effect in play, as it could be any of the parameters optimised by the R2oDNA Designer software, such as the orthogonality of the linkers, the absence of secondary structures, *etc*. For what concerns the "Functional" set, while all the linkers received a very bad score from the R2oDNA Designer analysis, one of them also had a particularly low GC content (25%). This is a very low value, unlikely to be optimal for annealing-based mechanisms, and we know that non-optimal GC content values can be highly disruptive for Gibson assembly. On the other hand we never tested such a low value and Gibson's reaction seems to favour low GC overlaps, so it is not possible to determine exactly which factor is mainly responsible for the low efficiency of Gibson assembly with the "Functional" set. Finally, in the "Scarless" set the presence of a large number of green and red colonies indicates that the lack of orthogonality between some of the overlaps was one of the main factors in play, but the GC content of the overlap regions varied between 40% and 60%, so it is possible that the presence of 60% GC content overlaps acted as bottleneck for the efficiency of the reaction.

The results with the “Random”, “Functional” and “Scarless” sets were obtained using Gibson isothermal assembly only and cannot be automatically extended to CPEC and yeast *in vivo* recombination or any other long overlap-based techniques: while all of these techniques are similar in the fact that they use long homology regions to guide the assembly reaction, their mechanisms of action are very different. These results suggest that certain features of the homology regions (secondary structures, orthogonality, GC content, *etc*) might have a significant impact on the performance of long overlap-based assembly methods, but it is likely that different methods will be more sensitive to different parameters and will have different optimal values for them.

3.3.2. Investigation of the biological neutrality of linker regions when located in intergenic or intragenic regions

The MODAL strategy leaves a 75 bp "scar" region between each of the parts being assembled, which contains the suffix-linker-prefix sequence. This synthetic DNA seems to be well-tolerated when located between expression cassettes (gene-level parts), as the results in Figure X appear to indicate that it does not affect expression levels of assembled parts. However, when assembling at the sub-gene-level (e.g. assembling a promoter and an ORF to make a gene), we observed a significant variation of the expression levels.

Similar effects have been found by several groups while investigating the role of local flanking sequences on gene expression: the sequence of the junctions between RBSs and ORFs in *E. coli* seem to dramatically affect part function[98–102]. To some extent this has also been observed in yeast[103]. All these local context effects involve DNA sequences that are transcribed to mRNA, exactly like in our tests, and are likely due to differences in local RNA folding within transcripts, modulating the efficiencies of elements such as RBS sequences[98] and changing the stability of the mRNA. Clearly these effects are problematic as they prevent predictability of gene expression, but recently three studies have tried to tackle the issue by designing RNA processing parts that alleviate these effects and improve predictability[100–102]. On the other hand the addition of synthetic DNA sequences that are computationally optimised to be functionally neutral (R2oDNA Designer[93] provides this to some extent) outside expression cassettes may actually be beneficial by providing some level of insulation against local context effects from neighbouring parts by acting as physical spacers. Previous work has shown that adding a spacer sequence upstream of a promoter improves predictability when reused in different constructs[104].

3.3.3. Conclusion

MODAL provides a framework that brings the benefits of standardisation and modularisation to long overlap-based DNA assembly techniques, which are some of the most advanced and powerful assembly methods currently available to synthetic biology researchers. Once a one-time formatting process, “Step 0”, is performed on a DNA biopart (which can be any DNA fragment amplifiable by PCR) this can be assembled with any other formatted part within one day of work, followed by transformation in either *E. coli* or *S. cerevisiae* depending on the specific technique adopted. This enables the user to easily reuse any formatted part for any new construct or variant without having to perform any additional transformation steps, saving days of work: any part can be easily moved around and/or its orientation inverted by simply using different linker oligonucleotides. The MODAL workflow can also be adapted to seamlessly perform targeted mutagenesis (via mutagenic PCR) on any of the parts being assembled.

The linker-based approach adopted by MODAL brings a number of other benefits as well: first of all it significantly improves the performance of the assembly reaction itself, both by increasing efficiency and by reducing chances of encountering unexpected problems. This, together with the efforts to make both Step 0 and Step 1 of the workflow as reliable as the techniques involved allow, aims to make this workflow more reliable than most alternatives.

Secondly, the presence of linkers between the assembled parts in the final construct helps reducing context effects which might affect the behaviour of the parts by physically spacing them away from each other. The linker sequences are computationally designed to be functionally neutral and they seem not to cause any context effects when located between TUs. Linkers do affect expression when located inside TUs, but this is unavoidable due to the extremely sensitive nature of mRNA to even small changes in the UTRs[98].

The flexible and fast nature of the MODAL workflow helps addressing the problem of context dependency by making it very easy for the user to explore different topologies of the final construct

and change any linker that might be causing problematic effects. It is important to note that context effects affecting UTRs will occur when any change happens within or near them, regardless of whether one is using a scarless or a linker-based approach. A flexible linker-based approach actually helps by facilitating troubleshooting, while with a scarless approach one might not have the freedom to change the sequence of the parts being assembled.

The work described in this chapter has been published in "One-pot DNA construction for synthetic biology: the Modular Overlap-Directed Assembly with Linkers (MODAL) strategy" by Casini *et al.*[105]

3.3.4. Future work

In the future it would be interesting to investigate thoroughly the influence of the features of linker sequences on the behaviour of assembly reactions, considering not only GC content but also its uniformity across the linker and different linker lengths, in order to find the optimal characteristics. It would also be important to compare different sets of linkers with the exact same parameters, to find out if there are any other features that are also affecting assembly efficiency that we do not know about yet. Another important factor to explore is whether the presence of repeated sequences in the constructs can cause undesired recombination events: this and many other frameworks for standardisation utilise sequences during the assembly process that remain in the final construct, and that are usually identical between every assembled part. It would be important to find out to what extent this is safe and when it begins to cause recombination events, keeping in mind that this effect might differ significantly from host to host. Finally MODAL would greatly benefit by making sure that linker sequences are as functionally neutral as possible, and a further step in this direction could be taken by integrating the R2oDNA Designer tool with other external tools that are able to predict and identify the presence of functional motifs in DNA sequences, as mentioned in **Chapter 2.3.4**.

Regarding the development of future DNA assembly workflows, our work on MODAL highlighted two major issues: the first is the necessity of moving away from a PCR-based approach, which is very quick and easy but suffers from the limitations mentioned above, both in terms of what is reliably amplifiable and in terms of fidelity. If we envision a future where DNA is assembled completely hands-off through robotic automation, PCRs are too unpredictable to be viable. On the other hand utilising strategies that require a large number of forbidden sequences is also unacceptably limiting, especially as projects scale up in size. This ties in with the fact that any assembly framework that aspires to become widely adopted in the synthetic biology community needs to be as simple as possible: requiring multiple pre-prepared plasmids, exotic enzyme kits or the absence of multiple forbidden sequences, is unlikely to be well received by most research groups. The second issue that

needs to be addressed is idempotency: the ability to run hierarchical assembly is extremely useful especially when scaling up to larger constructs that are difficult to assemble in one step, so it is important that any assembled construct can be reused for further cycles of assembly without adding too much complexity to the workflow.

4. BASIC: a Biopart Assembly Standard for Idempotent Cloning

Aims:

- Developing a long overlap-based DNA assembly method that achieves higher accuracy and reliability compared to the current state of the art in *in vitro* DNA assembly.
- Employing this novel DNA assembly method within a framework based on the MODAL strategy (Step 0-1-2 workflow and use of computationally designed linkers), where all the steps are designed to be extremely robust, reliable and suitable for automation.
- Devising a procedure that generates idempotent constructs within the same workflow, to allow hierarchical assembly.
- Leveraging all of the above to develop BASIC, a physical standard for DNA assembly that is also protein fusion-friendly and allows the user to alternatively employ the MODAL workflow.

4.1. Introduction

The development of long overlap-based techniques has brought significant improvements over traditional restriction/ligation cloning, mainly by allowing efficient parallel assembly of more than just two or three fragments simultaneously (see **Chapter 1.2**). Unfortunately, as **Chapter 3.2** and other works[3] have shown, even the best of these methods are subject to quite strong limitations: CPEC is the easiest to use, but also the least efficient and most likely to introduce errors. Gibson isothermal assembly is more efficient but its accuracy decreases very quickly when more than four or five parts are assembled simultaneously. Yeast *in vivo* recombination is the most efficient and accurate, but takes three or four days to perform, instead of two, and makes it very difficult to isolate assembled constructs.

Two methods have recently been developed that achieve higher efficiency and specificity: the first is Golden Gate, which employs type IIs restriction enzymes in a simultaneous digestion/ligation reaction. It has been shown to be able to assemble plasmids up to 33 kb made of up to seven parts with near-100% accuracy, and uses a colour-selection system that further simplifies post-assembly screening[33]. On the other hand this technique is not very flexible, and every attempt at exploiting it in a standard modular framework resulted in extremely complicated and unwieldy systems (see **Chapter 1.3.1**). The other method is LCR (Ligase Cycling Reaction), which uses thermal cycling and bridging oligonucleotides to fuse DNA fragments with great efficiency: the assembly of constructs as big as 20 kb and made of up to 20 parts has been demonstrated, again with near-100% accuracy[3]. Unfortunately there is no standard modular framework available for this method, and some of the practices proposed by the authors are less than optimal: parts are prepared for assembly with a PCR amplification, with all the limitations it entails, and the scarless approach implies the use of non-optimised homology regions which can cause a variety of problems (see **Chapter 3.2.2**). Other problems are intrinsic to the way the method works: the annealing of DNA strands becomes less and less efficient as these become longer, which might explain why the efficiency of LCR decreases so much going from 500 bp parts to 2 kb parts, limiting the usefulness of LCR in assembling very large

constructs. Additionally the method is subject to an element on unpredictability since misassembly events during the early cycles act as template for more misassemblies, propagating them through the rest of the reaction.

None of the currently available DNA assembly standards is completely satisfactory, either because too unwieldy, or too rigid, or because it employs non-optimal assembly techniques. Ultimately we decided to create our own standard and method, BASIC. It retains MODAL's linker-based structure, which allows modular and combinatorial assembly, supports optimised homology regions and is very simple, adaptable and fast. We then built on it by implementing hierarchical assembly and defining a new assembly method that achieves similar efficiency and accuracy as Golden Gate and LCR. Additionally we wanted this system to be as robust and predictable as possible by avoiding chain reactions (e.g. PCR or LCR), gel extractions, undefined "chew-back" reactions (e.g. Gibson assembly), *etc*. The goal is to minimise the need for *ad hoc* troubleshooting, which in turn makes BASIC a great choice for the majority of researchers, who do not have vast experience in DNA assembly, and for robotic automation, which intrinsically requires dependable protocols.

4.2. Results

4.2.1. The BASIC standard

The BASIC standard is defined at its core by the integrated prefix and suffix sequences (iP and iS, **Figure 35**): any DNA fragment is compliant with the standard when it is flanked by those sequences. The iP/iS have three roles: they allow assembly through the BASIC workflow, they allow assembly through the MODAL workflow and they encode for a protein fusion-friendly scar, regardless of the workflow adopted. Additionally, a BASIC-compliant part can be cloned in a storage plasmid that guarantees part stability and high DNA preparation yields, and allows sequence verification of the part. When the BASIC workflow is used, since it does not contain any PCR, DNA repair or transformation steps, chances of mutations are extremely low, and sequence integrity is essentially guaranteed.

Figure 35: sequence of the BASIC prefix and suffix. Bases in red constitute the recognition sequence for the BsaI enzyme, which cuts where the red arrows indicate. In green are the bases that will be left as scars in the final construct, between the linkers and the part, if the BASIC workflow is used. In grey are the amino acids encoded by the prefix and suffix if used in protein fusions.

Figure 35 shows the annotated sequences of the iP and iS. Their role in the BASIC workflow is to carry an inward-facing BsaI recognition site and its cutting site, where a 4 bp overhang is generated that guides the attachment of the linker to the DNA part, while the rest of the iP or iS falls off. The overhangs left in place of the iP and iS after digestion are different from each other in order to guide the ligation of the linkers specifically to the correct side of the part, but they are the same across every part so that the linkers can be reused universally within the standard (see chapter 4.2.4 for the full description of the linker oligonucleotides). The process of attaching the linkers to the part with

the BASIC workflow leaves a 4 bp scar between the part and the linker on the iP side, and a 6 bp scar on the iS side: this design is necessary in order for the iS side to encode for protein fusion-friendly amino acids.

The iP and iS can also be used to run the MODAL workflow: their full sequences can be used exactly like MODAL's prefix and suffix, as priming sites for the PCR in Step 1. They are optimised to promote reaction specificity while also at the same time encoding for a flexible protein linker, compatible with common protein fusion applications. In this case the scar that remains between the linker and the part is the full length of the iP or iS, 18 bp. Of course where using either workflow to generate protein fusions it is necessary to use linkers that also encode for adequate protein sequences, as the whole "iS-linker-iP" sequence that is left between the parts after assembly will constitute the protein bridge between the two proteins encoded by the parts. R2oDNA Designer can help designing such protein fusion-friendly linkers by verifying the quality of the sequences and choosing the best bases to introduce where more than one option is available (i.e. degenerate codons).

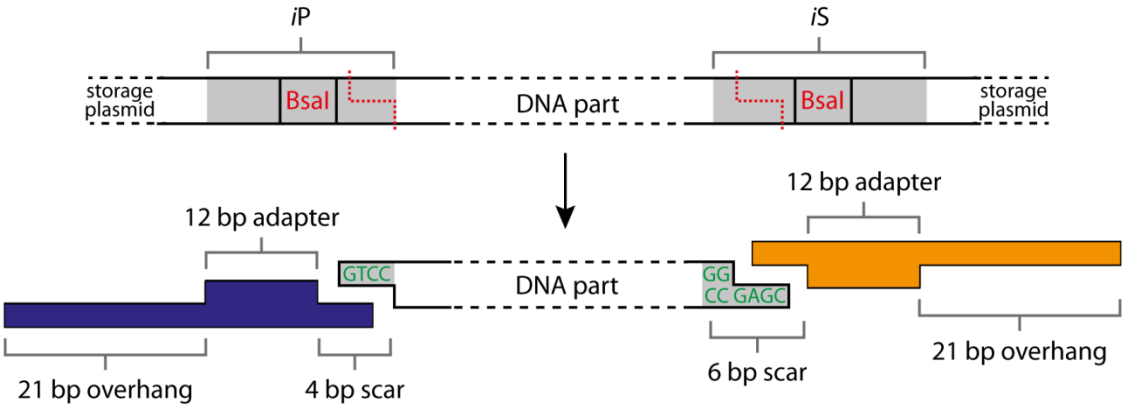


Figure 36: the diagram shows how the BASIC prefix and suffix work during assembly. Similarly to **Figure 35** in red arethe location of the BsaI recognition and cutting sites, and in green the scar sequences. The scar sequences also contain the 4 bp overhang that allow the attachment of the linkers to the parts. The linkers, in purple and yellow, have 4 bp 5' overhang that matches specifically the 4 bp overhang on one of the sides the parts, a central 12 bp double stranded region, and a 21 bp 3' overhang that guides the final assembly step.

4.2.2. Linkers design and specifications

The BASIC linkers are partially double stranded oligonucleotides composed of a short 12 bp strand and a long 37 bp strand, which anneal to generate a 4 bp 5' overhang and a 21 bp 3' overhang flanking the 12 bp double stranded region (**Figure 36**). They were designed starting from the MODAL "40% GC" linker set (expanded by generating three more linkers to suit the needs of the project) to facilitate comparisons with MODAL's results. As shown in **Figure 37**, the 45 bp sequence of the MODAL linkers was divided into a central 21 bp region, and two flanking 12 bp ones, which were used to compose the BASIC linkers.

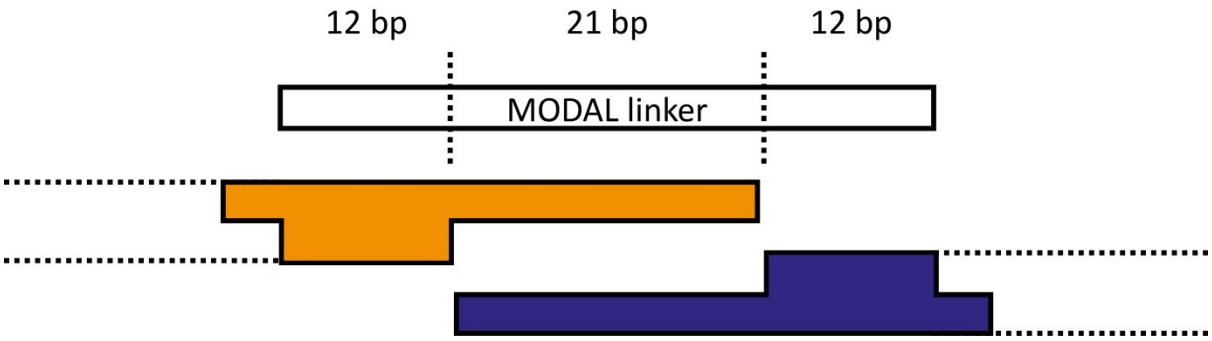


Figure 37: how the BASIC linkers were generated starting from the MODAL linkers. The 21 bp 3' overhang on the BASIC linkers corresponds to the central part of the MODAL linkers, while the 12 bp double stranded portions on cognate BASIC linkers complete the rest of the MODAL linker sequence. When two parts are joined using the BASIC workflow, the whole MODAL linker sequence is recreated between them.

The 4 bp 5' overhang guides the ligation of the linker to the correct side of the DNA part during Step 1 of the BASIC workflow, and thus needs to be phosphorylated. The 12 bp double stranded region is necessary to allow efficient ligation as it has been shown that T4 ligase requires at least 5 bp of double stranded DNA around a nick in order to be able to seal it efficiently[106]. The 21 bp 3' overhang finally directs the assembly of the parts during Step 2 of the BASIC workflow, where the final construct is built. The lengths of 21 bp for the single stranded region and 12 bp for the double stranded one were chosen to accommodate two needs: the single stranded regions need to be long enough to ensure efficient annealing of the parts at the final assembly (Step 2), and the double

stranded region needs to be long enough to ensure that the short strand of the oligonucleotide remains annealed to the long one while they are being attached to the DNA part (Step 1). A length of about 20 bp is commonly known to work well for target recognition and annealing of primers in PCR amplifications, which are reasonably similar to BASIC's Step 2 in terms of solution composition and annealing temperature: shorter sequences might have specificity problems or not bind strongly enough, while longer ones require more time to find their binding target for entropic reasons[107]. A length of 21 bp for the overhang allowed us to make the double stranded region 12 bp long, which we found to be enough to ensure the stability of the linker oligonucleotides. In order to enhance it further, these are prepared in advance by mixing the long and short fragment pairs in a high-salt buffer that promotes and stabilises annealing.

4.2.3. The BASIC assembly workflow

Step 0: the BASIC assembly workflow is modelled after MODAL's step-wise structure, and Step 0 is identical for both: any BsaI recognition site-free DNA fragment can be made compliant with the BASIC standard with a one-time procedure that attaches the iP/iS sequence respectively upstream and downstream of the part. The resulting standardised part is then cloned in a suitable storage plasmid. The particular method with which this is achieved can vary: for most DNA fragments this is most easily and quickly done by running a PCR amplification where the primers that amplify the desired part also carry tails that encode for the iP and iS. The product can then be purified and cloned in a pJET1.2 plasmid using a CloneJET PCR Cloning Kit (Thermo Scientific). Alternative methods could be necessary when dealing with parts that are difficult to amplify via PCR and these could include classic restriction/ligation cloning into blunt-cut vectors pre-prepared to have the iP/iS sequences at the respective ends. Additionally not all parts might be compatible with a single storage plasmid and alternative solutions might be necessary: for example the Ori part could be cloned into pJET1.2 because it carries the same replicon, PMB1, so we fused the part with the antibiotic resistance gene from pJET1.2 to produce a viable storage plasmid. Another example is the Ori+Kan part that contains PMB1 and a kanamycin resistance gene: we produced a viable storage plasmid by simply circularising it the part. Whatever method is used, after cloning the plasmid is isolated and the DNA part within is sequence verified.

Step 1: during Step 1 of the BASIC workflow the linker sequences are attached to the standardised part using a simultaneous restriction/ligation reaction (**Figure 38**). The storage plasmid containing the desired part is mixed with the appropriate pair of linkers in a solution containing BsaI and T4 ligase: BsaI cuts the iP and iS sequences generating 4 bp overhangs on the flanks of the part, which are recognised by the 4 bp overhang of the linkers. T4 ligase seals the nicks and covalently joins the linkers to the parts. The mix is incubated initially at 37°C for an hour to promote digestion of the storage plasmid, then 20 minutes at 20°C to promote linker ligation and finally 20 minutes at 65°C to inactivate the enzymes.

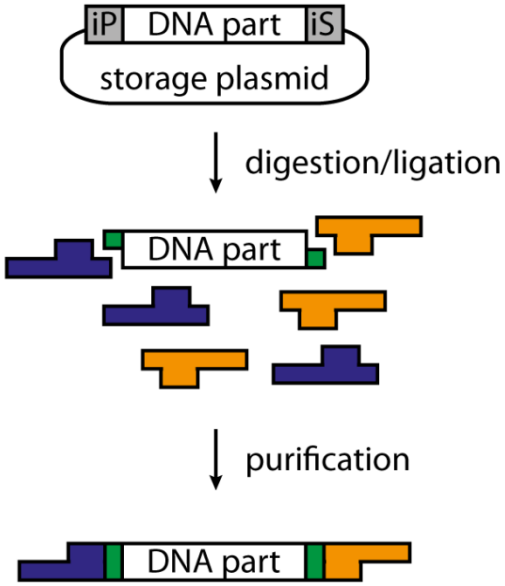


Figure 38: Step 1 of the BASIC workflow. A standardised BASIC part is prepared for assembly by attaching the linker sequences to it. This is achieved with a simultaneous digestion and ligation reaction that releases the part from the storage vector, exposing the 4 bp overhangs (in green) that specifically match those on the linker oligonucleotides. These anneal to the part and are permanently ligated to it. If the part is ligated to the storage plasmid backbone, the BsaI sites are reconstituted and the part can be cut and released again, driving the reaction to completion. Finally a magnetic beads-based PCR purification step is used to remove excess linker oligonucleotides from the solution.

It is important to note that both enzymes are active to some degree throughout the whole reaction, both at 37°C and 20°C, which helps drive the reaction to completion: the DNA parts that are ligated back to the storage plasmid reform the original BsaI recognition site and can be cut again,

while those that are ligated to the linkers do not. We observed that one of the critical factors in this reaction is the ratio of the concentrations of the two enzymes, and particularly an excess of T4 ligase compared to BsaI can significantly increase the number of colonies containing the storage plasmid that carries the antibiotic resistance part, instead of the desired final construct (unless a double selection design is employed, where the final construct contains two antibiotic resistance genes carried by two different storage plasmids). This is likely due to a large number of parts being ligated back to the storage plasmid backbones compared to those that are digested again. After the reaction unligated linkers are removed using a beads-based PCR purification kit (Agencourt AMPure XP). We also tested the column-based QIAquick PCR purification kit as a more commonly adopted substitute for the beads-based purification kit, but we found that it is not compatible with our design, as it decreased the number of colonies obtained to zero on most occasions. We hypothesize that this might be due to inefficient removal of the partially double stranded oligonucleotides.

Step 2: during the final assembly step (**Figure 39**) all the parts flanked by the appropriate linkers and purified are mixed together in a single tube with an ionic buffer and incubated for 45 minutes at 50°C. No enzymes are required since the BASIC linkers are already single stranded and available to spontaneously anneal to each other, leaving nicks that will be repaired by the cells. After incubation the solution can be used directly for cell transformation.

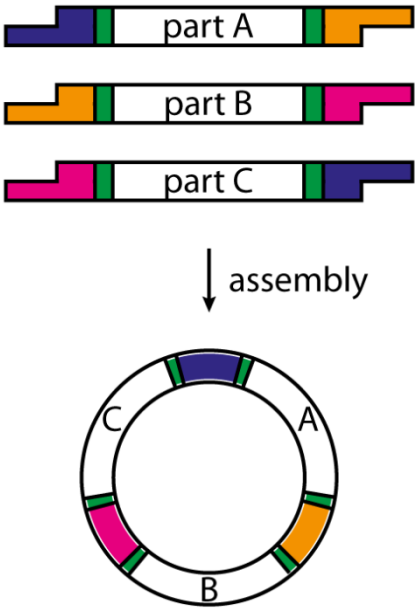


Figure 39: Step 2 of the BASIC assembly workflow. All the parts to be assembled, equipped with the linkers and purified during Step 1 are mixed in a single tube and annealed to each other, generating the final construct. The nicks are automatically repaired *in vivo* after transformation of the host.

Harry Trewhitt, a third year undergraduate student under the supervision of Dr. Geoff Baldwin, reported that performing the reaction at 37°C with T4 ligase gives a significantly higher number of incorrect assemblies, which might be due to inaccurate annealing of linkers caught in thermodynamic traps that may be stabilised by non-productive binding of T4 ligase. These incorrect junctions would be repaired by the cells after transformation and sequencing demonstrated that a single linker sequence is present, but joining incorrect parts. Harry Trewhitt also tested incubating at 50°C with the addition of Taq ligase, but found this gives no significant improvement over no-ligase reactions, which are then preferable because they are simpler and cheaper.

4.2.4. Evaluation of assembly efficiency

We tested the MODAL workflow by assembling eleven constructs of various sizes (**Table 5**) starting from eight different standardised parts that encode for easily detectable phenotypes, including antibiotic resistances and fluorescent reporters (**Table 4**). We determined the total number of colonies obtained as a measure of the efficiency with which parts are joined to each other to assemble plasmids, and calculated the percentage of colonies expressing the expected fluorescent reporters as an estimate of the accuracy of assembly.

Part name	Part size	Part function
Ori	615 bp	PMB1 origin of replication
Kan	977 bp	Kanamycin resistance gene
Ori+Kan	1667 bp	A single part containing both the “Ori” and the “Kan” parts
Cm	959 bp	Chloramphenicol resistance gene
P102	153 bp	Constitutive promoter
GFP	925 bp	GFP gene containing promoter, RBS, ORF and terminator
GFP ORF	881 bp	Same as the “GFP” part, but without promoter
RFP	912 bp	RFP gene containing promoter, RBS, ORF and terminator
RFP ORF	869 bp	Same as the “RFP” part, but without promoter

Table 4: list of the parts used during testing of the BASIC workflow.

Selection marker	Number of parts	Construct size	Construct composition	Expected colony colour
Single	2	2689 bp	Ori+Kan – RFP	Red
	3	2669 bp	Kan – Ori – RFP	Red
	4	3649 bp	Kan – Ori – RFP - GFP	Yellow
	5	3813 bp	Kan – Ori – RFP – P102 - GFP ORF	Yellow
	6	3978 bp	Kan – Ori – P102 – RFP ORF – P102 – GFP ORF	Yellow
Double	2	2736 bp	Cm – Ori+Kan	White
	3	3703 bp	Cm – Ori+Kan – RFP	Red
	4	3683 bp	Cm – Kan – Ori – RFP	Red
	5	4663 bp	Cm – Kan – Ori – RFP – GFP	Yellow
	6	4827 bp	Cm – Kan – Ori – RFP – P102 – GFP ORF	Yellow
	7	4992 bp	Cm – Kan – Ori – P102 – RFP ORF – P102 – GFP ORF	Yellow

Table 5: list of the constructs built during testing of the BASIC workflow.

The results in **Figure 40** show that the BASIC workflow is able to reliably assemble constructs of up to seven parts, with the efficiency decreasing exponentially as the number of parts increases: tens of colonies are produced for 6 -7-part plasmids, hundreds for 4-5 parts and thousands for smaller constructs. The accuracy of assembly varies greatly between single and double antibiotic

selection plasmids: in the former it decreases rapidly as the number of parts increases from >99% with two parts down to 27% with six, while in the latter it remains above 95% regardless of the number of parts involved.

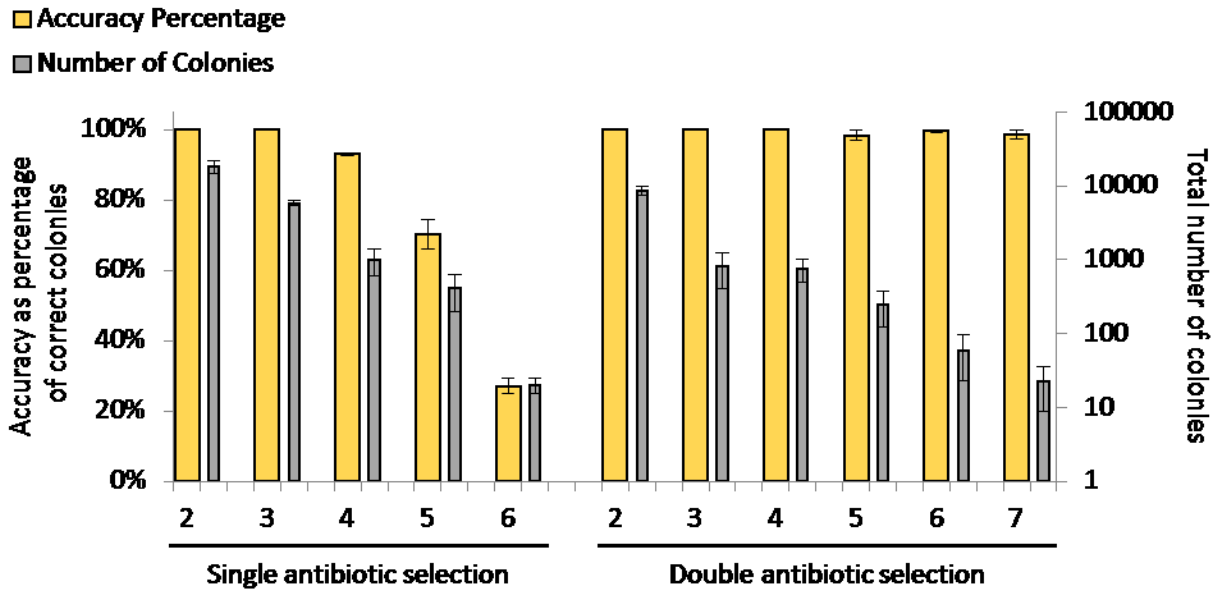


Figure 40: results of the BASIC efficiency test. The total number of colonies and the percentage of those containing correctly assembled plasmids (accuracy) were calculated from image analysis of each plate for the various assembled constructs (n=2). The numbers on the X axis correspond to the number of parts contained in the relative construct, as listed in **Table 5**. Assembly accuracy for the two-part double-selection construct could not be determined experimentally due to the fact that the correctly assembled construct does not contain any fluorescent reporter, and is thus indistinguishable from any other non-fluorescent plasmid. It is estimated to be very close to 100% due to the fact that all other double selection constructs showed near-100% accuracy. Error bars indicate standard error.

We investigated the mechanisms involved in determining the accuracy of assembly by looking at the incorrect colonies produced in the experiment. **Figure 41**, based on the same data set as **Figure 40**, shows the number and colour of the incorrect colonies for each assembly, where the colour correlates with the fluorescent reporters expressed. The experiment was repeated twice and the results show a high degree of variability, possibly due to the fact that cell transformation steps tend to be very sensitive to a large number of factors such as little variations in the incubation times, or how recently the competent cells were prepared, *etc*. The experiment was also repeated twice by

Dr. Marko Storch, a postgraduate research assistant in Dr. Geoff Baldwin's group, whose data set shows a higher level of variability too, but still seems to agree with ours. The results show that coloured incorrect colonies (the result of misassembled plasmids lacking one of the fluorescent markers) are quite rare, while the majority is white: these can be produced both by misassembled plasmids lacking all fluorescent markers and by carryover storage plasmids containing the antibiotic resistance part, which are viable in the selective medium used for the final transformation of single selection plasmids (but not for that of double selection plasmids, because no storage plasmid carries both antibiotic resistance genes).

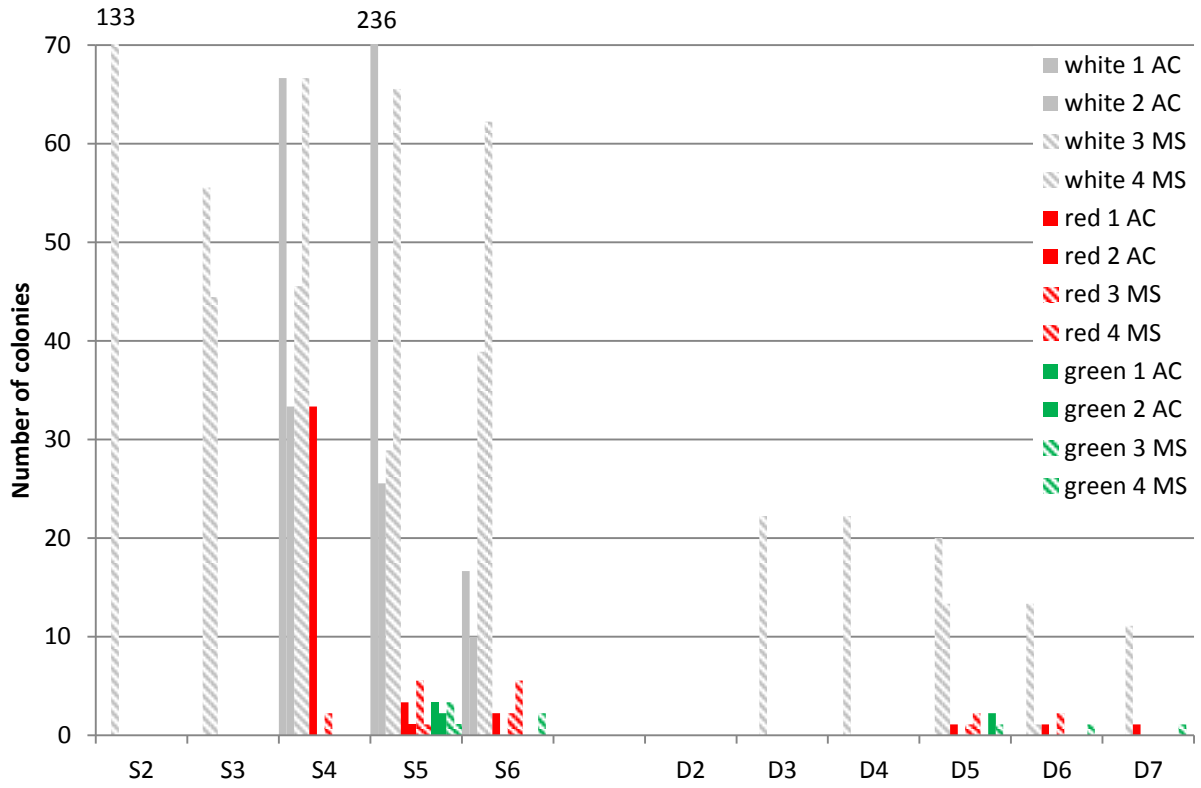


Figure 41: results of the BASIC efficiency test. The plot uses the same data set as **Figure 40** but here are shown the number and colour of the incorrect colonies that were obtained after the assembly of each construct. On the X axis again are the various test constructs (S or D stands for single or double antibiotic selection, the number indicates how many parts compose the construct). The colour of the colonies depends on the presence and type of fluorescent reporter in the incorrect plasmids, refer to **Table 5** for the expected colour of the correctly assembled plasmids. The data has been gathered in four replicates, shown separately: 1 and 2 by the author of this work Arturo Casini (AC), and 3 and 4 by Dr. Marko Storch (MS). Data is shown respectively as full and barred for easier distinction. To facilitate visualisation two data points are left off the scale and their values are shown on top of their bars. This plot shows that the majority of the incorrect colonies are white, an indication that they could be a result of storage plasmid carryover rather than incorrect assembly.

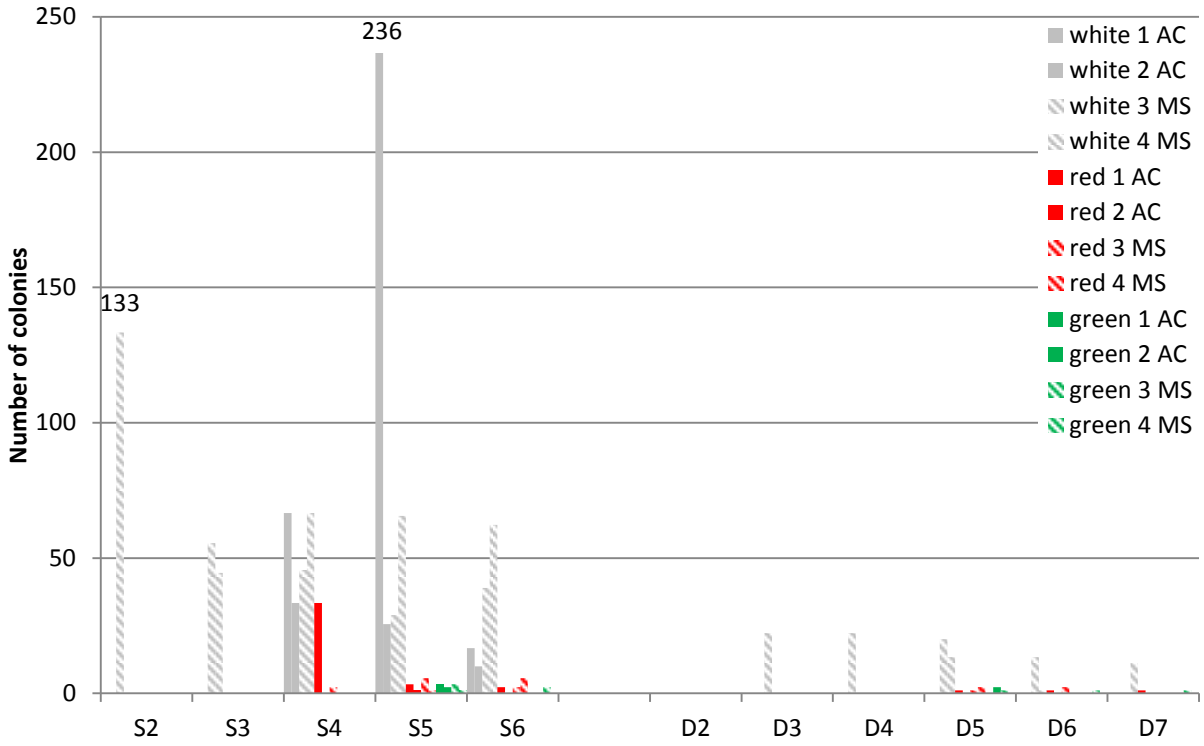


Figure 42: same as **Figure 41** but with adjusted Y axis scale to include the data points that were not shown previously.

The same data set as **Figure 40** can also be looked at in terms of number of correct and incorrect colonies per number of parts (**Figure 43** and **Figure 44** respectively): the data reveals that these have a markedly different trend. As the number of parts per plasmid increases, the number of correct colonies seems to fall exponentially while the number of incorrect colonies seems to fluctuate independently, remaining more or less constant. This suggests that the two are generated by two different mechanisms, one that is influenced by the number of parts being assembled simultaneously and one that is not.

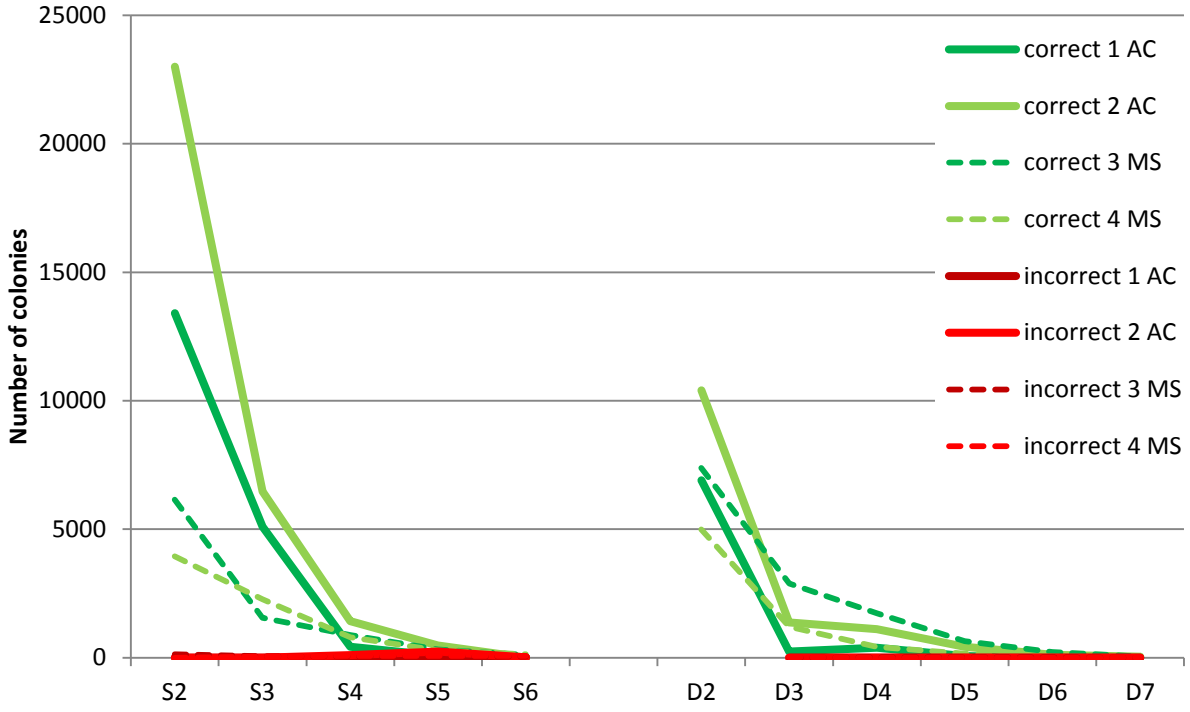


Figure 43: results of the BASIC efficiency test. The plot uses the same data set as **Figure 40** but shows the total number of correct colonies for each sample (green lines). Total number of incorrect colonies is also shown as transparent red lines for scale comparison (see **Figure 44** for greater detail). On the X axis again are the various test constructs (S or D stands for single or double antibiotic selection, the number indicates the number of parts that compose the construct). The data has been gathered in four replicates, shown separately: 1 and 2 by the author of this work Arturo Casini (AC), and 3 and 4 by Dr. Marko Storch (MS). Data is shown respectively as continuous and dashed lines for easier distinction. The plot shows that the number of correct colonies decreases in an exponential-like fashion as the number of parts being assembled simultaneously increases.

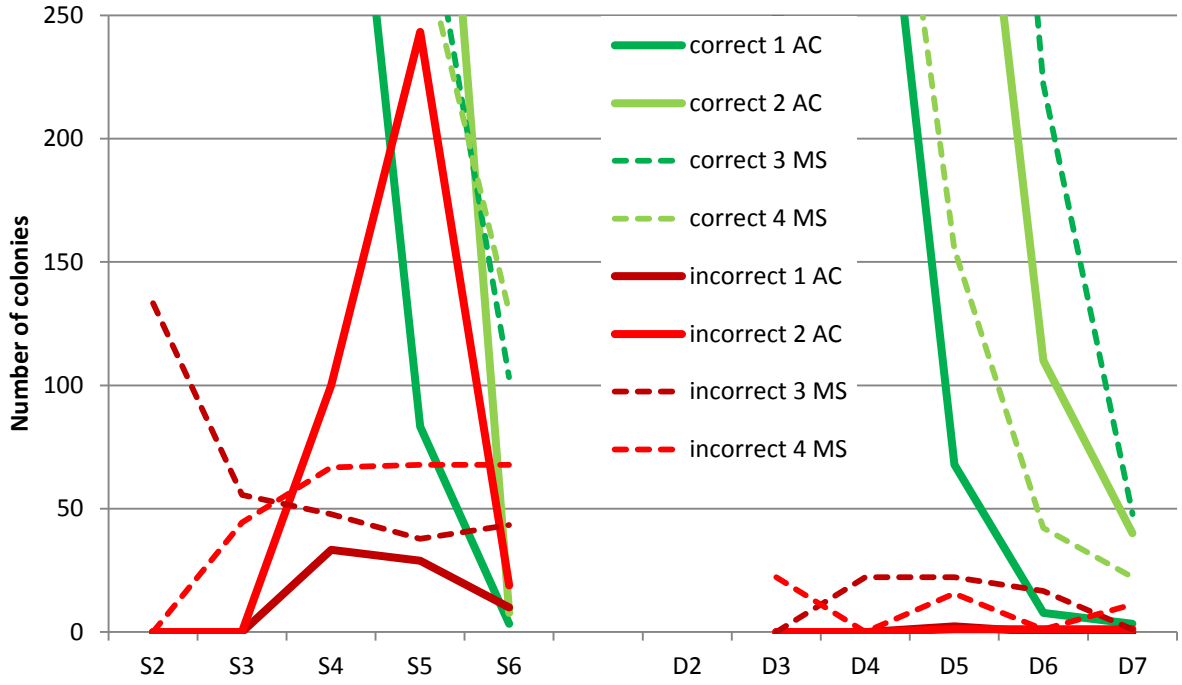


Figure 44: results of the BASIC efficiency test. The plot uses the same data set as **Figure 40** but shows the total number of incorrect colonies for each sample (red lines). Total number of correct colonies is also shown as transparent green lines for scale comparison (see **Figure 43** for greater detail). On the X axis again are the various test constructs (S or D stands for single or double antibiotic selection, the number indicates the number of parts that compose the construct). The data has been gathered in four replicates, shown separately: 1 and 2 by the author of this work Arturo Casini (AC), and 3 and 4 by Dr. Marko Storch (MS). Data points are shown respectively as continuous and dashed lines for easier distinction. The plot shows that the number of incorrect colonies fluctuates independently of the number of parts being assembled simultaneously, unlike the number of correct colonies.

The presence of storage plasmids among the white colonies in the single selection assemblies was also confirmed by diagnostic PCR: we prepared the plasmid DNA from a few colonies and ran a PCR using the pJET1.2 sequencing primers from the commercial kit which would only give a product when used on a pJET1.2 plasmid.

4.2.5. Implementing idempotency to allow hierarchical assembly

BASIC can be adapted to perform idempotent assembly, which means that any newly assembled construct will have the same standard format as any BASIC part (i.e. will be flanked by the iP/iS sequences) and will be reusable in further assemblies using the same workflow. The only adaptation that is required is the use of two pairs of special linker oligonucleotides (linkers A and B, **Figure 45**) that respectively restore the iP and iS sequences in the final construct. As shown in **Figure 46**, in the final construct these linkers would be located between the area that contains all the parts constituting the new "composite part", and the area that contains the "backbone" genes for plasmid replication and survival. The process of creating a plasmid containing a "composite part" and a "backbone" is essentially equivalent to inserting a "part" in a "storage plasmid" as it happens in Step 0 of the workflow, hence the idempotency of the process. Since linker A and B necessarily contain a BsaI recognition site that would be cut during Step 0 of the workflow, disrupting the assembly process, we decided to protect the recognition site through methylation. The BsaI restriction modification system normally employs a C-5 methyltransferase, but its target within the BsaI recognition sequence is not known[108]. Dr. Marko Storch determined that methylation of either C in the top strand effectively protects the DNA from digestion, so we decided to test linkers methylated on the first one.

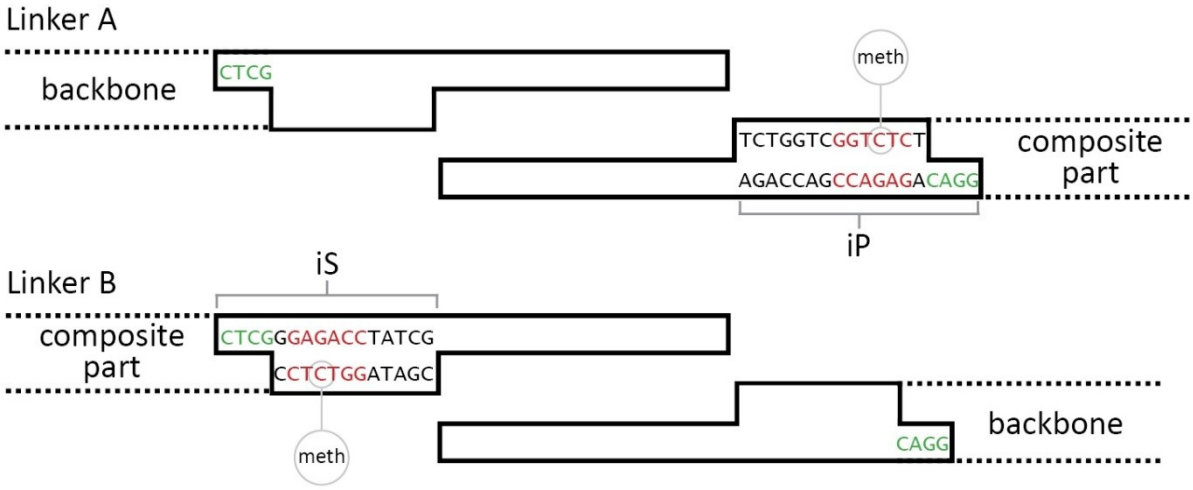


Figure 45: schematic of the methylated linker oligonucleotides employed for hierarchical assembly in the BASIC workflow. When assembling intermediate constructs for hierarchical assembly, the plasmids contain two or more parts that represent the composite part to be carried forward in the next round of assembly, and one or more backbone parts. Special linkers, called linker A and B, are used to join the parts at the ends of the composite part with the backbone. These special linkers contain the full iP/iS sequences, but with methylated BsaI recognition sequences (in red) so that they cannot be cleaved during the assembly of the intermediate plasmid. After transformation of this intermediate construct methylation is spontaneously lost due to normal *in vivo* replication, so that the plasmid can be isolated and used like a normal storage plasmid in a new round of assembly.

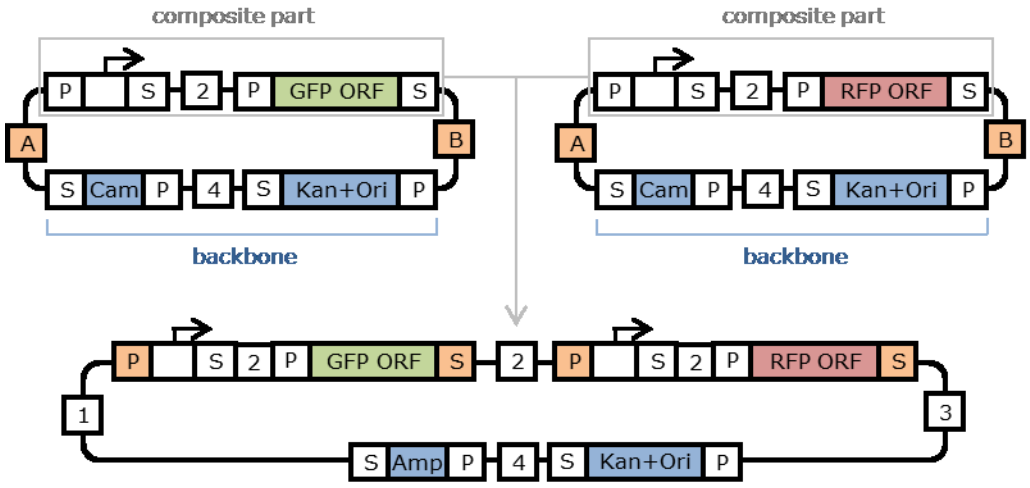


Figure 46: diagram of hierarchical assembly with the BASIC standard. The intermediate constructs contain a backbone (blue) and the composite part that will be carried forward to the next round of assembly. In the example the intermediates are 4-part plasmids that contain a composite part made of a promoter part and a fluorescent reporter part. Linkers A and B (orange) are located between the two and contain the iP and iS sequences. Once these are assembled and isolated, they are used like a normal storage plasmid in the next round of assembly. In the example the final construct contains 6 parts but is assembled from 4: two composite parts and two backbone parts.

We tested idempotent assembly by building two plasmids, one containing a GFP gene and the other an RFP gene (**Figure 46**), both composed of two separate BASIC parts: the promoter and the ORF. We assembled them using linkers A and B between the reporter genes and the backbone, and also using normal linkers, in order to investigate whether methylated linkers have any negative impact on the assembly process. The results in **Figure 47** show that there is essentially no difference in number of colonies or assembly accuracy between normal and methylated linkers, and these results are also compatible with our previous results for four-part plasmids with double antibiotic (**Figure 40**). Final proof that hierarchical assembly work flawlessly was obtained by Dr. Marko Storch, who verified that these plasmids are effectively equivalent to any BASIC storage plasmid: since linkers A and B, in orange, restore the iP and iS sequences in the first-level plasmids they can be used directly in a further round of assembly with the normal BASIC workflow to build a second-level plasmid as shown in **Figure 46**. Dr. Marko Storch also used double antibiotic selection, changing one type of resistance compared to the plasmids carrying the composite parts (chloramphenicol for

ampicillin), in order to prevent the first-level plasmids from being viable in the transformation medium of the second-level plasmids, which would cause background transformation.

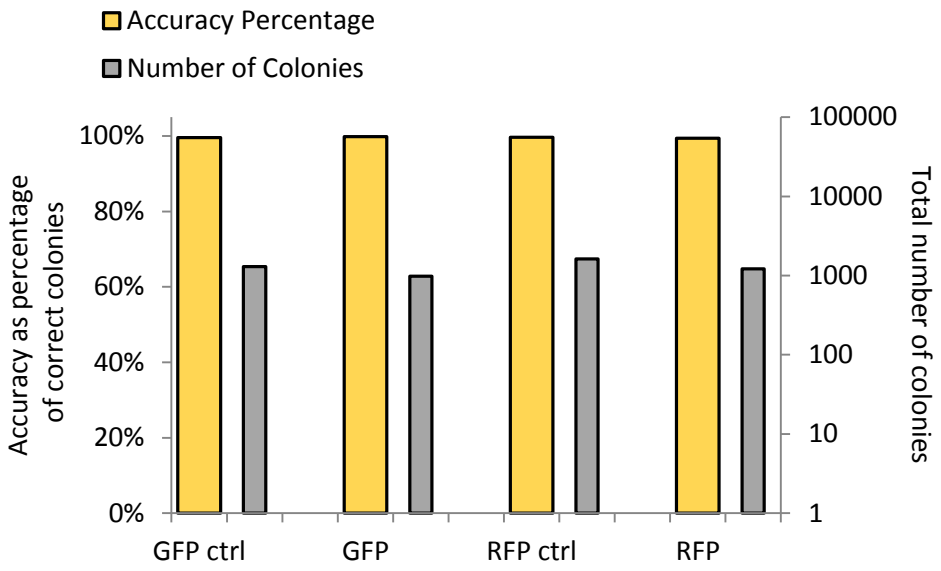


Figure 47: testing the effect of linker A and B on assembly efficiency. The total number of colonies and the percentage of those containing correctly assembled plasmids (accuracy) were calculated from image analysis of each plate for the various assembled constructs (n=1). The X axis indicates the constructs assembled, whose structure is shown in **Figure 46**: they are all 4-part plasmids containing a part with kanamycin resistance and PMB1 origin of replication, a chloramphenicol resistance part for double selection, a constitutive promoter part and either a GFP or RFP ORF part. The first two represent the backbone, the second two the composite part. Those named GFP ctrl and RFP ctrl have been assembled using normal linkers (1-4). Those named GFP and RFP have been assembled using linker A and B between the backbone and the composite part. The results show no difference in assembly efficiency, suggesting that linker A and B and behave like normal BASIC linkers: methylation effectively prevents BsaI cleavage and does not seem to affect the assembly reaction in any way.

4.3. Discussion

4.3.1. An issue with background transformation reveals BASIC assembly's high accuracy

Accuracy is an extremely important factor for DNA assembly methods, as the higher the accuracy, the less effort is required to screen for correctly assembled constructs. The difference in assembly accuracy between single selection plasmids and double selection plasmids shown in **Figure 40** is quite marked: for the former it decreases very fast as the number of parts increases, while for the latter it remains constant and very high. This prompted us to investigate its causes, so we took a closer look at the number and colour (which is representative of the fluorescent reporter genes present in the construct) of the incorrect colonies for each sample. **Figure 41** shows that in all samples the majority is white, which can be produced either by misassembled plasmids lacking all fluorescent reporters or by carryover storage plasmids containing an antibiotic resistance part (background transformation). It is important to note that the second case can only happen when building single antibiotic selection plasmids, since when building double selection plasmids cells are grown in the presence of two antibiotics and there is no single storage plasmid that carries all the necessary antibiotic resistances to produce viable colonies. The results in **Figure 41** seem to indicate that most of the white colonies are produced through the second mechanism, since there is a high number of white colonies in the single selection samples and none in the double selection samples. Additional repeats of this experiment performed by Dr. Marko Storch show a similar trend, with the number of white colonies remaining very low for the double selection plasmids, while being consistently higher for the single selection ones.

Similar conclusions can be drawn from **Figure 43** and **Figure 44**, which indicate that correct and incorrect colonies might be generated through different mechanisms: only the number of correct colonies seems to depend on the number of parts being assembled, as it is expected, since the probability of all the parts coming together is inversely proportional to the number of parts. The number of incorrect colonies on the other hand does not seem to depend on the complexity of the

final construct, and this is consistent with the hypothesis that most of the incorrect colonies are produced by carryover of undigested storage plasmids containing the antibiotic resistance part, as this effect is not influenced by the number of parts being assembled. This would also explain the rapid decrease of accuracy of the single selection constructs: as the number of parts increases the number of correct colonies decreases (due to the increasing number of simultaneous annealing reactions required to make the construct) while the number of white background colonies remains constant and high. The double selection constructs do not show the same trend because the number of background colonies remains zero or close to zero.

These observations seem indicate that most of the incorrect colonies produced by the BASIC workflow do not contain plasmids made of incorrectly joined parts, but carryover storage plasmids. If incorrect colonies were produced mainly by misassembly (incorrectly joined parts) we would instead expect to see a higher number of coloured incorrect colonies, a smaller difference in accuracy between single and double selection assemblies and an increase in incorrect colonies as the number of parts in the constructs increases. This means that the method BASIC employs to join DNA molecules, which is simply annealing DNA fragments flanked by single stranded optimised linker sequences at a high temperature, is extremely accurate: it yields over 95% correct colonies when assembling up to seven parts simultaneously.

4.3.2. BASIC aims to overcome MODAL's limitations

The BASIC workflow is based on the same structure as MODAL, but overcomes some of its limitations especially in terms of reliability. In Step 1, where the linker sequences are attached to the part, MODAL employs a PCR amplification. PCRs are a very unpredictable type of reaction: they have a chance of introducing mutations and they might have difficulties amplifying certain sequences (e.g. GC rich or repetitive sequences). Coping with these problems requires expensive and time consuming processes such as sequence verification and gel extractions. BASIC adopts instead a simultaneous restriction/ligation reaction, inspired by the methods that were used by researchers to introduce new restriction sites or sticky ends on DNA fragments at the times when PCR had not been invented yet[109]. This approach does not have any of the issues PCRs have, and has been widely adopted in synthetic biology and molecular biology in general for a variety of purposes because of its simplicity, precision and reliability: the BioBrick assembly method is a famous example of how many researchers prefer an old and simple but extremely reliable technique to more advanced and powerful ones[110]. In addition to this, simultaneous restriction/ligation reactions have been adopted in a very successful series of DNA assembly techniques and standards, which can boast very high precision and reliability[32,33,74,75]. The only limitation the adoption of this kind of reaction brought to BASIC is that it requires the absence of the 6 bp recognition sequence of the BsaI restriction enzyme from all DNA parts. While this is a significant limitation, we believe it is easier to cope with this than with the limitations and issues brought by the use of PCR, because it is a more predictable limitation: one can easily find out which sequences contain a BsaI recognition site, while one can at best guess which sequences might be difficult to amplify with a PCR amplification. We believe this in itself is already a big help but in addition to this removing unwanted BsaI sites is almost certainly easier than changing the GC content of a sequence or removing secondary structures, since the former only requires the modification of a single base pair, while the latter requires more extensive changes and it is hard to predict whether they will be sufficient to fix the issue. Finally, although we did not test it, it might be possible that this limitation will be also alleviated by the fact that BsaI recognition sites

will only be problematic if, when cut, they generate the same overhangs generated in the iP/iS or in other undesired BsaI sites on the same part. This is because this scenario would cause these cut areas to ligate to the iP/iS or to the linkers or to other fragments of the part, while if they generate unique overhangs they will simply be ligated back to themselves. This would likely lower the overall efficiency of Step 1, but in theory it should not completely break it or compromise its accuracy.

Another significant improvement of BASIC over MODAL is in Step 2, where the parts that are prepared and purified in Step 1 are assembled together, guided by the linker sequences. MODAL can employ virtually any long overlap-based DNA assembly technique to perform this step, which gives it great flexibility and the ability to adapt to specific requirements, but even though we tested three of the best techniques available none of them was completely satisfactory. We mentioned before (**Chapter 4.3.1**) that accuracy of assembly is definitely one of the most important factors, especially concerning the overall reliability of the assembly workflow, but when we tested MODAL to assemble plasmids made of four 1 kb parts (**Chapter 3.2.2**), a commonly used format, only yeast recombination achieved >90% accuracy. Unfortunately yeast recombination is also the most unwieldy of the techniques we tested: it requires up to four days to grow colonies after transformation, as opposed to only one for bacterial transformations, and it is very difficult to prepare plasmids from yeast cells, making it difficult to move them in different hosts or perform hierarchical assembly. The other techniques, Gibson and CPEC, are faster but achieve lower accuracy and employ expensive enzyme reagents like DNA polymerases, which also have a chance of introducing sequence errors. Moreover Gibson is known to be very inefficient at assembling small (<200 bp[111]) parts, and CPEC, besides being the least efficient technique, also carries all the limitations of PCR, since it is essentially based on the same principle.

BASIC instead uses a single method in Step 2, specifically designed to continue along the same themes of the overall workflow: simple, robust and predictable reactions. The method is based on the spontaneous annealing of the parts which are already equipped with single stranded linkers, at

an appropriate temperature to minimise unspecific duplex formation. Since there are no enzymes involved the reaction is extremely simple and cheap to prepare and the annealing mechanisms of PCR primer-length DNA strands are also very well studied, making the process extremely predictable. The simplicity of the step makes it also intrinsically very robust to changes in incubation times and temperatures that might be caused by human error. It compares very positively to yeast recombination as well, not only because it employs much faster bacterial transformations, but also because it achieves similar accuracy: double selection plasmids obtained more than 90% correct colonies using parts ranging from 153 to 1667 bp in length to build constructs of up to seven parts. The number of colonies obtained is also similar: when assembling four-part plasmids the best MODAL results, both by yeast recombination and Gibson assembly, were around the 10^3 mark, and BASIC obtained essentially the same numbers of colonies for the same number of parts. While we do know that yeast recombination, with all its limitations, can produce colonies with much higher number of parts[54], observations by the authors and various discussions with colleagues indicate that the efficiency of Gibson isothermal assembly decreases very rapidly as the number of parts increases. Five parts normally being reported as the highest number of parts that can be routinely be assembled with any reliability[112], even though it is likely that the use of computationally designed linkers would improve this (see **Chapter 5.2**). Finally, BASIC is specifically designed to support hierarchical assembly: all that is required is the use of two special pairs of methylated linkers (linker A and B) that reconstitute the iP/iS sequences in the final construct, effectively conferring it the same format as any storage plasmid. Since they are transformed into bacterial cells, these first-level plasmids can be easily isolated in large amounts and used for further rounds of assembly with no modifications to the workflow.

4.3.3. Conclusion

We developed here BASIC, a new standard for DNA assembly that includes both a physical format for DNA parts and a method that allows reliable, accurate and efficient assembly of standardised parts into new DNA constructs. The standard format allows a high degree of flexibility in the design of new constructs and in the choice of methods to build them: similarly to MODAL, new standard parts can be obtained by simply adding two short sequences, called integrated prefix and suffix (iP/iS), at the flanks of any DNA fragment and cloning them in an appropriate storage plasmid. The iP and iS sequences allow these parts to be assembled together using either the MODAL workflow, with all the advantages described in **Chapter 3**, or the BASIC workflow described in this chapter. The BASIC workflow retains the same stepwise structure and the use of linkers that constitute the core of MODAL, but introduces a series of radical changes and improvements that aim to obtain a significantly more accurate and reliable process. At Step 1 this is achieved by moving away from PCR amplifications and employing restriction/ligation reactions instead, while at Step 2 we chose to use a simple *in vitro* annealing step instead of a pre-existing DNA assembly method. Step 0 remains instead identical, except for using the newly designed iP/iS sequences. In order to expand the range of applications with which BASIC is compatible, the iP/iS sequences are also designed to be compatible with protein fusions by encoding amino acids commonly used in flexible protein arms, regardless of which assembly workflow is being used. Another important improvement of BASIC over MODAL is the possibility of performing hierarchical assembly by utilising specifically designed methylated linker oligonucleotides that allow new constructs to retain the exact same format as any standardised part: these composite parts can then easily be reused as normal parts to build larger constructs. This brings a number of advantages, such as the possibility of recycling previously assembled constructs in new constructs, or the possibility of progressively assembling very complex constructs through a hierarchical series of simpler constructs.

4.3.4. Future work

Future improvements to BASIC would first of all have to tackle the background transformation issue: using double antibiotic selection is a very efficient way of making all storage plasmids non-viable while keeping the workflow unmodified, but it has the drawback of requiring the assembly of an additional part. As our data shows, increasing the number of parts seems to cause an exponential decay of the number of colonies obtained, so it is worth exploring other options. A possible avenue is making the storage plasmids toxic for the cells one transforms the products of BASIC into, which can be obtained in a number of ways. A commonly used approach is introducing a restriction enzyme gene in the storage plasmid backbone, and then having these plasmids replicate in cells equipped with the cognate methylation enzyme gene in the genome. This way the storage plasmids would kill any cell able to express the restriction enzyme but not equipped with the appropriate methylase, which is the large majority of *E. coli* strains. Unfortunately there is a possibility that the presence of a good number of suicide plasmids in the transformation mix will kill some of the cells that also received a correctly assembled plasmid, lowering the overall efficiency of the process.

Another approach that is similar but easier to implement would be to introduce a LacZ gene instead of a lethal gene in the storage plasmid backbone, so that any colony containing a storage plasmid would appear blue on the transformation plates in the presence of X-gal, making screening very easy. This would not require any genomic modification in the cells hosting the storage plasmids, but the downside would be that this would not stop these colonies from growing in the BASIC transformation plates. They would probably largely outnumber correct colonies in very difficult assemblies, which might have undesired effects such as slowing or impeding the growth of cells containing the desired construct by taking resources away from them. Their presence might also be problematic for automated colony-counting software or colony-picking robots.

Background transformation could also be fixed by not utilising plasmids to carry the antibiotic genes, so that they could not be replicated and maintained in cells. This could be achieved by

introducing a sort of “Step 0.5”, where the antibiotic genes flanked by the iP/iS are amplified from the storage plasmids using a PCR amplification, purified, and then used in the BASIC workflow as normal. Unfortunately this brings some degree of unreliability in the workflow, since PCR amplifications have a chance of introducing mutations in these genes, and linear fragments are not as stable as circular plasmids, so they would need to be periodically re-prepared to avoid problems.

Finally background transformation could be avoided by removing the promoter from the antibiotic resistance genes in the storage plasmids, and reintroducing it in the final constructs by placing them in the linker upstream. The downside is that this would complicate the standard, since for example MODAL-compatible resistance genes would need to be equipped with promoters, making these parts compatible with one of the workflows only, special linkers would need to be generated for this purpose, and great care should be taken in making sure that the antibiotic resistance is not expressed on the storage plasmids through other promoters present in the backbone.

Another issue that needs to be addressed by future work is assessing the impact on assembly efficiency and accuracy of the presence of unwanted BsaI recognition sites within the parts. It is clear that they will prevent the assembly from running correctly if they generate overhangs compatible with those of the linkers, of the iP/iS or of any other unwanted BsaI sites present on the same part, as during Step 1 this would generate truncated or mutated parts. On the other hand, if they generated unique overhangs, it is possible that they would simply be ligated back to themselves, reconstituting the original part, without influencing the normal linker ligation at all. It is also possible that some of these parts would be religated in time, leaving them cut and preventing the formation of circular products during Step 2, thus reducing the overall efficiency of the workflow.

Future improvements to BASIC will also explore the possibility of using the linkers as composable parts, by introducing functions encoded by short sequences inside them, such as RBSs, promoters,

RNAse recognition sites, terminators, *etc*. This would be especially useful with RBSs since need to be no more than a few base pairs away from the beginning of the relative ORF, and they are essentially impossible to isolate from the context: anything up to about 30 bp upstream of the ATG is known to influence translation initiation efficiency[98,113]. Of course it will be important to make sure that the sequences introduced in the linker oligonucleotides are compatible with their requirements: the double stranded region needs to bind strongly enough to remain double stranded during Step 1, and the long single stranded region needs to have all the features necessary to ensure accurate and efficient assembly during Step 2. R2oDNA Designer can help designing such sequences by evaluating their quality and choosing the best bases to introduce in the positions where there is freedom of choice.

5. Discussion

5.1. Two worlds combined

We have presented here a new approach to DNA assembly, developed through two different standards and workflows: we defined MODAL first, which was then expanded and improved in BASIC. We also developed Linker, a script to help designing the necessary homology regions, which was the basis for R2oDNA designer, a more advanced web-based tool that we helped design. Even though MODAL and BASIC are different in the details of the reactions they employ, they share certain basic principles and goals.

The main motivation for this project comes from our everyday experience in the laboratory: we noticed that when cloning most of the researchers were using either BioBrick assembly or Gibson isothermal assembly. Those using the first liked the simplicity and the predictability of the system: the standard is easy to use, it always works, and it is idempotent, so that anything assembled with it can be re-used in further assemblies with the exact same process. Unfortunately, however, it is also slow and laborious, since it only allows pairwise assembly and often requires gel extractions. Those who preferred Gibson isothermal assembly on the other hand were looking for the exact opposite: a fast and powerful method which can assemble many parts simultaneously. The downside here is that it requires *ad hoc* designing of the cloning process, of the primers and of the PCR amplifications involved, which are all prone to cause unexpected problems.

This project aims to combine the best of the two worlds: the confidence and simplicity of BioBrick assembly with the speed and power of Gibson assembly. At the same time we realised that DNA assembly is inevitably becoming a commodity, just like DNA oligonucleotide synthesis, which means that it needs to stop being one of the main daily duties for many scientists, and be instead carried out by a machine. We believe that automation will play a major role in DNA assembly, so we aimed to make our DNA assembly strategies as compatible with automation as possible, by keeping two things in mind: the first is that we should only use reactions and protocols that a liquid handling

robot can prepare and run (which for example excludes gel extractions), and the second is that we should avoid reactions and protocols that require a lot of *ad hoc* adjustments, which may be easy for a human to perform, but very hard or impossible for a robot.

We believe that there are three parameters that need to be taken into consideration in order to evaluate the success of our approach compared to other currently available methods and standards. The first is efficiency, which is essentially how good the method is at joining pieces of DNA correctly. The second is flexibility, which is the ability of adapting to different needs without having to make substantial changes to the standard workflow. The third is reliability, which concerns the likelihood of encountering unexpected problems during any stage of the assembly process.

5.2. Efficiency

Assembly efficiency is not only determined by the number of colonies obtained after transformation, but also by the percentage of these that contains a correctly assembled construct. These factors are equally important because, for example, producing a high number of correct colonies but mixed with a high number of incorrect ones will require extensive post-assembly colony screening, which is expensive and time consuming. On the other hand an assembly process that has very high accuracy but low colony yield can be problematic because, for example, difficult assemblies might yield no colonies at all, and some applications such as library cloning or combinatorial assembly require the recovery of many colonies to cover a good portion of the library or of the combinatorial space.

5.2.1. Homology region optimisation and DNA assembly efficiency

Assembly efficiency is inevitably highly dependent on the number and size of the parts being assembled, but there are many other factors, including the type of reaction and the nature of the homology regions (which are in turn usually determined by the standard adopted) that also play a crucial role. As mentioned in **Chapter 1.5** it is widely known that the features of the sequences involved play an extremely important role in the outcome of many molecular biology reactions, such as PCR amplifications and oligonucleotide microarrays. In DNA assembly the same issue has only been investigated properly for what concerns 4 bp sticky ends, while there is much less work done on longer homology regions.

The two main publications that tackle problem of designing optimised linker sequences are Guye *et al.*[78] and Torella *et al.*[79], and they both take an approach very similar to what we did with the Linker script (and later in the R2oDNA Designer collaboration): they designed a software tool that randomly generates a pool of sequences and then refines it according to a set of rules, to arrive eventually to a small but high quality set.

There is some degree of consensus regarding the optimisation rules, with three main points that both us and the other two publications cover: the first regards the general features of the sequences, which define the way they are expected to behave in the assembly reactions, such as length, GC content and melting temperature. The second aspect is about the ability of these sequences to anneal incorrectly to themselves or to other sequences (that are not their intended targets), in ways that prevent them from behaving correctly during assembly. The third point concerns the behaviour of the sequences after assembly, such as the presence of undesired restriction sites, functional regions like promoters, RBSs, transcription initiation sites, and regions that are too similar to the host's genome.

While all three tools cover these aspects, the level of attention that they reserve for them varies significantly: the one by Guye *et al*. focuses more on preventing undesired annealing, leaving the first point quite relaxed and ignoring the third almost completely. Torella *et al*. on the other hand do the opposite: they aim to generate linkers that can also act as biological insulators so they give the third point a great deal of attention, but employ quite superficial checks for incorrect annealing. Our tool takes a middle ground for the second and third points: regarding annealing, we have a comprehensive list of different type of undesired annealing events that we screen for, similar to Guye *et al.* (**Figure 19**), but our algorithm does not employ stricter checks for the annealing of the terminal regions of the linkers. Regarding the third point and the elimination of functional motifs, while we do use a BLAST search against the host genome and look for forbidden restriction sites, transcription initiation sites and RBS-like regions, we do not employ any external computational tool to find promoter-like regions like Torella *et al.* do. Also, differently from both other tools, we use much stricter rules for the first point, regarding the GC content, which in our tests turned out to be an extremely important and sensitive variable.

Unfortunately it is very difficult to compare the quality of the linker sequences generated by the three tools, or establish whether those by Torella *et al.* and Guye *et al.* are better than using non-

optimised sequences. Even though we have all used these optimised sequences with the same technique (Gibson isothermal assembly), none of the other papers includes optimisation tests to find the best parameters, or a comparison of the performance of their linker sequences against controls in assembly reactions. Comparing the assembly efficiency data they report is not very fruitful because both the number of colonies produced and the accuracy of assembly are heavily influenced by a large number of factors that differ between the three studies, such as cell competence, transformation protocol adopted, final construct size, DNA purification methods used, amount of DNA in the assembly reaction, sequence similarity between the fragments, *etc.*

Nevertheless our results (**Chapter 3.2.2**) prove that our optimised linkers are beneficial both in terms of accuracy and of number of colonies, and it is very likely that is true for the other two studies as well. Both the Gibson assembly commercial kit's manual from NEB and personal experience, confirmed by discussions with many colleagues, suggest that this technique usually struggles when assembling more than five DNA fragments simultaneously, while both Torella *et al.* and Guye *et al.* demonstrate highly efficient parallel assembly (>80% correct colonies) of five and eight fragments respectively: this discrepancy might be due to the use of optimised linker sequences.

It is also important to note that while ours is the only work where linkers are successfully tested across different techniques, suggesting that their benefits might extend to different types of reactions, it is likely that this applies to the other two works as well, due to the similarity of the design rules adopted. Lastly there is another advantage to the use of these optimised sequences for DNA assembly, beyond obtaining an increment in efficiency, which is reducing the occurrence of unexpected and often reaction-breaking problems that can be caused by human error or by the presence of undetected emergent features in the homology sequences. Delegating their design to a computational tool that generates homogeneous results is likely to be very helpful in this regard,

increasing the overall reliability of the process, which is one of our major aims with this project (see **Chapter 5.4**).

5.2.2. Assembly efficiency of BASIC and its competitors

After establishing the usefulness of optimising linker sequences, one of our goals was to carry these benefits over to a better DNA assembly technique that would surpass the current state of the art of DNA assembly in terms of efficiency, flexibility and reliability. In order to achieve this we developed BASIC, a new DNA assembly standard that employs an original assembly technique, based on a simple annealing reaction. This technique has produced very exciting results, yielding tens of colonies with >95% accuracy in 7-part assemblies, and a comparison against our results with Gibson assembly and yeast *in vivo* recombination under very similar conditions shows that BASIC is considerably more accurate than the former and much faster than the latter, while producing a similar number of colonies (see **Chapter 4.3.2**).

Keeping in mind the *caveat* mentioned above regarding the comparability of assembly efficiency across different techniques and constructs, other competitor techniques obtained similar results: Golden Gate, under the MoClo standard, yielded around 10^4 colonies for 4-part plasmids, 10 times more than BASIC, with similar near-100% accuracy. The gap shrinks when assembling more complex constructs, since 7-part assemblies produced only 150 colonies with 90% accuracy[33]. LCR instead seems to hold on better as the number of parts increases: under analogous conditions (using chemical transformation and 1 kb parts) it obtained around 10^3 colonies for 4-part constructs and around 10^2 for 7-part constructs, with 100% efficiency in both cases. LCR has also been shown to be able to assemble up to ten 1 kb parts before accuracy suddenly starts to decrease rapidly[3].

It is difficult to predict how BASIC will behave when increasing the number of parts, the size of the parts, or the size of the final construct, as this varies substantially depending on the mechanisms of the assembly reactions. The MoClo paper for example shows that Golden Gate is much less influenced by part size than LCR, possibly because LCR involves complete denaturation and re-

annealing of all DNA in the solution, which becomes progressively harder as the size of the fragments increases. We expect BASIC to be more similar to Golden Gate than LCR in regard to part size, due to the fact that we both rely on single stranded overhangs of a defined length to guide assembly, even though they are of different sizes.

Regarding the influence of the number of parts, our data shows no decrease in assembly accuracy as the number increases up to seven, while the number of colonies decreases substantially, so we expect the latter rather than the former to become the limiting factor. The reason for this might be that BASIC's last step does not employ ligation and the DNA fragments are only kept together by the strength with which they anneal to each other. Mis-annealed linkers would have a significantly lower binding strength, which would let the fragments come apart more easily, explaining why misassembled constructs are so rare. On the other hand it is possible that some of the correctly annealed fragments also might come apart and become unable to produce viable colonies, explaining why BASIC's colony yield is on the lower end of the spectrum among the latest DNA assembly methods. If this hypothesis is confirmed, this might be a positive feature after all, since colony numbers can be easily increased in a number of ways, such as using more efficient competent cells, better transformation protocols, running more transformation reactions in parallel *etc.*, which are all relatively simple solutions and amenable to automation. Dealing with reduced accuracy instead would require more laborious and *ad hoc* solutions such as extensive post-transformation colony screening and sequencing.

5.3. Flexibility

The flexibility of a DNA assembly standard is defined by its ability to comply with different needs without requiring *ad hoc* changes or slowing down the workflow significantly. Briefly, this includes things such as re-using parts and constructs in other assemblies, integrating with other techniques (such as combinatorial assembly, fusion proteins, mutagenesis, *etc.*), compatibility with different types of parts, the kind of scars it leaves (if any), the presence of hard limits on the number of parts it can assemble simultaneously, and so on.

5.3.1. Linkers and scars: benefits and drawbacks

The use of linker sequences to guide the assembly of DNA fragments implies that these remain in the final construct as "scars", as shown in **Figure 18**. Scar sequences have been cause of heated debate, especially after the development of high-efficiency scarless techniques such as SLIC, CPEC and Gibson assembly. The publications describing these techniques present the absence of scars as one of the main selling points, explaining that scars are likely to have undesired and unpredictable interactions with the neighbouring parts. We do not agree with this for the reasons explained in **Chapters 3.3.2** and **3.3.3**, and we believe that linker sequences are extremely useful not only to enhance assembly efficiency, but also to improve the flexibility of an assembly standard.

One of the main disadvantages of scarless methods is that a DNA part from one assembly reaction cannot be used in another assembly without altering the homology sequences at its flanks. This means that parts are not modular but rather "bespoke" and require a modification step to be re-used, typically done via PCR with newly synthesized primers. A linker-based approach instead allows modular assembly by defining purpose-built standard homology regions which can be attached to any part, so that they can be re-used across different assemblies.

This kind of linker-based modular format can also be extended to allow idempotent assembly: the same format is shared not only by parts, but also by newly assembled constructs, so that they can be re-used as parts in further rounds of assembly in a hierarchical fashion. This is extremely

helpful not only because entire constructs can be re-used instead of re-built from scratch, saving a great amount of time, but also because large assemblies can be split in two or more smaller and simpler intermediate ones which can be merged successively.

Some techniques and standards take a further step in this direction, and allow post-assembly modifications, such as swapping parts in and out of previously assembled constructs. This is an extremely useful feature, especially when small modifications are required on constructs that are particularly difficult to build from scratch. Unfortunately these systems have only been developed using classic restriction/digestion cloning[69,70] or recombinase-based cloning[59]: the first has very low assembly efficiency, while the second can only assemble constructs made of up to six parts. Both these drawbacks essentially mean that these systems cannot be used to build very complex constructs in the first place, so that post-assembly modification is not as relevant as it could be.

Regarding the concerns for the context effect that these linker sequences might cause, it is important to remember that context effects can appear between any two neighbouring sequences, even if assembled scarlessly. As mentioned in **Chapter 3.3.2** spatially separating functional regions can actually help insulating them, either by simply spacing them away or by using active mechanisms. Employing user-defined synthetic scars also means that if they are found to cause undesired context effects they can be easily modified or replaced. This would not be possible when adopting a scarless approach because usually none of the parts can be freely modified, or when using a DNA assembly standard that employs fixed scar regions, such as those that use recombinase-based assembly techniques (see **Chapters 1.2.3** and **1.3.3**).

Finally scarless assembly can be extremely problematic when dealing with inter-part repeated sequences[79], as shown in **Chapter 3.2.2**: the presence of very similar or identical sequences at the extremities of a DNA part inevitably cause the relative scarless homology regions to also be very similar or identical, causing misassembly of the parts.

For these reasons we chose to adopt a linker-based approach for both MODAL and BASIC, which seems to be becoming an increasingly popular choice in synthetic biology, being also adopted by all recently developed standards (see **Chapter 1.3**). This is most likely because the advantages of modularity and part insulation outshine those of scarless assembly in the eyes of the community, and this will only improve as better models and tools for the prediction of functional motifs will become available, helping to generate better functionally neutral sequences.

BASIC also supports hierarchical assembly, which we believe is extremely useful feature. It is not always present in DNA assembly standards due to the fact that it can be quite difficult to implement but it is indeed included by some of the most successful standards, such BioBricks[64], Golden Braid 2.0[75], MoClo[33] and the one developed by Guye *et al.*[78]. For what concerns post-assembly modifications, at the moment there are no viable methods for achieving it without increasing the number of forbidden sequences or adopting recombinase-based cloning, so we decided not to implement it.

Even though this means that with MODAL and BASIC constructs must be re-built from scratch every time a modification is necessary, their workflows are quite fast, especially compared to most other DNA assembly standards such as MoClo, Golden Braid ,Torella *et al.* and Guye *et al.*. This is because these standards use a cloning step in order to attach the appropriate homology regions to parts, which requires cell transformation, colony screening and construct isolation. MODAL and BASIC instead do not require any of this, going from standard parts to final construct transformation in a single day, saving at least two days of work.

5.3.2. Single vs. multiple tiers of assembly

The linker-based approach in MODAL and BASIC is completely part-agnostic: any type of part is assembled exactly in the same way, in a single-tier fashion, even across multiple rounds of hierarchical assembly. As described in **Chapter 1.3** many other DNA assembly standard adopt a multi-tier approach instead, which is typically implemented in a hierarchical system: usually transcription units are built during the first round of assembly, from a standard set of sub-gene components (e.g. promoter, ORF, terminator) in a fixed order. These transcription units are then assembled into multi-gene constructs during the following round, which typically allows complete freedom of choice regarding their position.

The advantages of a multi-tiered approach often come from the limitations that it adopts, especially in the first tier: for example in HomeRun[82] and Guye *et al.*[78] transcription units are assembled using the recombinase-based Multisite Gateway cloning system, which is very efficient but can only assemble up to five fragments at a time. Normally this would be a very undesirable limitation, but in these standards transcription units always contain a fixed number of sub-gene parts (five and three respectively) so it is not an issue at all. Another example comes from MoClo[33] and Golden Braid[74,75], where the definition of the position of a part in a construct comes from the particular positional vector it is cloned in, and changing position requires cloning in a different vector. A multi-tiered approach with fixed-order transcription units alleviates this inconvenient because sub-gene parts only need to be cloned in a single positional vector, since they are always placed in a fixed order when assembling transcription units. This cannot be done for the following tiers, where positional freedom is required, which is why MoClo requires an extraordinary number of positional vectors and Golden Braid a complicated looped design.

There are a few important downsides to this approach: in our opinion the main one is that the pre-determined design of the transcription units cannot be changed without requiring extensive *ad-hoc* adaptations of the experimental design. Multi-tiered standards are inevitably designed with a

very specific project (or type of project) in mind, so their usefulness and applicability are intrinsically limited. Adapting their structure to comply with alternative designs is a step back towards non-standard cloning, with all the problems discussed earlier, and reduces the usefulness of adopting a standard in the first place. Additionally different tiers intrinsically require different experimental designs and often employ different techniques, making the overall process more complicated and possibly more expensive, in terms of both time and money (*e.g.* multiple sets of enzymes might be required).

Single-tier DNA assembly standards such as MODAL and BASIC can freely assemble any part in any position, and the only limit on the number of parts that can be assembled simultaneously derives from assembly efficiency. This guarantees a much higher level of flexibility and universality, which we believe are essential features of a DNA assembly standard. They can be used for a wider range of experimental designs and they can adapt more easily to changing needs, which is a very common occurrence even within a single project. In addition to this, as discussed in **Chapter 1**, one of the advantages of adopting a standard for DNA assembly is facilitating the exchange of parts and constructs between research groups. This can only happen if they all adopt the same standard, which in turn is more likely if the standard in question has a wider applicability. On a side note, as mentioned in **Chapter 4.3.4**, we are planning on implementing an optional "pseudo multi-tier" element in BASIC by allowing the assembly of short functional sequences as part of the linker regions. This does not involve any of the drawbacks mentioned above, and rather adds another element of flexibility.

5.3.3. Compatibility with parts

The freedom to choose the placement of all DNA parts in a construct is not the only factor that determines the universality of a standard: some techniques carry specific limitations regarding the features of the DNA sequences that participate in the reactions, so that they might be compatible with certain DNA parts but not with others. The use of PCR amplification and restriction digestion are the most common causes of incompatibility: the former has difficulties with DNA fragments that present certain features such as high GC content, repetitive regions, excessive length, *etc.*, while the latter requires the absence of additional restriction sites within the parts, which can sometimes be very difficult to achieve (*e.g.* some sequences cannot be modified without significantly changing their behaviour, or a certain restriction site might be extremely common).

There have been attempts at developing standards that minimise or avoid the use of either technique, which typically employ recombinase-based systems and homing endonucleases, such as the one proposed by Guye *et al.*[78] and HomeRun[82]. Unfortunately due to the limited number of parts that recombinase-based techniques can assemble simultaneously, they both adopt a multi-tier structure, with the disadvantages discussed in **Chapter 5.3.2**. Additionally homing endonucleases are not as reliable as restriction enzymes in terms of recognition sequence specificity, adding an undesirable element of unpredictability, and they leave long non-modifiable scars, which can sometimes be an issue as explained in **Chapter 5.3.1**.

Torella *et al.*[79] propose a middle-ground solution instead, that employs normal restriction enzymes but with redundancy: each restriction site that must be cut to release the DNA part from the storage plasmid can also be cut with a type IIs enzyme that has a different recognition site. This approach increases the chances of finding a set of unique recognition sites, but does not guarantee it, and including a large number of enzymes in the standard might expose it to the reliability issue mentioned in **Chapter 5.4.2**, whereby some of these enzymes might be problematic to use.

MODAL employs PCR amplification in Step 1, to attach the linker sequences to the parts, and is thus subject to all the limitations that come with it. This choice was made because we prioritised simplicity and speed over universality, but nonetheless we made every effort possible to define a strong and reliable protocol for the PCR that ensures the highest quality results possible. BASIC on the other hand aims to be a widely applicable assembly standard, and this is one of the reasons why we replaced the Step 1 PCR with a restriction digestion & ligation reaction. In order to limit the problems with forbidden sequences we only employ a single restriction enzyme, BsaI, unlike other recent standards that employ two or more[33,74,75]. BsaI is also a very commonly used enzyme and thus less likely to be found inside parts that researchers routinely work with: there is even a DNA synthesis company, Gen9[114], which requires all submitted sequences to be BsaI-free due to the needs of their synthesis process. Finally, even though it has not been tested yet, it is possible that BsaI sites inside BASIC parts could be tolerated as long as they generate overhangs that are completely orthogonal to the iP/iS overhangs. They are expected to be cut and re-ligated periodically during the reaction, producing at least a fraction of intact products, as mentioned in **Chapter 4.3.4**.

Another cause of part incompatibility in DNA assembly standards is the use of DNA purification protocols, since they can only retrieve fragments within a certain size range: the lower limit for part size in both MODAL and BASIC is set by the PCR purification step in the workflow, which removes any primer-sized part or smaller (about <100 bp). The upper limit for MODAL is set by the use of PCR amplification, which means that parts larger than 5 kb might require *ad hoc* troubleshooting or be impossible to amplify correctly. BASIC does not have this problem, but the beads-based purification protocol has a risk of shearing very large DNA molecules. The purification kit manufacturer suggested via personal communication that fragments as large as 10-15 kb can be retrieved as long as mechanical stress is kept to a minimum during pipetting.

5.3.4. Integration with other techniques

DNA assembly is often performed together or as preparation for the use of other experimental techniques, and how easily they integrate is an important factor in determining the flexibility of a standard. Combinatorial assembly is one of the most common examples, as it is employed for a variety of applications: when building a construct one or more of the parts are not introduced as a solution of identical molecules, but as a pool of variants that have different sequences and functionalities but are all assembled in the same position (e.g. assembling a pool of promoter mutants driving the same GFP gene as in **Chapter 3.3.5**). The result is that the colonies on the final transformation plate will contain different versions of the same construct, depending on which part variants they assembled.

All DNA assembly techniques are in theory able of integrating this technique, but it is much easier with linker-based standards: the sequences of the part variants will be at least slightly different, if not completely different. This means that assembling them scarlessly requires *ad hoc* homology regions for every single variant, which can be quite time consuming to perform with small libraries (10^1-10^2 elements) and very hard or impossible with large ones (10^3 or more variants), even though there are software tools that can assist with this[84].

The use of linkers solves this problem by using always the same homology regions for parts that are to be placed in the same position so that all variants are assembled identically and can just be directly mixed in the assembly reaction. Another important factor in determining whether a DNA assembly method or standard can perform combinatorial assembly is its efficiency: the number of possible assembly combinations can be very high, especially when mixing two or more variant pools for different parts of a construct. This means that the reaction needs to produce a very high number of correctly assembled constructs and colonies in order to retrieve a decent portion of the combinatorial variants of the construct.

Both MODAL and BASIC are well suited for combinatorial assembly, as they both adopt a linker-based approach, which greatly facilitates the experimental design. MODAL also improves the efficiency of the existing assembly techniques involved by using optimised linker sequences, making them able to retrieve a larger fraction of the combinatorial space, and is the only DNA assembly standard currently available that seamlessly supports the generation of mutant libraries of parts via mutagenenic PCR during the normal assembly workflow. BASIC similarly ensures an even greater level of assembly efficiency, further improved by its near-100% accuracy. Colony screening for combinatorial assembly can be extremely time-consuming because researchers are not looking for only one correctly assembled construct, but for as many variants of it as possible. Having near-100% accuracy means that essentially almost every colony produced contains a correctly assembled member of the combinatorial library, greatly reducing or removing the need for screening.

Protein fusion is another very common technique which consists of forming a single hybrid protein coding sequence by consecutively assembling two DNA parts that contain two separate protein coding sequences. It is very strongly tied to DNA assembly because any scar sequence left in between will also be translated to amino acids, so it is important that it does not prevent the hybrid protein from folding or working correctly. Very few DNA assembly standards support this natively: the short homology regions in the various Golden Gate-based standards[33,74,75] have to be modified *ad hoc* to allow it, and protein fusion-compatibility is one of the main goals behind development of the BioBrick standard variant BglBricks[71]. Scarless assembly techniques can achieve this very easily, but they also carry all the disadvantages mentioned previously. The BASIC standard instead has been designed to be natively compatible with protein fusions, when using either the BASIC or the MODAL workflow, as described in **Chapter 4.2.1**.

It is also important to underline the fact that the BASIC standard can indeed be used to perform both the BASIC and the MODAL workflow: this means that, given the same standardised parts, the user can choose between two different ways of attaching linkers to the parts and four different

assembly techniques in total, depending on their specific needs. Such a wide variety of options makes BASIC one of the most flexible and universally applicable standards currently available.

5.4. Reliability

In the previous chapters we discussed the importance of standardisation for DNA assembly, intended as stepping away from *ad hoc* solutions and moving towards established formats and protocols. While flexibility allows a standard to be applicable in as many cases as possible, its reliability ensures that there are no unexpected exceptions to it: a reliable standard should always work as expected within its known limits, and should not require any *ad hoc* troubleshooting.

5.4.1. PCR

The most common source of unreliability in DNA assembly is the use of PCR: even when amplifying sequences that comply with all the usual limitations (*e.g.* not too long, not too GC rich, no repetitive regions, *etc.*) there is always a chance that the reaction will introduce sequence mutations or that it will encounter other problems. All DNA polymerases, even the best high-fidelity ones available, will introduce sequence mutations at a certain rate (4.4×10^{-7} for Phusion DNA polymerase[115], used in this work), which means that sequence integrity can never be assumed when working with PCR-amplified DNA. This introduces and element of uncertainty that can only be eliminated by sequencing all constructs assembled using PCR products, which is time consuming and quite expensive.

Another significant problem is that the annealing mechanisms that guide primer-target recognition are never 100% accurate, so there is always a chance of amplifying the wrong region. Due to the fact that PCR is a chain reaction where the products act as template to generate more products in the following cycles, both sequence errors and annealing errors that occur during the early cycles will propagate exponentially for the rest of the reaction.

Nevertheless PCR amplification is widely employed in DNA assembly: it is used for example to isolate fragments from a plasmid or chromosomal source, to perform site directed mutagenesis (*e.g.* to remove undesired restriction sites) and to attach short sequences to the flanks of existing DNA fragments. The latter in particular is exploited by nearly all assembly methods, as it can be used to

attach restriction sites, recombinase recognition sites and long homology regions, and the alternatives are either slow and difficult (classic digestion/ligation using spontaneously present restriction sites) or very expensive (total synthesis). The most recent DNA assembly standards take this into account by cloning amplified DNA parts in vectors where they can be safely maintained and sequence-verified, and they do not use PCRs for the rest of the workflow[33,74,75,78,79,82].

It is also important to note that certain DNA assembly methods exploit mechanisms that are very similar to PCR and are thus subject to similar problems: CPEC[1] is essentially identical to a PCR amplification, and is inevitably equally unreliable. Gibson assembly[2] uses DNA polymerisation to repair gaps left in the homology regions after joining two fragments, so there is a chance that sequence errors will be introduced. Finally LCR[3] adopts a PCR-like chain reaction mechanism where fragments that are incorrectly joined during the first cycles can act as templates for other misassemblies, propagating the error during the rest of the reaction.

Both MODAL and BASIC at Step 0, where the prefix and suffix sequences are attached to the parts, employ the strategy described above of using a PCR amplification followed by cloning, screening and sequence verification to overcome the reliability problems. MODAL then proceeds by using another PCR to attach linkers to the parts (Step 1) in order to keep the workflow as simple and easy as possible but, being aware of the issues just described, we made all possible efforts to maximise its reliability. As described in **Chapter 3.2.1,** the protocol uses a high fidelity DNA polymerase and the conditions are optimised to achieve maximum specificity. At the assembly stage (Step 2) MODAL gives the user the choice between three techniques: CPEC and Gibson assembly are simple and fast but quite unreliable, while yeast *in vivo* recombination has the opposite qualities. We believe that MODAL strikes an acceptable compromise, keeping all the advantages of using PCR amplifications while minimising the downsides, and letting the users choose the trade-off that suits their needs best at the assembly step.

BASIC instead was developed with the main aim of maximising reliability, so the PCR amplification at Step 1 had to be replaced with a digestion/ligation step. As mentioned above there were no alternative viable methods available to attach linkers to parts, so we devised a novel method that is inspired by Golden Gate assembly and by pre-PCR era methods used to add restriction sites at the end of a DNA fragments. The result is a highly efficient, reliable and simple protocol whose only downside is requiring one forbidden restriction site (see **Chapter 5.3.3**). Similarly we designed an original assembly protocol for Step 2 of the BASIC workflow, which consists of a very simple and fast annealing reaction that does not present any elements of unreliability.

5.4.2. Other unreliability factors

Beside PCR amplification, anther common source of unreliability in DNA assembly is the use of reactions that are very sensitive to small changes in conditions such as temperature, incubation time and reagents concentration. Exonuclease chew-back reactions are a prime example of this: in Gibson assembly[2] and SLIC[4] DNA digestion proceeds indefinitely until the exonuclease enzyme spontaneously denatures or the reaction tube is chilled on ice, respectively. Different incubation times or temperatures greatly affect the amount of DNA that is digested, and thus the outcome of the assembly reaction.

Certain restriction endonucleases can also present unpredictability factors, such as star activity (off-target cleaving), instability at certain temperatures, rapid loss of activity in storage, *etc.* Extensive efforts by commercial companies have been directed towards developing more reliable versions of the most commonly used enzymes (*e.g.* the High Fidelity series by NEB), but digestion/ligation-based standards that use a large number of different enzymes often include one or more problematic ones[69,70,95]. Homing endonucleases, employed in many recently developed standards[78,82], are also prone to cleaving unintended sites, even under optimal conditions, because their recognition sequences a allow a certain level of variability.

MODAL employs Gibson assembly as one of the three options for the final assembly reaction (Step 2) but, as mentioned previously (**Chapter 5.3**), we believe that its unreliability is compensated by MODAL's overall speed and simplicity, and by the fact that the users can easily switch to using yeast *in vivo* recombination instead, if a higher level of reliability is required.

BASIC on the other hand does not contain any of the unreliability factors presented in this chapter, as both Step 1 and 2 of its workflow employ reactions that are extremely resilient and not easily compromised. Step 1 uses BsaI from NEB's High Fidelity series and T4 DNA ligase, while Step 2 is a simple annealing reaction. None of these enzymes has unpredictable or unspecific activities, or is

particularly sensitive to inactivation at low or medium temperatures (<37°C approximately), and neither Step 1 nor Step 2 are greatly affected by changes in incubation time or temperature.

Finally human error inevitably plays a large role in compromising the reliability of DNA assembly, especially when performing protocols that require a high level of manual skill such as DNA purification by gel extraction. Gel extraction is a very powerful technique that selectively isolates DNA fragments of any chosen size, but its success varies a lot depending on a number of practical details, such as how good is the separation of the bands during electrophoresis, how precisely is the desired band cut out of the gel, and so on. Even though this is one of the least reproducible protocols in molecular biology, it is required by many DNA assembly standards, from BioBricks[64] to Torella *et al.*[79]. One of the main reasons for this is that gel extraction is one of the few methods that can be used to isolate a DNA fragment that is contained in a plasmid vector: typically the fragment is separated from the backbone with a restriction digestion, and then isolated via gel extraction. This is very often required when preparing DNA fragments for assembly because leaving the backbone in the same solution means that when the fragment is used in a DNA assembly reaction there is a chance that it is re-joined to the backbone, reconstituting the original plasmid and causing background transformation.

Both MODAL and BASIC are exposed to this problem, because DNA parts are cloned in storage plasmids from which they have to be extracted (during Step 1) in order to use them in the assembly process. In MODAL this problem has an easy solution, because Step 1 uses a PCR amplification: the desired part is simply amplified from the storage plasmid which is then destroyed with a DpnI digestion, leaving nothing that could produce undesired viable colonies. BASIC instead uses a simultaneous digestion/ligation reaction in Step 1, during which DNA parts are cut away from the storage backbone and made available for linker ligation. Reconstitution of a large quantity of original storage plasmid is made less likely by the nature of the reaction, since reconstituted plasmids can be

cut again during the reaction while the DNA parts that are ligated with the linkers cannot, and are thus unable to be re-joined to their original backbone.

Nevertheless, as seen in **Chapter 4.2.4**, this is not sufficient to completely eliminate background transformation, as there are always a certain number of viable plasmids being produced, either by regenerating the original storage plasmids or by unspecific ligation of the storage plasmid backbones to themselves. In order to prevent background transformation we adopted a double antibiotic selection system, so that no single storage plasmid backbone contains all required antibiotic resistance genes to produce viable plasmids. This allows us to avoid both PCR amplifications and gel extractions, a solution that has also been adopted by Guye *et al.*[78]. Golden Gate-based standards instead adopt a hybrid system that uses both antibiotic resistances and colour selection[33,74,75], which is also very successful, but requires the preparation of special agar plates for colour selection and does not prevent incorrect colonies from growing, but merely marks them visually.

5.5. Automation

Being able to perform DNA assembly protocols on automated liquid handling platforms would be highly beneficial, first of all because it would dramatically increase the throughput compared to manual work, and secondly because it would reduce the level of human involvement in the process. This will eventually transform DNA assembly from one of the most difficult parts of many synthetic biology projects to a commodity technology, delegated to machines or specialised service providers.

The synthetic biology community has already developed software tools that produce machine-readable assembly plans [83,84], even though the development of automation-friendly assembly protocols has been much slower, for a number of reasons. The first obvious problem is that liquid handling platforms cannot perform certain protocols, depending on the equipment available. Protocols involving gel electrophoresis are generally excluded, such as gel extraction-based DNA purification and post-assembly screening based on restriction or amplification patterns. This is already a very strict limitation, since a large number of DNA assembly workflows require gel extraction or do not achieve a sufficient assembly accuracy to do without screening. Liquid handling platforms are also rarely equipped with centrifuges, which excludes a number of protocols such as column-based DNA purification, even though there are magnetic beads-based kits that can be used instead.

Another difficulty comes from the fact that robots often perform protocols differently from how humans would: they cannot precisely pipette very small volumes (usually <5 µl), they work with 96-well plates instead of single tubes, they perform certain pipetting steps much faster or much slower, and so on. This essentially means that protocols need to be adapted, and thus they need to employ reactions that are robust to the unfavourable changes in conditions that are sometimes necessary. Additionally robots often cannot react or even notice when unexpected problems occur, so the assembly protocols need to be extremely predictable and reliable.

Due to all these issues, implementing automated DNA assembly protocols is very challenging and only done when manual assembly is extremely difficult. A notable example is the assembly of TAL effector genes, which encode for DNA binding proteins made of modular domains, each recognising a specific nucleotide. Typically they are assembled to recognise a 10-20 bp target sequence using 10-20 different DNA fragments, and various combinations are usually generated to compare binding efficiency, target different sites, fuse them to different functional domains, *etc.* Manual assembly of tens or hundreds of variants is clearly excessively time-consuming, but fortunately the best methods to assemble TAL effectors genes are Golden Gate-based protocols, which are also very automation-friendly due to their accuracy, adaptability and reliability. For these reasons TAL effector gene assembly is one of the few areas were automated DNA assembly has been successfully applied[30,116].

BASIC was designed with automation in mind: both Step 1 and Step 2 of the workflow only contain protocols that can be run entirely by a liquid handling robot with standard equipment, and as discussed in **Chapter 5.4.2** all the reactions are highly tolerant to changes and adaptations. Step 1 uses a simultaneous digestion/ligation reaction that is very similar to those employed by the automated protocols for the assembly of TAL effectors mentioned above. The reaction is followed by a PCR purification step that employs a magnetic beads-based kit that is actually specifically designed to be used on liquid handling platforms. Step 2 is a simple annealing reaction that should not present any obstacles to automation, and finally transformation of chemically competent cells is routinely automated for a variety of applications, up to the stage where transformed cells are spread on agar plates, which is usually done manually. Post-assembly screening, also typically done manually, should not be required due to BASIC's high accuracy. Step 0 on the other hand has to be performed manually, but it is only a one-time procedure for each part, and it is actually important that this step is curated manually to ensure that all BASIC-formatted parts are thoroughly verified before using them for assembly.

In conclusion, we expect that BASIC can be easily adapted to be run on a liquid handling platform completely hands-off from Step 1 up to the cell plating stage using standard equipment such as 8-way pipetting, 96-well plate handling, heated and cooled plate holders, a magnetic plate holder, a shaking plate holder and a thermocycling block. Unexpected problems that require human intervention should be rare due to the high reliability of all the reactions involved (see **Chapter 5.4**). We believe that BASIC can be the foundation for the development of automated assembly protocols that are simple enough to use to be useful not just in very specific cases but to a wide range of projects.

5.6. Conclusion

We developed two DNA assembly workflows, MODAL and BASIC, that can be used for the assembly of gene or sub-gene DNA parts (such as promoters, RBSs and ORFs) into multi-gene constructs. MODAL is intended as a way to modularise and improve commonly used long overlap-based assembly methods, while BASIC proposes a new assembly standard for synthetic biology, complete with an original assembly method. Both are based on the idea that computationally designed homology regions, called linkers, are extremely beneficial to DNA assembly reactions, just like careful primer design is for PCR amplifications. For this reason we developed Linker, a MATLAB script that automatically generates these linkers, which was later used as the foundation for the development of a more advanced and publicly available web tool called R2oDNA Designer. Our work on MODAL acted as a stepping stone to verify our ideas about linker sequences, context effects, long overlap-based DNA assembly, *etc.*, which informed the development of BASIC.

We had a number of ideas about how BASIC should be: we wanted it to be very efficient, which means being able to assemble at least five or more parts simultaneously, yield a large number of colonies and most of all be as close to 100% accurate as possible. We wanted it to be extremely reliable, to solve the age-old problem of DNA assembly: the requirement of constant *ad hoc* troubleshooting. We also wanted it to be universally applicable, so that it could actually be useful to the synthetic biology community at large, by being compatible with a vast range of DNA parts, design needs, other techniques, *etc.* And finally we wanted it to be able to be run by robotic liquid handling platforms, which are the inevitable (and highly desirable) future of DNA manufacturing. We believe we achieved all of these goals, some of them quite well and some of them less well, but we still tried to propose a plausible solution. We hope that this work will be valuable to the synthetic biology community and to molecular biology research in general, and that it will be used to inform the future developments of DNA assembly standards, techniques and practices.

5.7. Future work

We believe that DNA assembly in general will greatly benefit first of all from an in-depth exploration of the features that determine the performance of the homology sequences in assembly reactions, similar to what has been done for other techniques such as PCR and oligonucleotide microarrays. Secondly this knowledge should be used to develop advanced software tools that generate highly optimised linker sequences. These two points will also be crucial for the future development of BASIC, since its main competitors (LCR[3] and the Golden Gate-derived standards MoClo and Golden Braid[33,74,75]) all present similar or slightly better assembly efficiency. BASIC's advantage resides in its simple and powerful modular framework, because LCR at the moment is presented as a scarless technique, not supported by any modular standard, while all Golden Gate-derived standards are extremely complicated and difficult to adopt. We believe future work on BASIC should initially focus on bridging the assembly efficiency gap, both by improving BASIC's assembly protocol and by addressing the two issues mentioned above regarding the computational design of optimised linker sequences.

Additionally, in order to further reduce any elements of uncertainty and unpredictability, it will be important to deal with the context effects caused by long scar sequences: future work will need to deepen our understanding of these effects and devise strategies to minimise them or predict them reliably in order to be able to integrate them in the design process. Another outstanding technical issue regards BASIC's compatibility with automation: in the future it will be very interesting to formulate an adaptation of the BASIC workflow for automated liquid handling platforms and test it experimentally. Finally, in order to promote the adoption of the BASIC standard in the synthetic biology community, it will be necessary to curate a database of BASIC-formatted parts cloned in appropriate storage plasmids and to make sure that is well annotated and easy to distribute.

6. Materials and methods

6.1. Cells manipulation

6.1.1. Strains and media

Escherichia coli DH10B (Invitrogen) and DH5α (New England Biolabs) strains were used as the host to clone bacterial plasmid DNA. DH5α chemically competent cells prepared by Marko Storch were used to transform BASIC-assembled plasmids, while DH10B cells were used for all other purposes. Liquid cultures were grown at 37°C in Luria-Bertani (LB) medium (Sigma-Aldrich) in a shaking incubator, solid cultures were grown at 37°C on plates prepared with LB-agar medium (Sigma-Aldrich). SOC medium prepared by Marko Storch (2% w/v tryptone, 0.5% w/v yeast extract, 10 mM NaCl, 2.5 mM KCl, 10 mM MgCl2, 20 mM glucose) was used in the transformation of DH5α cells. Both liquid and solid cultures were supplemented with kanamycin (50 mg/ml), ampicillin (50 mg/ml), chloramphenicol (25 mg/ml) or combinations thereof as needed to select for cells transformed with the plasmid of interest.

Saccharomyces cerevisiae YPH500 strain[117] was used as the host for yeast plasmid DNA and was grown at 30°C in Yeast extract Peptone Dextrose (YPD) rich medium or synthetic complete drop-out medium lacking uracil (SC-Ura)[50] to select for transformed cells.

6.1.2. Competency protocols

E. coli cells were made electro competent using the following protocol. In order to maintain the efficiency of the cells, it is vital that once the mid-exponential culture is chilled, the cells remain at a low temperature for the rest of the procedure. Only use sterile solutions and vessels.

Day one:

1. Inoculate a 10 ml LB culture with a single colony and incubate shaking overnight at 37 °C.
2. Incubate a conical flask containing 1 litre of LB overnight at 37 °C.

Day two:

1. Use the overnight culture to inoculate the LB flask and incubate shaking at 37 °C.
2. Chill on ice 1 litre of distilled H2O, at least 40 ml of 10% w/v glycerol, two 500 ml centrifugation bottles, a centrifuge rotor for 500 ml bottles, and about a hundred 1.5 ml Eppendorf microtubes.
3. When the OD600 value of the culture reaches ~0.5, transfer the liquid to two 500 ml centrifugation bottles and chill on ice for 30 minutes.
4. Set the centrifuge to 4 °C and spin the bottles in the pre-chilled rotor at 4000 rpm for 15 minutes.
5. Discard the supernatant and re-suspend the cells in 500 ml of pre-chilled water. To facilitate the re-suspension of the pellet use only 20 ml at first and add the rest later.
6. Centrifuge as before (step 4).
7. Discard the supernatant, re-suspend the cells in 20 ml of pre-chilled 10% glycerol and transfer the liquid to two 50 ml Falcon tubes.
8. Centrifuge as before (step 4).
9. Discard the supernatant and re-suspend the cells in 2.5 ml of pre-chilled 10% glycerol.
10. Transfer the cells into the pre-chilled microtubes in 50 µl aliquots and store immediately at -80 °C.

E. coli cells were made chemically competent using the same protocol but using 0.1 M $CaCl_2$ instead of water and 0.1 M $CaCl_2$ plus 15% w/v glycerol instead of 10% w/v glycerol. Final aliquots were 200 µl instead of 50 µl.

S. cerevisiae cells were made competent during the transformation protocol (see **Chapter 6.1.3**).

6.1.3. Transformation protocols

Bacterial electroporation was performed using a Biorad Micropulser electroporator set on "Bacteria" and "Ec1" according to the following protocol:

1. For each transformation reaction put one tube of competent cells (containing 50 µl) and one electroporation cuvette on ice. Also put a tube containing at least 1 ml of LB medium per reaction and the necessary plates in the 37 °C incubator.
2. When the cells are thawed, transfer 40 µl of them to the electroporation cuvette and add the appropriate amount of DNA. Mix gently by moving the tip of the pipette.
3. Place the cuvette in the electroporator and activate it using the appropriate settings for *E. coli* transformation.
4. Immediately add 950 µl of pre-heated LB medium to the cuvette, mix gently and transfer to a 1.5 ml Eppendorf microtube.
5. Incubate the microtubes in a 37 °C shaking incubator for one hour.
6. Spin the cells for 1 second at 13000 rpm on a benchtop centrifuge, then remove 900 µl of supernatant and resuspend the pellent in the remaining liquid.
7. Transfer the desired amount of culture on the pre-heated plates and spread evenly until all liquid is absorbed.
8. Incubate the plates overnight at 37 °C, making sure that the agar side of the plate is facing up.

Bacterial chemical transformation was performed using the following protocol. When transforming into Dr. Marko Storch's DH5α cells, SOC medium instead of LB was used at step 5.

1. Put one 14 ml Falcon tube for each transformation and one tube of competent cells (containing 200 µl) for each four transformations on ice. Also put a tube containing at least 1 ml of LB medium per reaction and the necessary plates in the 37 °C incubator. Prepare a 42 °C water bath.
2. When the cells are thawed, transfer 40 µl of them to the Falcon tube and add the appropriate amount of DNA.
3. Incubate the Falcon tubes on ice for 30 minutes.
4. Place the Falcon tubes in the water bath for exactly 45 seconds, then immediately transfer back on ice for 2 minutes.
5. Add 950 µl of pre-heated LB medium to the cuvette, mix gently and transfer to a 1.5 ml Eppendorf microtube.
6. Incubate the microtubes in a 37 °C shaking incubator for one hour.
7. Spin the cells for 1 second at 13000 rpm on a benchtop centrifuge, then remove 900 µl of supernatant and resuspend the pellet in the remaining liquid.
8. Transfer the desired amount of culture on the pre-heated plates and spread evenly until all liquid is absorbed.
9. Incubate the plates overnight at 37 °C, making sure that the agar side of the plate is facing up.

S. cerevisiae transformation reactions were performed using the following protocol, adapted from Gietz *et al*. The cells are made competent during the procedure. This protocol was used both to introduce transform plasmids and to assemble them through *in vivo* recombination[50].

Day one:

1. Inoculate 5 ml of YPD liquid medium in a 14 ml Falcol tube using a single colony grown on a YPD-agar plate. Incubate at 30 °C overnight.

Day two:

1. For each transformation reaction inoculate a 3 ml YPD liquid culture in a 14 ml Falcon tube using 30 µl of the culture from day one. Place the cultures in a shaking incubator at 30 °C for 4 hours.
2. Prepare single stranded carrier DNA by boiling a tube of 2 mg/ml salmon sperm DNA (New England Biolabs) for 5 min and then chilling on ice
3. Transfer 2 ml of the cultures in 2 ml Eppendorf microtubes and spin them on a benchtop centrifuge at 13000 rpm for 30 seconds. Discard the supernatant and resuspend the pellets in 1 ml of sterile water.
4. Spin the microtubes again as in step 3 and discard the supernatant.
5. Add the following reagents to the pellets in the microtubes in the order listed. Make sure they have all been sterilised.
    a. 240 µl of 50% w/v PEG 3350
    b. 36 µl of 1 M LiAc
    c. 50 µl of single stranded carried DNA, vortex prior to addition
    d. 34 µl of distilled water plus any plasmid DNA or DNA fragments
6. Resuspend the pellet by vortexing, then incubate at 42 °C for one hour.
7. Spin the microtubes again as in step 3, discard the supernatant and resuspend in up to 1 ml of sterile water to obtain the desired dilution.
8. Plate 100 µl of cells on the appropriate selective medium plates and incubate for 3-4 days at 30 °C.

6.2 General DNA manipulation

6.2.1. Isolation and purification

Plasmid DNA was isolated using the Qiagen QiaPrep Spin kit, DNA fragments were purified using either the Qiagen QiaQuick Spin or the Qiagen QiaQuick Gel Extraction kit. Manufacturer's protocols were followed for all. The Agencourt AMPure XP kit was used to purify DNA fragments in Step 1 of the BASIC workflow, as explained in **Chapter 6.3.3**.

6.2.2. Digestion and ligation

All DNA digestions were performed using NEB restriction enzymes according to manufacturer's protocol. Diagnostic digestions were usually set up in a total volume of 20 µl with 10 U of enzyme per 500 ng of DNA, and incubated for 1 hour at the required temperature. DpnI digestions to destroy template plasmid DNA from PCR amplifications were performed by adding 0.5 µl of enzyme directly in the PCR mix at the end of the thermal cycling, and incubating for 1 hour at 37 °C.

DNA ligations were performed using T4 DNA ligase supplied by NEB. DNA insert and vector were ligated using T4 DNA ligase (NEB). A 3:1 molar ratio of insert to vector was used, with 50 ng of vector. DNA was added to a 20 µl reaction mix, with 400 U of T4 DNA ligase and 1× T4 DNA ligase buffer, and incubated for 2 hours at room temperature. Ligase was inactivated by incubation at 65 °C for 10 minutes. Linearised vectors were treated with Antarctic Phosphatase (NEB) following the manufacturer's protocol to reduce self-ligation. Circularisation of linear fragments was performed similarly using 50 ng of DNA fragment only.

The setup of the digestion/ligation reaction in Step 1 of the BASIC workflow is detailed in **Chapter 6.3.3**.

6.2.3. Oligonucleotides preparation

All oligonucleotides were synthesised by Integrated DNA Technologies (IDT) and dissolved in water at a concentration of 100 µM. Working stocks of PCR primers were prepared by diluting them to 10 µM. BASIC's partially double stranded linkers were prepared at a concentration of 1 µM in a total volume of 50 µl by mixing 49 µl of annealing buffer (10 mM TRIS buffer, 100 mM NaCl, 10 mM CaCl2, HCl to pH 7.9), 0.5 µl of the long oligonucleotide and 0.5 µl of the short oligonucleotide. The solution is incubated 30 mins at room temperature to let the oligonucleotides anneal before storing it at -20 °C.

6.2.4. PCR

Routine PCR amplifications were performed using Phusion DNA polymerase using the following mix: 1X Phusion HF buffer (NEB), 5% DMSO (NEB), 200 µM dNTPs (Sigma-Aldrich), 0.25 µM primers, 5-50 ng template DNA, 0.02 U/µl Phusion DNA polymerase (NEB). Total volume used was 20 µl for diagnostic PCRs and 50 µl for the amplification of fragments for downstream use. Thermal cycling consisted of the following steps: (i) initial denaturation at 98 °C for 30 seconds, (ii) 20-30 cycles of denaturation at 98 °C for 10 seconds, annealing at 55-65 °C for 30 seconds, extension at 72 °C for 30 seconds/kb, (iii) final extension at 72 °C for 5 minutes. Annealing temperatures were calculated using the IDT OligoAnalyzer web tool with the following parameters: Na+ 50 mM, Mg++ 1.5 mM, dNTPs 0.2 mM, and increased by 3 °C following NEB's recommendation. Reactions using primers able to anneal at 72 °C or above were run without the annealing step.

Colony PCR amplifications were performed using TAQ Dna polymerase using the following mix in a total volume of 20 µl: 1X Standard TAQ buffer (NEB), 5% DMSO (NEB), 200 µM dNTPs (Sigma-Aldrich), 0.25 µM primers, DNA, 0.025 U/µl Taq DNA polymerase (NEB) and 1 ul of template obtained by dissolving the colony in 20 µl of water. Thermal cycling consisted of the following steps: (i) initial denaturation at 98 °C for 5 minutes, (ii) 20-30 cycles of denaturation at 98 °C for 10 seconds, annealing at 55-65 °C for 30 seconds, extension at 72 °C for 1 minute/kb, (iii) final extension at 72 °C for 5 minutes. Annealing temperatures were calculated without the +3 °C adjustment.

Mutagenic PCR was performed adapting the protocol by Zaccolo *et al.*[97] and is composed of two subsequent amplification reactions. The reaction mix for the first PCR contains 1x Standard Taq (Mg-free) buffer (NEB), 50 mM MgCl2 (Sigma-Aldrich), 200 µM dNTPs (Sigma-Aldrich), 200 µM dPTP (Trilink Biotechnologies), 200 µM 8-oxo-dGTP (Trilink Biotechnologies), 1 mg/ml gelatine (Sigma-Aldrich), 50 ng template and 5 U/µl Taq DNA polymerase (NEB). Thermal cycling settings were: (i) initial denaturation at 98 °C for 2 minutes, (ii) 30 cycles of denaturation at 98 °C for 1 minute, annealing at 55 °C for 1.5 minutes, extension at 72 °C for 5 minutes, (iii) final extension at 72 °C for 5 minutes. 1 Annealing temperatures were calculated without the +3 °C adjustment. The second PCR was run similarly but without including dPTP or 8-oxo-dGTP in the mix, and using 1 µl of the previous reaction as template, after removal of the previous template via DpnI digestion. Gel extraction was necessary to obtain a clean product.

The PCR protocol used in Step 1 of the MODAL workflow is detailed in **Chapter 6.3.2**.

6.2.5. Long overlap-based DNA assembly

Gibson isothermal reactions were performed as recommended[118]. Equimolar amounts of DNA fragments were added except for those that were about 200 bp long or less that were added at a 5x higher concentration. CPEC reactions were performed as recommended[96], with 30 cycles and replacing the annealing and extensions steps with a single combined annealing/extension step of 3 minutes at 72 °C. The amount of reaction mix used for cell transformation was the same for both methods: 1 µl for electroporation and 5 µl for chemical transformation. Yeast *in vivo* recombination was performed as described in **Chapter 6.1.3**. The assembly reactions that were performed for the collection of data for MODAL and BASIC through colony counting used a standard amount of 0.1 pmol of DNA for each part.

6.3. DNA Assembly workflow protocols

6.3.1. Storage plasmids preparation (Step 0)

The preparation of storage plasmids containing DNA parts in a standard format (also called Step 0) was performed identically for both MODAL and BASIC with the only difference being the sequences added at the flanks of the part: prefix and suffix for MODAL, and integrated prefix and integrated suffix for BASIC. A PCR was run to amplify the desired part, which was then purified using either DpnI digestion followed by PCR purification, if the amplification had been highly specific, or gel extraction if there were additional undesired products. This PCR is performed using primers that carry 5' tails encoding for the prefix/suffix or iP/iS, so that these are automatically added to the flanks of the part. The parts are then cloned in the storage plasmid, pJET 1.2, using the CloneJET PCR Cloning Kit (Thermo Scientific) according to manufacturer's protocol and are finally transformed into *E. coli* DH10B cells. As mentioned in **Chapter 4.2.3** the parts containing a PMB1 origin of replication could not be cloned in pJET 1.2 so we used different strategies. The Ori part was cloned by fusing it with the ampicillin resistance gene from pJET 1.2. The resistance gene was amplified with a PCR which also added the iP/iS sequences to its flanks, so that they would match those on the PMB1 part. The two fragments were joined using a Gibson isothermal reaction, with the iP/iS sequences acting as homologous recombination regions. The Ori+Kan part was cloned by simply circularising it which was achieved with a blunt self-ligation.

6.3.2. MODAL workflow protocol

Step 1 consists of a PCR amplification optimised for high specificity, prepared as follows:

	Volume	Stock concentration	Final concentration
H_20	33.0 µl		
HF Buffer	10.0 µl	5x	1x
DMSO	2.5 µl	100%	5%
dNTPs	1.0 µl	10 mM	0.2 mM
Forward primer	1.0 µl	10 µM	0.2 µM
Reverse primer	1.0 µl	10 µM	0.2 µM
DNA template	1.0 µl	76 nM	1.5 nM
Phusion polymerase	0.5 µl	2 U/µl	0.02 U/µl
Total	50.0 µl		

Table 6: composition of the MODAL Step 1 PCR mix.

The reaction mix is incubated in a thermal cycler and the following PCR program is run: (i) initial denaturation at 98 °C for 30 seconds, (ii) 25 cycles of denaturation at 98 °C for 10 seconds, annealing/extension at 72 °C for 30 seconds/kb, (iii) final extension at 72 °C for 5 minutes. The DNA template consists of isolated storage plasmid diluted to a concentration of 50 ng/µl for each kb of length of the whole plasmid. For example a 1 kb part cloned in pJET 1.2, which is about 3 kb long, will be stored at a concentration of 200 ng/µl. The prefix and suffix sequences that here act as priming regions are designed so that these PCR amplifications can always be run merging the annealing and extension steps in a single 72 °C step. The reagents were sourced as follows: HF Buffer (NEB), DMSO (NEB), dNTPs (Sigma-Aldrich), Phusion DNA polymerase (NEB).The products of the reaction are then purified using a DpnI digestion and a PCR purification kit, eluting in 40 µl of water.

Step 2 is performed differently depending on which reaction is chosen between Gibson isothermal, CPEC and yeast recombination, as described in **Chapter 6.2.5**. When Gibson isothermal or CPEC were chosen, 5 µl of reaction mix were used to transform chemically competent *E. coli* DH10B cells.

6.3.3. BASIC workflow protocol

Step 1 of the BASIC workflow begins with a simultaneous digestion and ligation containing the following reagents:

	Volume	Stock concentration	Final concentration
H_20	8.5 µl		
ATP	3.0 µl	10 mM	1 mM
NEBuffer 4	3.0 µl	10x	1x
BSA	3.0 µl	10x	1x
iP linker	5.0 µl	1 µM	166.6 nM
iS linker	5.0 µl	1 µM	166.6 nM
DNA part	1.0 µl	76 nM	2.5 nM
BsaI-HF	1.0 µl	20 U/µl	0.66 U/µl
T4 DNA ligase	0.5 µl	400 U/µl	6.6 U/µl
Total	30.0 µl		

Table 7: composition of the BASIC Step 1 digestion/ligation mix.

The mix is incubated in a thermocycler with the following program: 37 °C for 1 hour, 20 °C for 20 minutes, 65 °C for 20 minutes. The DNA part is carried on a storage plasmid, and the stock solution of isolated plasmid is prepared at a concentration of 50 ng/µl for each kb of length of the whole plasmid. The linker oligonucleotides are chosen to add the appropriate linker sequence on the flanks of the part. The other reagents were sourced as follows: NEBuffer 4 (NEB), ATP (Sigma-Aldrich), Bovine Serum Albumin (NEB), BsaI-HF (NEB), T4 DNA ligase (NEB). After the incubation the reaction mix is purified using an Agencourt AMPure XP kit following manufacturer's protocol. 54 µl of beads are used for each reaction and the DNA is eluted in 40 µl of water. When transferring the eluate to a clean microtube it is recommended to move 30 µl only, leaving 10 µl behind, to ensure that no beads are carried over.

The Step 2 assembly mix is prepared as follows:

	Volume	Stock concentration	Final concentration
Part (each)	1.0 µl	~1.5 nM	~0.15 nM
BSA	1.0 µl	10x	1x
NEBuffer 4	1.0 µl	10x	1x
H_2O	up to 10.0 µl		
Total	10.0 µl		

Table 8: composition of the BASIC Step 2 assembly mix.

Incubate the mix at 50 °C for 45 minutes. The parts prepared during Step 1 are estimated to have approximately a 1.5 nM concentration assuming 80% recovery efficiency of the purification step. The volume of water is adjusted depending on the number of parts assembled simultaneously. To assemble more than 8 parts increase the total volume of the reaction and adjust all other reagents accordingly. After incubation 5 µl of the mix were used to transform chemically competent *E.coli* DH5α cells.

6.4. Data collection

6.4.1. Quantification of GFP and RFP expression

Plate reader readings for population-level GFP and RFP expression were taken on *E. coli* cultures grown to mid-exponential phase in LB liquid medium using a BMG Labtech Polarstar Omega plate reader. The excitation and emission settings were as follows: GFP excitation 485nm, GFP emission 510nm, RFP excitation 584nm, RFP emission 610nm.

Flow cytometry assays for single-cell GFP and RFP expression were taken with a modified Becton Dickinson FACScan flow cytometer equipped with both a blue laser (488 nm) for GFP excitation and a green laser (561 nm) for RFP excitation. Green fluorescence was detected with a 530 nm band pass filter (FL1) with gain 890. Red fluorescence was detected with a 610 nm filter (FL5) with gain 850. Data analysis for *E. coli* cultures was performed using Cyflogic software (CyFlo Ltd.), applying a gate on forward and side scatter to only include readings from bacterium-sized single particles. Mean FL1 and FL5 values were used as a measure of the amount of GFP and RFP per cell, respectively. Data analysis for *S. cerevisiae* cultures was performed using FlowJo software (Treestar Inc.), using appropriate forward and side scatter gating for yeast-sized single particles. Plasmids containing a 2-μ origin of replication show large variation of copy-number in S. cerevisiae, making it difficult to accurately measure gene expression. To account for this, we normalised our data using constitutive RFP expression from the same plasmid: a gate was applied to all samples for mid-range RFP expression (FL5), and the geometric mean of FL1 values for these particles was used as a measure of the amount of GFP per cell. RFP-based gating was used to normalize for plasmid copy number variation within the population as described previously[119]. An example of this gating is shown in **Figure 48**.

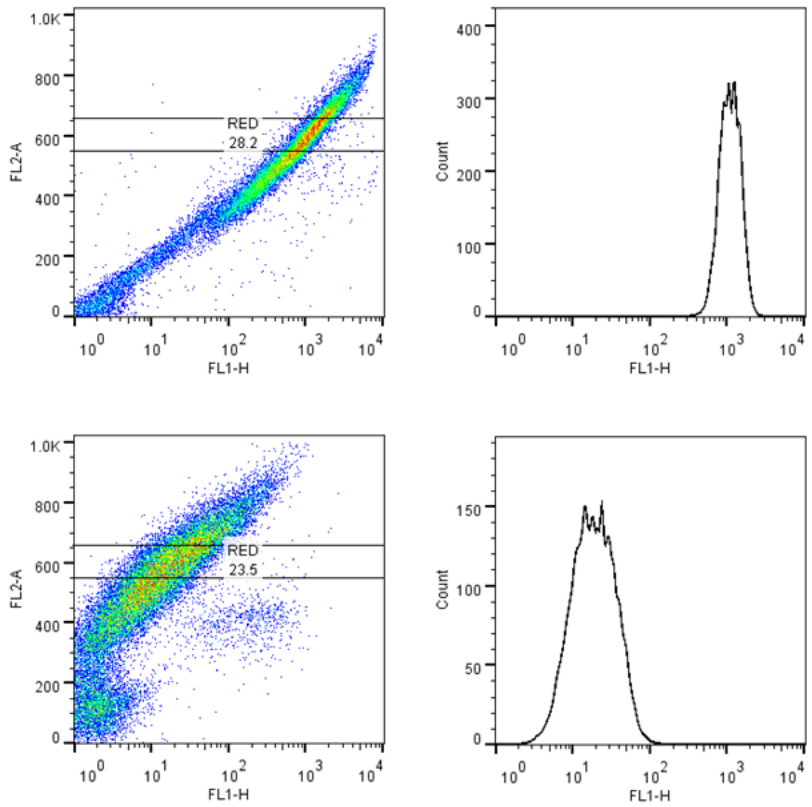


Figure 48: flow cytometry gating for the MODAL mutation library experiment. 2-micron yeast plasmids show large copy number variation in *S. cerevisiae*, making it difficult to accurately quantify gene expression from such plasmids. To account for copy number variations when using flow cytometry, we normalised the data using constitutive RFP expression from the same plasmid. The examples here show two promoters giving different gene expression (as measured by GFP in the FL1 channel). By gating samples for a narrow mid-range region of FL5 (shown here in the 2D dot-plots as FL2-A) expression, we only sample cells within a defined RFP expression level. The FL1 measurement of these gated cells (histograms) is then used as the GFP measurement per cell. Data analysis was performed in FlowJo (Treestar Inc.). Figure from Casini *et al.*[105]

6.4.2. Colony counting

Colony counting was employed to gather the data for three separate experiments, shown in **Figure 28**, **Figure 40** and **Figure 47**. All assembly reactions part of the same experiment were run on the same day using the same batch of competent cells and plates. A standard amount of 0.1 pmol of DNA for each part was used for CPEC, Gibson isothermal and yeast recombination. The amount of DNA used in BASIC reactions is about 0.15 fmol, which is much lower because it is not amplified through PCR. For all experiments, transformed cells were plated in three different dilutions: 100 µl of solution were spread, containing 90%, 9% or 1% cells in LB or SOC liquid medium. After growth the plate with the most colonies that were still clearly distinguishable was chosen for imaging using a Fuji FLA-5000 scanner: a blue (473 nm) laser and Fluorescein isothiocyanate (FITC) filter were used to visualise GFP expressing colonies and a green (532 nm) laser and a long pass green (LPG) filter for RFP expressing colonies. Images were overlaid and aligned to correct for aberration using ImageJ software (NIH). The colonies on each plate were counted manually, gathering both the total number and the number of colonies of each colour separately, representing the fluorescent reporters being expressed in the cells: white for those not expressing any reporter, green for those expressing GFP, red for RFP and yellow for both GFP and RFP. The numbers were then adjusted according to the dilution factor used to estimate the number of colonies from the whole transformation. For each plate, representing a single assembly reaction, assembly efficiency was measured as the total number of colonies obtained from the transformation, and the assembly accuracy was calculated as the percentage of total colonies that had the correct colour.

7. Bibliography

1. Quan, J. & Tian, J. Circular polymerase extension cloning of complex gene libraries and pathways. *PLoS One* **4,** e6441 (2009).

2. Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. **6,** 12–16 (2009).

3. Kok, S. De *et al.* Rapid and Reliable DNA Assembly via Ligase Cycling Reaction. (2014).

4. Li, M. & Elledge, S. Harnessing homologous recombination in vitro to generate recombinant DNA via SLIC. *Nat. Methods* **4,** 251–256 (2007).

5. Keasling, J. D. Synthetic biology for synthetic chemistry. *ACS Chem. Biol.* **3,** 64–76 (2008).

6. Goeddel, D. V *et al.* Direct expression in Escherichia coli of a DNA sequence coding for human growth hormone. *Nature* **281,** 544–8 (1979).

7. Perlak, F. J. *et al.* Insect resistant cotton plants. *Biotechnology. (N. Y).* **8,** 939–43 (1990).

8. Gardner, T. S., Cantor, C. R. & Collins, J. J. Construction of a genetic toggle switch in Escherichia coli. *Nature* **403,** 339–42 (2000).

9. Elowitz, M. B. & Leibler, S. A synthetic oscillatory network of transcriptional regulators. *Nature* **403,** 335–8 (2000).

10. Tabor, J. J. *et al.* A synthetic genetic edge detection program. *Cell* **137,** 1272–81 (2009).

11. Prindle, A. *et al.* A sensing array of radically coupled genetic “biopixels”. *Nature* **481,** 39–44 (2012).

12. Purnick, P. E. M. & Weiss, R. The second wave of synthetic biology: from modules to systems. *Nat. Rev. Mol. Cell Biol.* **10,** 410–22 (2009).

13. Khalil, A. & Collins, J. Synthetic biology: applications come of age. *Nat. Rev. Genet.* **11,** 367–379 (2010).

14. Menzella, H. G. & Reeves, C. D. Combinatorial biosynthesis for drug development. *Curr. Opin. Microbiol.* **10,** 238–45 (2007).

15. Paddon, C. J. *et al.* High-level semi-synthetic production of the potent antimalarial artemisinin. *Nature* **496,** 528–32 (2013).

16. Amyris. at <http://www.amyris.com/>

17. Ginkgo Bioworks. at <http://ginkgobioworks.com/>

18. Synthetic Genomics. at <http://www.syntheticgenomics.com/>

19. Cameron, D. E., Bashor, C. J. & Collins, J. J. A brief history of synthetic biology. *Nat. Rev. Microbiol.* **12,** 381–90 (2014).

20. Church, G. M., Elowitz, M. B., Smolke, C. D., Voigt, C. a & Weiss, R. Realizing the potential of synthetic biology. *Nat. Rev. Mol. Cell Biol.* **15,** 289–94 (2014).

21. Cohen, S. N., Chang, A. C. Y., Boyert, H. W. & Hellingt, R. B. Biologically Functional Bacterial Plasmids In Vitro. **70,** 3240–3244 (1973).

22. International Genetically Engineered Machine Competition. at <http://igem.org/>

23. Isaacs, F. J. *et al.* Precise manipulation of chromosomes in vivo enables genome-wide codon replacement. *Science* **333,** 348–53 (2011).

24. Jiang, W., Bikard, D., Cox, D., Zhang, F. & Marraffini, L. A. RNA-guided editing of bacterial genomes using CRISPR-Cas systems. *Nat. Biotechnol.* **31,** 233–9 (2013).

25. Gibson, D. G. *et al.* Creation of a bacterial cell controlled by a chemically synthesized genome. *Science* **329,** 52–6 (2010).

26. Arkin, A. Setting the standard in synthetic biology. *Nat. Biotechnol.* **26,** 771–4 (2008).

27. Blake, W. J. *et al.* Pairwise selection assembly for sequence-independent construction of long-length DNA. *Nucleic Acids Res.* **38,** 2594–602 (2010).

28. Chen, W.-H., Qin, Z.-J., Wang, J. & Zhao, G.-P. The MASTER (methylation-assisted tailorable ends rational) ligation method for seamless DNA assembly. *Nucleic Acids Res.* **41,** e93 (2013).

29. Engler, C., Gruetzner, R., Kandzia, R. & Marillonnet, S. Golden gate shuffling: a one-pot DNA shuffling method based on type IIs restriction enzymes. *PLoS One* **4,** e5553 (2009).

30. Liang, J., Chao, R., Abil, Z., Bao, Z. & Zhao, H. FairyTALE: A High-Throughput TAL Effector Synthesis Platform. *ACS Synth. Biol.* (2013). doi:10.1021/sb400109p

31. Jakobi, A. J. & Huizinga, E. G. A rapid cloning method employing orthogonal end protection. *PLoS One* **7,** e37617 (2012).

32. Engler, C., Kandzia, R. & Marillonnet, S. A one pot, one step, precision cloning method with high throughput capability. *PLoS One* **3,** e3647 (2008).

33. Weber, E., Engler, C., Gruetzner, R., Werner, S. & Marillonnet, S. A modular cloning system for standardized assembly of multigene constructs. *PLoS One* **6,** e16765 (2011).

34. Horton, R. M., Hunt, H. D., Ho, S. N., Pullen, J. K. & Pease, L. R. Engineering hybrid genes without the use of restriction enzymes: gene splicing by overlap extension. *Gene* **77,** 61–8 (1989).

35. Shevchuk, N. a *et al.* Construction of long DNA molecules using long PCR-based fusion of several fragments simultaneously. *Nucleic Acids Res.* **32,** e19 (2004).

36. Cha-aim, K., Fukunaga, T., Hoshida, H. & Akada, R. Reliable fusion PCR mediated by GC-rich overlap sequences. *Gene* **434,** 43–9 (2009).

37. USER Cloning. at <https://www.neb.com/applications/cloning-and-synthetic-biology/user-cloning>

38. Bitinaite, J. *et al.* USER friendly DNA engineering and cloning method by uracil excision. *Nucleic Acids Res.* **35,** 1992–2002 (2007).

39. Nørholm, M. H. H. A mutant Pfu DNA polymerase designed for advanced uracil-excision DNA engineering. *BMC Biotechnol.* **10,** 21 (2010).

40. Zou, R., Zhou, K., Stephanopoulos, G. & Too, H. P. Combinatorial Engineering of 1-Deoxy-D-Xylulose 5-Phosphate Pathway Using Cross-Lapping In Vitro Assembly (CLIVA) Method. *PLoS One* **8,** e79557 (2013).

41. Too, P. H.-M., Zhu, Z., Chan, S.-H. & Xu, S. Engineering Nt.BtsCI and Nb.BtsCI nicking enzymes and applications in generating long overhangs. *Nucleic Acids Res.* **38,** 1294–303 (2010).

42. Wang, R.-Y., Shi, Z.-Y., Guo, Y.-Y., Chen, J.-C. & Chen, G.-Q. DNA fragments assembly based on nicking enzyme system. *PLoS One* **8,** e57943 (2013).

43. Schmid-Burgk, J. L. *et al.* Rapid hierarchical assembly of medium-size DNA cassettes. *Nucleic Acids Res.* **40,** e92 (2012).

44. Daniel G Gibson, Lei Young, Ray-Yuan Chuang, J Craig Venter, Clyde A Hutchison, H. O. S. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6,** 343–345 (2009).

45. Gibson, D. G., Smith, H. O., Hutchison, C. A., Venter, J. C. & Merryman, C. Chemical synthesis of the mouse mitochondrial genome. *Nat. Methods* **7,** 901–3 (2010).

46. Vroom, J. A. & Wang, C. L. Modular construction of plasmids through ligation-free assembly of vector components with oligonucleotide linkers. *Biotechniques* **44,** 924–6 (2008).

47. Ramon, A. & Smith, H. O. Single-step linker-based combinatorial assembly of promoter and gene cassettes for pathway engineering. *Biotechnol. Lett.* **33,** 549–55 (2011).

48. Pachuk, C. J. *et al.* Chain reaction cloning: a one-step method for directional ligation of multiple DNA fragments. *Gene* **243,** 19–25 (2000).

49. Chao, R., Yuan, Y. & Zhao, H. Recent Advances in DNA Assembly Technologies. *FEMS Yeast Res.* (2014). doi:10.1111/1567-1364.12171

50. Gietz, R. D. & Schiestl, R. H. Quick and easy yeast transformation using the LiAc/SS carrier DNA/PEG method. *Nat. Protoc.* **2,** 35–7 (2007).

51. Gietz, R. D. & Schiestl, R. H. Large-scale high-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. *Nat. Protoc.* **2,** 38–41 (2007).

52. Gibson, D. G. Synthesis of DNA fragments in yeast by one-step assembly of overlapping oligonucleotides. *Nucleic Acids Res.* **37,** 6984–90 (2009).

53. Shao, Z., Zhao, H. & Zhao, H. DNA assembler, an in vivo genetic method for rapid construction of biochemical pathways. *Nucleic Acids Res.* **37,** e16 (2009).

54. Gibson, D. G. *et al.* Complete chemical synthesis, assembly, and cloning of a Mycoplasma genitalium genome. *Science* **319,** 1215–20 (2008).

55. Noskov, V. N. *et al.* Assembly of Large, High G+C Bacterial DNA Fragments in Yeast. *ACS Synth. Biol.* **1,** 267–273 (2012).

56. Kuijpers, N. G. *et al.* A versatile, efficient strategy for assembly of multi-fragment expression vectors in Saccharomyces cerevisiae using 60-bp synthetic recombination sequences. *Microb. Cell Fact.* **12,** 47 (2013).

57. Turan, S. *et al.* Recombinase-mediated cassette exchange (RMCE): traditional concepts and current challenges. *J. Mol. Biol.* **407,** 193–221 (2011).

58. Hartley, J. L., Temple, G. F. & Brasch, M. A. DNA cloning using in vitro site-specific recombination. *Genome Res.* **10,** 1788–95 (2000).

59. Colloms, S. D. *et al.* Rapid metabolic pathway assembly and modification using serine integrase site-specific recombination. *Nucleic Acids Res.* 1–10 (2013). doi:10.1093/nar/gkt1101

60. Moriarity, B. S. *et al.* Modular assembly of transposon integratable multigene vectors using RecWay assembly. *Nucleic Acids Res.* **41,** e92 (2013).

61. Turan, S. *et al.* Expanding Flp-RMCE options: the potential of Recombinase Mediated Twin-Site Targeting (RMTT). *Gene* **546,** 135–144 (2014).

62. Missirlis, P. I., Smailus, D. E. & Holt, R. a. A high-throughput screen identifying sequence and promiscuity characteristics of the loxP spacer region in Cre-mediated recombination. *BMC Genomics* **7,** 73 (2006).

63. Sasaki, Y. *et al.* Evidence for high specificity and efficiency of multiple recombination signals in mixed DNA cloning by the Multisite Gateway system. *J. Biotechnol.* **107,** 233–243 (2004).

64. Shetty, R. P., Endy, D. & Knight, T. F. Engineering BioBrick vectors from BioBrick parts. *J. Biol. Eng.* **2,** 5 (2008).

65. Ellis, T., Adie, T. & Baldwin, G. S. DNA assembly for synthetic biology: from parts to pathways and beyond. *Integr. Biol. (Camb).* **3,** 109–18 (2011).

66. Shetty, R., Lizarazo, M., Rettberg, R. & Knight, T. F. *Assembly of BioBrick standard biological parts using three antibiotic assembly. Methods Enzymol.* **498,** 311–26 (Elsevier Inc., 2011).

67. Xu, P., Vansiri, A., Bhan, N. & Koffas, M. a G. ePathBrick: a synthetic biology platform for engineering metabolic pathways in E. coli. *ACS Synth. Biol.* **1,** 256–66 (2012).

68. Norville, J. E. *et al.* Introduction of customized inserts for s-treamlined assembly and optimization of BioBrick synthetic genetic circuits. *J. Biol. Eng.* **4,** 17 (2010).

69. Silva-Rocha, R. *et al.* The Standard European Vector Architecture (SEVA): a coherent platform for the analysis and deployment of complex prokaryotic phenotypes. *Nucleic Acids Res.* 1–10 (2012). doi:10.1093/nar/gks1119

70. Litcofsky, K. D., Afeyan, R. B., Krom, R. J., Khalil, A. S. & Collins, J. J. Iterative plug-and-play methodology for constructing and modifying synthetic gene networks. *Nat. Methods* **9,** 1077–80 (2012).

71. Anderson, J. C. *et al.* BglBricks: A flexible standard for biological part assembly. *J. Biol. Eng.* **4,** 1 (2010).

72. Leguia, M., Brophy, J. A., Densmore, D., Asante, A. & Anderson, J. C. 2Ab Assembly: a Methodology for Automatable, High-Throughput Assembly of Standard Biological Parts. *J. Biol. Eng.* **7,** 2 (2013).

73. Engler, C. *et al.* A Golden Gate Modular Cloning Toolbox for Plants. *ACS Synth. Biol.* (2014). doi:10.1021/sb4001504

74. Sarrion-Perdigones, A. *et al.* GoldenBraid: an iterative cloning system for standardized assembly of reusable genetic modules. *PLoS One* **6,** e21622 (2011).

75. Sarrion-Perdigones, A. *et al.* GoldenBraid2.0: A comprehensive DNA assembly framework for Plant Synthetic Biology. *Plant Physiol.* (2013). doi:10.1104/pp.113.217661

76. Sleight, S. C., Bartley, B. A., Lieviant, J. A. & Sauro, H. M. In-Fusion BioBrick assembly and re-engineering. *Nucleic Acids Res.* **38,** 2624–36 (2010).

77. Sleight, S. C. & Sauro, H. M. Randomized BioBrick Assembly: A Novel DNA Assembly Method for Randomizing and Optimizing Genetic Circuits and Metabolic Pathways. *ACS Synth. Biol.* (2013). doi:10.1021/sb4000542

78. Guye, P., Li, Y., Wroblewska, L., Duportet, X. & Weiss, R. Rapid, modular and reliable construction of complex mammalian gene circuits. *Nucleic Acids Res.* **41,** e156 (2013).

79. Torella, J. P. *et al.* Rapid construction of insulated genetic circuits via synthetic sequence-guided isothermal assembly. *Nucleic Acids Res.* 1–9 (2013). doi:10.1093/nar/gkt860

80. Shao, Z., Luo, Y. & Zhao, H. Rapid characterization and engineering of natural product biosynthetic pathways via DNA assembler. *Mol. Biosyst.* **7,** 1056–9 (2011).

81. Shoemaker, D. D., Lashkari, D. A., Morris, D., Mittmann, M. & Davis, R. W. Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar-coding strategy. *Nat. Genet.* **14,** 450–6 (1996).

82. Li, M. V *et al.* HomeRun Vector Assembly System: A Flexible and Standardized Cloning System for Assembly of Multi-Modular DNA Constructs. *PLoS One* **9,** e100948 (2014).

83. Appleton, E., Tao, J., Haddock, T. & Densmore, D. Interactive assembly algorithms for molecular cloning. *Nat. Methods* **11,** 657–62 (2014).

84. Hillson, N. J., Rosengarten, R. D. & Keasling, J. D. j5 DNA Assembly Design Automation Software. *ACS Synth. Biol.* **1,** 14–21 (2012).

85. Chuang, L.-Y., Cheng, Y.-H. & Yang, C.-H. Specific primer design for the polymerase chain reaction. *Biotechnol. Lett.* **35,** 1541–9 (2013).

86. Pozhitkov, A. E., Tautz, D. & Noble, P. A. Oligonucleotide microarrays: widely applied--poorly understood. *Brief. Funct. Genomic. Proteomic.* **6,** 141–8 (2007).

87. Xu, Q., Schlabach, M. R., Hannon, G. J. & Elledge, S. J. Design of 240,000 orthogonal 25mer DNA barcode probes. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 2289–94 (2009).

88. Untergasser, A. *et al.* Primer3--new capabilities and interfaces. *Nucleic Acids Res.* **40,** e115 (2012).

89. Wernersson, R. & Nielsen, H. B. OligoWiz 2.0--integrating sequence feature annotation into the design of microarray probes. *Nucleic Acids Res.* **33,** W611–5 (2005).

90. Pfleger, B. F., Pitera, D. J., Smolke, C. D. & Keasling, J. D. Combinatorial engineering of intergenic regions in operons tunes expression of multiple genes. *Nat. Biotechnol.* **24,** 1027–32 (2006).

91. Ng, D. T. W. & Sarkar, C. a. NP-Sticky: a web server for optimizing DNA ligation with non-palindromic sticky ends. *J. Mol. Biol.* **426,** 1861–9 (2014).

92. SantaLucia, J. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl. Acad. Sci. U. S. A.* **95,** 1460–5 (1998).

93. Casini, A. *et al.* R2oDNA Designer: Computational Design of Biologically Neutral Synthetic DNA Sequences. *ACS Synth. Biol.* **3,** 525–8 (2014).

94. Andronescu, M., Aguirre-Hernández, R., Condon, A. & Hoos, H. H. RNAsoft: A suite of RNA secondary structure prediction and design software tools. *Nucleic Acids Res.* **31,** 3416–22 (2003).

95. Torella, J. P. *et al.* Unique nucleotide sequence-guided assembly of repetitive DNA parts for synthetic biology applications. *Nat. Protoc.* **9,** 2075–2089 (2014).

96. Quan, J. & Tian, J. Circular polymerase extension cloning for high-throughput cloning of complex and combinatorial DNA libraries. *Nat. Protoc.* **6,** 242–51 (2011).

97. Zaccolo, M., Williams, D. M., Brown, D. M. & Gherardi, E. An approach to random mutagenesis of DNA using mixtures of triphosphate derivatives of nucleoside analogues. *J. Mol. Biol.* **255,** 589–603 (1996).

98. Salis, H. M., Mirsky, E. A. & Voigt, C. A. Automated design of synthetic ribosome binding sites to control protein expression. *Nat. Biotechnol.* **27,** 946–50 (2009).

99. Mutalik, V. K. *et al.* Quantitative estimation of activity and quality for collections of functional genetic elements. *Nat. Methods* **10,** 347–53 (2013).

100. Mutalik, V. K. *et al.* Precise and reliable gene expression via standard transcription and translation initiation elements. *Nat. Methods* **10,** 354–60 (2013).

101. Lou, C., Stanton, B., Chen, Y.-J., Munsky, B. & Voigt, C. A. Ribozyme-based insulator parts buffer synthetic circuits from genetic context. *Nat. Biotechnol.* **30,** 1137–42 (2012).

102. Qi, L., Haurwitz, R. E., Shao, W., Doudna, J. A. & Arkin, A. P. RNA processing enables predictable programming of gene expression. *Nat. Biotechnol.* **30,** 1002–6 (2012).

103. Crook, N. C., Freeman, E. S. & Alper, H. S. Re-engineering multicloning sites for function and convenience. *Nucleic Acids Res.* **39,** e92 (2011).

104. Davis, J. H., Rubin, A. J. & Sauer, R. T. Design, construction and characterization of a set of insulated bacterial promoters. *Nucleic Acids Res.* **39,** 1131–41 (2011).

105. Casini, A. *et al.* One-pot DNA construction for synthetic biology: the Modular Overlap-Directed Assembly with Linkers (MODAL) strategy. *Nucleic Acids Res.* **42,** e7 (2014).

106. Horspool, D. R., Coope, R. J. & Holt, R. a. Efficient assembly of very short oligonucleotides using T4 DNA Ligase. *BMC Res. Notes* **3,** 291 (2010).

107. Dieffenbach, C. W., Lowe, T. M. & Dveksler, G. S. General concepts for PCR primer design. *PCR Methods Appl.* **3,** S30–7 (1993).

108. Roberts, R. J., Vincze, T., Posfai, J. & Macelis, D. REBASE--a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res.* **38,** D234–6 (2010).

109. Sambrook, J. & Russell, D. W. *Molecular Cloning: A Laboratory Manual*. 1.160–1.161 and 1.88–1.89 (Cold Spring Harbor Laboratory Press,U.S., 2001).

110. Kahl, L. J. & Endy, D. A survey of enabling technologies in synthetic biology. *J. Biol. Eng.* **7,** 13 (2013).

111. FAQ: Can ≤ 200 bp dsDNA fragments be assembled by this method? at <https://www.neb.com/faqs/1/01/01/can-le-200-bp-dsdna-fragments-be-assembled-by-this-method>

112. FAQ: How many fragments of DNA can be assembled in one reaction? at <https://www.neb.com/faqs/1/01/01/how-many-fragments-of-dna-can-be-assembled-in-one-reaction>

113. Salis, H. M. The ribosome binding site calculator. *Methods Enzymol.* **498,** 19–42 (2011).

114. Gen9. at <https://www.gen9bio.com/>

115. FAQ: What is the error rate of Phusion® High-Fidelity DNA Polymerase? at <https://www.neb.com/faqs/2012/09/06/what-is-the-error-rate-of-phusion-reg-high-fidelity-dna-polymerase>

116. Reyon, D. *et al.* FLASH assembly of TALENs for high-throughput genome editing. *Nat. Biotechnol.* **30,** 460–5 (2012).

117. Sikorski, R. S. & Hieter, P. A system of shuttle vectors and yeast host strains designed for efficient manipulation of DNA in Saccharomyces cerevisiae. *Genetics* **122,** 19–27 (1989).

118. Gibson, D. One-step enzymatic assembly of DNA molecules up to several hundred kilobases in size. *Protoc. Exch.* (2009). doi:10.1038/nprot.2009.77

119. Liang, J. C., Chang, A. L., Kennedy, A. B. & Smolke, C. D. A high-throughput, quantitative cell-based screen for efficient tailoring of RNA device activity. *Nucleic Acids Res.* **40,** 1–14 (2012).