Rewriting the genome of *Escherichia coli*

This thesis is submitted to the University of Cambridge for the degree of Doctor of Philosophy

by

Daniel de la Torre Martín

Homerton College

March 2020

Preface

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text.

It does not exceed the prescribed word limit set by the Biology Degree Committee.

Thesis summary

Rewriting the genome of *Escherichia coli* - Daniel de la Torre Martín

Our recently acquired ability to synthesize DNA at large scale is opening the door to writing entire genomes; this constitutes a powerful approach to address fundamental biological questions, and may enable the creation of designer organisms with useful properties.

One interesting avenue for investigation is the creation of recoded genomes, where codons are substituted by their synonyms. Compression of synonymous codon boxes may provide blank spaces in the quasi-universal genetic code, and these may be amenable for reassignment into unnatural amino acids, in synergy with parallel efforts to engineer the protein translation machinery. Recoding genomes is subject to both biological and technical challenges. First, synonymous codon choice genome-wide is not trivial, and identifying suitable synonymous replacements is challenging. Second, synthetic DNA pieces need to be assembled into fragments of increasing size, and ultimately implemented inside a target host.

Here, recently reported strategies for genome engineering in *E. coli* (REXER and GENESIS) are extended and used to create a synthetic, recoded *E. coli* genome. **Chapter 2** describes a strategy for assembling large natural genomic DNA pieces into BACs that are substrates for genome replacement. Experiments with these BACs serve to validate and improve REXER and GENESIS, and lay out a strategy for genome replacement. **Chapter 3** employs these strategies for the synthesis and assembly of a recoded *E. coli* genome where all annotated instances of serine codons TCG and TCA, and stop codon TAG, are systematically replaced by their synonyms. The resulting strain, Syn61, provides a unique platform for exploring sense codon reassignment *in vivo*, and reassignment of the TCG codon to both natural and unnatural amino acids is demonstrated. Finally, **Chapter 4** extends the toolkit for genome engineering in *E. coli*, and provides technologies for splitting the genome into pairs of chromosomes, as well as performing precise inversions and translocations. These technologies are used to precisely combine synthetic sections from distinct strains into a single genome. Together, these technologies may provide a foundation for future genome synthesis endeavours.

Acknowledgments

Firstly, I would like to thank Professor Jason Chin for giving me the opportunity to work in his laboratory, and for the invaluable guidance, knowledge and opportunities that he has provided me with over the course of my PhD.

I have been fortunate to share my time in the lab with a wonderful team of people. I especially want to thank Kai, who inspired me every day with his crazy creative genius. Julius, for pushing our science forward, the laughs and the good vibes. Louise, with whom I shared the PhD learning journey, successes and frustrations. Wes, for providing infinite knowledge, occasionally eggs, and for being way ahead of the game.

I have made some invaluable friends during my time at the LMB. The list is long, but I must mention Daniele (an endless source of entertainment), Charlie (who almost drowned with me in Yorkshire), Shan, Wolfgang, Nicolas, Zakir, Gianluca, Donny, Chris, Jakob, Sebastian, Julian, Vaclav, Inja, Solomon, Franz, Adam, Yonka, Mart, everyone at the GSA… They have looked after me, made sure I stayed happy and sane, and the time we shared together has infinitely enriched my time in Cambridge.

I also want to thank Emma for taking care of me, as well as all the LMB administrative and support staff, including those at the media kitchen, glass wash, stores, reception, IT, the workshops, operations, lab and domestic services, the restaurant and the scientific facilities, with special thanks to the mass spectrometry facility. They take care of everything so we can focus on the science, and their work is greatly appreciated.

I am immensely grateful to Eze for being there for me unconditionally, and for being tremendo perret. I can't imagine the last few years without him. Also to Gonzalo, who sparks joy in my life every day. Edo, for coming in the same indivisible pack as me, and for being my partner in crime for way too long. Kima, for being an amazing housemate and periodically sending me into self-isolation. Pamela and Ana, Pena y Pánico, for maintaining basal levels of latinness in my life, for the support when I needed it, and for the great times we shared together. Luz, who I definitely have not spent enough time with.

Finally, and most importantly, I want to thank my family, for their unconditional love, and for always going beyond what was conceivable to support me in every way they could, both inside and outside my academic journey. This dissertation is, effectively, the product of their work over the years, and it is dedicated to them.

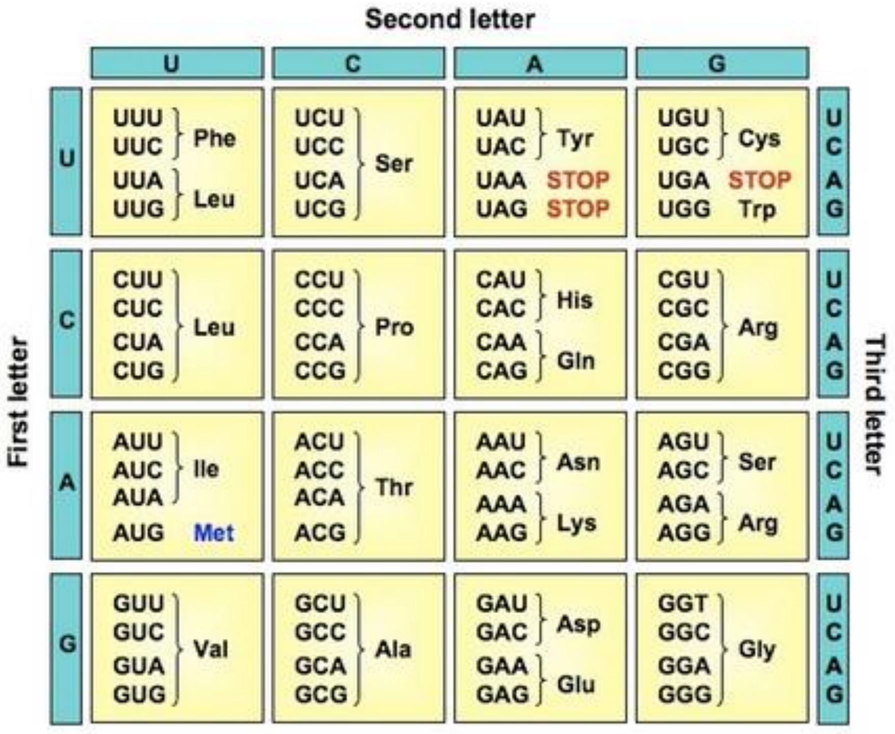
Table of contents

Chapter 1 – Introduction

1.1 – The genetic code

The genetic code defines the relationship between the information encoded in nucleic acids and amino acid identity. During ribosomal protein translation, genetic information is read in groups of three nucleotides from a messenger RNA (mRNA) sequence; these informational units are termed codons. At the ribosomal decoding centre, complementarity between the bases of a codon and the anticodon of its corresponding transfer RNA (tRNA), charged with its cognate amino acid, result in the template-directed addition of amino acids to polypeptide chains.

Because codons are composed of three nucleotides, and genetic polymers (DNA and RNA) result from the concatenation of four distinct bases, the genetic code is composed of 64 (4^3) codons. There are, however, 20 canonical amino acids; this results in a non-equitable relationship between the number of codons and protein building blocks. Indeed, 18 out of the 20 amino acids are coded for by multiple codons, which are considered synonymous. Additionally, three codons encode translation termination signals – their recognition by their cognate release factors stops translation, and results in the release of the polypeptide chain from the ribosome.



	Second letter				
First letter	U	C	A	G	Third letter
U	UUU Phe	UCU Ser	UAU Tyr	UGU Cys	U
	UUC	UCC	UAC	UGC	C
	UUA Leu	UCA	UAA STOP	UGA STOP	A
	UUG	UCG	UAG STOP	UGG Trp	G
C	CUU Leu	CCU Pro	CAU His	CGU Arg	U
	CUC	CCC	CAC	CGC	C
	CUA	CCA	CAA Gln	CGA	A
	CUG	CCG	CAG	CGG	G
A	AUU Ile	ACU Thr	AAU Asn	AGU Ser	U
	AUC	ACC	AAC	AGC	C
	AUA	ACA	AAA Lys	AGA Arg	A
	AUG Met	ACG	AAG	AGG	G
G	GUU Val	GCU Ala	GAU Asp	GGT Gly	U
	GUC	GCC	GAC	GGC	C
	GUA	GCA	GAA Glu	GGA	A
	GUG	GCG	GAG	GGG	G

Figure 1.1 – The standard genetic code. Adapted from www.vce.bioninja.com.au.

The genetic code is remarkably conserved across all domains of life; generally, the same codons encode the same amino acids in all organisms. As the nature of the genetic code was being elucidated in the 60's[1], multiple theories emerged in an attempt to explain this phenomenon. The stereochemical theory of genetic code origin postulated that codon-amino acid relationships were intrinsically defined by specific chemical interactions between amino acids and their corresponding tRNAs[2,3]. The 'frozen accident' theory proposed that that the current genetic code may have originated serendipitously very early in evolution, and that subsequent deviations from its canonical form would cause mistranslation, which would be selected against[4]. Other theories suggest that over the course of evolution selective pressures may have shaped the code into a most-efficient configuration, while minimizing the fitness cost of mistranslation[5,6]. Even the implications of horizontal gene transfer in shaping the early genetic code have been explored[7,8]. These explanations are not mutually exclusive, and it seems likely that several of these factors played a role; the origin of the genetic code remains a fundamental biological question[9,10].

As our ability to read DNA sequences has improved it has become increasingly clear that, while the genetic code is near-universal, it is also in continuous evolution, and several deviations from its canonical form are now known. For instance, in vertebrate mitochondrial genomes, UGA encodes tryptophan instead of a termination signal, and AUA encodes methionine instead of isoleucine[11]. The reassignment of UGA as tryptophan has also been observed in the *Mycoplasma* genus[12]. Moreover, in *Mycoplasma capricolum* the CGG codon, which normally encodes arginine, appears to be unassigned[13]. Alternative codon assignments are also observed, for instance, in the *Candida* genus[14,15] and in ciliates[16].

Spontaneous mutations in tRNAs may result in decoding of alternative codons. This can result in the generation of missense and nonsense tRNA suppressors, which lead to perturbations in the relationship between codons, amino acids and stop signals[17,18]. Mutations and modifications to the anticodons of tRNAs are considered to be one of the major driving forces behind codon reassignment, and may be under selective pressures by factors such as genome minimisation[19]. These have been reviewed previously[20].

Several models have been proposed to explain deviations from the canonical genetic code. The 'codon capture' hypothesis postulates codons are first erased from the genome through mutations, generating blank spaces which can then be occupied by alternative translational components[21]. An alternative hypothesis is that of the 'ambiguous intermediate', which states that during genetic code evolution there is a transition state in which single codons can encode more than one amino acid[22,23]. Under certain circumstances, ambiguity may confer a selective advantage[24], and subsequent adaptations may consolidate reassignment to a particular amino acid. Either way, these examples illustrate the natural flexibility of the genetic code. Over time, they have encouraged work aimed at recapitulating and expanding such evolutionary processes in the laboratory.

1.2 – Reprogramming the genetic code

Generally, living organisms synthesize proteins using 20 amino acids; their combination generates countless different and unique proteins, and gives rise to vast phenotypic diversity. However, the 20 amino acids constitute a rather limited set of building blocks, which occupy a minor fraction of the possible chemical space. Consequently, a growing body of work has focused on expanding this paradigm. As the mechanisms of the processes underlying protein translation were elucidated, opportunities for their reprogramming became apparent. This led to efforts to engineer tRNAs, aminoacyl-tRNA synthetases and the ribosome, with views of ultimately repurposing the translational machinery towards the biologically encoded synthesis of unnatural polymers[25–27].

1.2.1 Genetically encoding non-canonical amino acids

Much of the work on genetic code expansion has focused on genetically encoding unnatural amino acids into proteins. In cells, this requires the introduction of additional aminoacyl-tRNA synthetase/tRNA pairs, engineered to charge an unnatural amino acid into the tRNA[28]. Aminoacyl-tRNA synthetases recognise specific sequences and structural motifs within their cognate tRNA molecules, known as identity elements[29], which allow them to discriminate and aminoacylate only the correct tRNA among a pool of structurally similar tRNA molecules. Failure to do so would result in mis-synthesis of the proteome. Consequently, any novel aminoacyl-tRNA synthetase/tRNA pair must be orthogonal with respect to the host's cellular machinery; the aminoacyl-tRNA synthetase must not charge the endogenous tRNAs, and the tRNA should not be aminoacylated by endogenous aminoacyl-tRNA synthetases[30]. In addition, an available codon is required to encode the novel amino acid[26,28].

Generally, once a pair is known to be orthogonal in a given organism, the aminoacyl-tRNA synthetase is engineered by directed evolution. Libraries of mutations in and around the active site are created, aimed at generating structural variants that can accommodate the unnatural amino acid of interest. Subsequent rounds of positive and negative selection identify aminoacyl-tRNA synthetase variants that can acylate their cognate tRNAs only with the unnatural amino acid, while not cross-reacting with the host's endogenous tRNAs[27,28,30–34]. In order to identify starting aminoacyl-tRNA synthetase/tRNA pairs for further engineering, scientists have searched for natural instances of orthogonality, including natural examples of genetic code expansion.

Selenocysteine, a rare selenium derivative of cysteine considered to be the 21st amino acid, is co-translationally incorporated in response to the UGA (opal) stop codon[35,36]. In *E. coli*, a unique selenocysteine tRNA (tRNA^{Sec}) is first charged with serine (seryl-tRNA^{Sec}), and subsequently converted to selenocysetyl-tRNA^{Sec} in a two-step process[37]. selenocysetyl-tRNA^{Sec} is then delivered to the ribosome by a specialized elongation factor analogous to EF-Tu[38], but is only incorporated at particular mRNA contexts, immediately preceding a hairpin-like motif known as the selenocysteine incorporation sequence[39]. This process is observed, with variations, across the three domains of life[40], and serves to encode selenocysteine into a number of selenium-dependent enzymes[41]. While the incorporation of selenocysteine in response to codons other than UGA has been explored[42], its intricate translational pathway has precluded its repurposing as a general tool for genetic code expansion. This is in contrast to pyrrolysine, the 22nd amino acid, which has proved to be a useful platform for unnatural amino acid incorporation.

Pyrrolysine is a lysine derivative, encoded into methylamine methyltransferases in certain archaeas[43]. In contrast to selenocysteine, its translational pathway is analogous to that of other natural amino acids; it is co-translationally incorporated into polypeptide chains in response to the TAG (amber) stop codon, through a dedicated aminoacyl-tRNA synthetase/tRNA (PylRS/tRNA^{Pyl}) pair[44–47]. *pylT*, the gene encoding tRNA^{Pyl}, effectively acts as an amber suppressor. There is no validated consensus sequence for the incorporation of pyrrolysine. Instead, the genome of these archaea is thought to be configured to tolerate some level of amber suppression and extension of polypeptides beyond the amber stop codon[43].

Several tRNA^{Pyl}/PylRS pairs from methanogenic archaea (notably from *Methanosarcina barkeri* and *Methanosarcina mazei*) have been imported into heterologous hosts, and shown to be orthogonal with respect to *E. coli* and mammalian[48] tRNAs and aminoacyl-tRNA synthetases[49,50]. By evolving derivatives of PylRS, its substrate specificity has been shifted to permit the incorporation of a wide range of pyrrolysine analogues and unnatural amino acids[48,51–55]. The pyrrolysine system uses TAG as a blank codon; this is usually the least abundant codon in genomes, which allows its suppression to be tolerated with minor deleterious effects derived from the extended translation of polypeptides that end in TAG. Moreover, as the amber suppressor tRNA is in competition for TAG decoding with release factor 1 (RF-1), the result of translation is often either a full-length or a truncated product, which facilitates functional selection for incorporation into TAG.

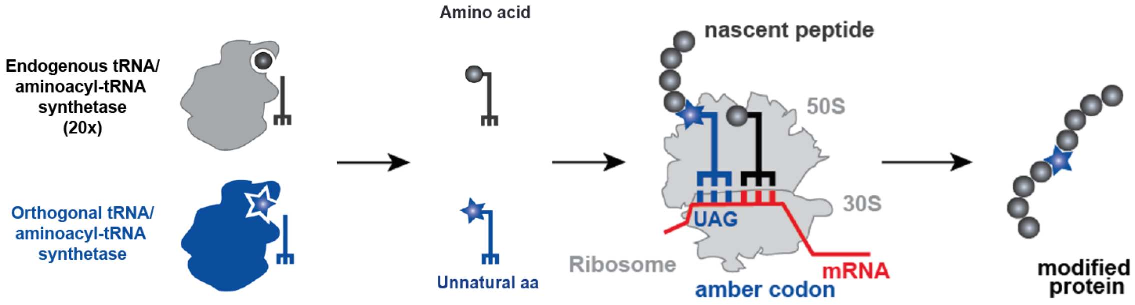


Figure 1.2 – Schematic of genetic code expansion using an orthogonal tRNA/aminoacyl-tRNA synthetase pair. Schematic provided by Kaihang Wang.

Other natural pairs that incorporate canonical amino acids have been shown to be orthogonal when expressed in heterologous hosts. The tRNAs of these pairs do not naturally recognize TAG; in order to act as amber suppressors, their anticodon needs to be mutagenized to CUA, and this must not destroy their affinity for their cognate synthetase. The archaeal tyrosyl-tRNA synthetase/tyrosyl-tRNA pair from

Methanocalcococcus jannaschii has been shown to be orthogonal in *E. coli*[56], and extensively derivatized to incorporate unnatural amino acids in response to the amber stop codon[33,34]. Some *E. coli* pairs, such as those for tyrosine and leucine, are orthogonal in *S. cerevisiae* and have been used to encode non-canonical amino acids in this organism[57,58]. Conversely, several pairs derived from yeast have been used to expand the genetic code of *E. coli*[59–61]. In some cases, engineering of the tRNA was required to achieve full orthogonality to the host machinery[62] and increase the efficiency of non-canonical amino acid incorporation[63].

Genetically encoding unnatural amino acids allows for installing synthetic, useful chemical moieties at virtually any position within a polypeptide, which constitutes a powerful approach for probing and controlling protein function. Several biologically relevant post-translational modifications have been genetically encoded by evolving aminoacyl-tRNA synthetase/tRNA pairs that have specificity for methylated[64–66], acetylated[51,67] and phosphorylated[68–73] amino acids, as well as their analogues[74]. The introduction of bio-orthogonal chemical handles, which react quickly and specifically with particular functional groups but not those of biomolecules, has enabled the site-specific labelling of proteins with fluorophores and affinity tags, with applications in cellular imaging[75–77] and proteomics[78–80]. These have extended and increased the versatility of previous approaches, which relied on the permissivity of natural synthetases to incorporate amino acids bearing chemoselective handles[81]. The site-specific incorporation of synthetic amino acids containing photo-cleavable protecting groups has enabled precise temporal control of protein function *in vivo*[74,82–84]. Finally, the incorporation of unnatural amino acids in the active site of enzymes may endow them with new useful reactivities that are not currently accessible to natural proteins[85,86].

The creation of orthogonal aminoacyl-tRNA synthetase/tRNA pairs capable of aminoacylating tRNAs with amino acids bearing chemically diverse side chains must be accompanied by progress in the genetic systems that enable their efficient incorporation – these are discussed in the following sections.

1.2.2 Improving amber suppression

Unnatural amino acid incorporation in response to TAG is in competition with termination by RF-1, which intrinsically limits its efficiency. This is particularly limiting when two or more amber-encoded unnatural amino acids are incorporated within the same polypeptide, as the efficiency of full-length protein production drops cumulatively. Strategies for improvement have been primarily based on attenuation of RF-1 mediated termination, RF-1 deletion and removal of a subset or all amber codons from genomes.

One study reported the engineering of an *E. coli* ribosome with decreased affinity for RF-1[87]. This was facilitated by the creation of orthogonal ribosomes, which selectively read orthogonal mRNAs through altered Shine-Dalgarno base-pairing[88]. Because orthogonal ribosomes do not translate endogenous mRNAs, and hence do not synthesize the proteome, they provide a suitable starting point for ribosome evolution without compromising cellular viability. The creation of A-site mutational libraries, followed by selection with a chloramphenicol resistance assay, identified orthogonal ribosome mutants with lower RF-1 release activity, resulting in a relative increase in efficiency of amber suppression[87].

RF-1 has been inactivated in *in vitro* experiments. Translation systems reconstituted from individually purified components offer control over the precise constituents of the translation mixture; this allows for selectively excluding RF-1, which removes competition with amber suppressors at TAG[89]. RF-1 depletion has also been achieved in crude cell-free extracts by employing *E. coli* strains carrying heat-sensitive RF-1 variants and subjecting them to mild heat treatment after lysis[90]. Other methods for partially inactivating RF-1 in cell-free extracts have included treatment with RF-1 specific antibodies[91] or RNA aptamers[92] (with limited success), but these solutions are *a priori* not applicable to *in vivo* systems that rely on TAG termination for proteome synthesis.

In *E. coli*, RF-1 (encoded by *prfA*) was thought to be essential. However, studies have shown that RF-1 can be deleted in *E. coli* following the introduction of compensatory mutations in RF-2[93–95] (encoded by *prfB*), which does not normally terminate translation at TAG. An A246T mutation in RF-2 restores the genotype of ancestral *E.*

coli derivatives and increases its termination activity at TAA up to five-fold[96], presumably helping compensate for ineffective termination at TAA caused by the lack of RF-1. While excess RF-2 is reportedly toxic in *E. coli*[96], A246T is sufficient to support RF-1 deletion in *E. coli* BL21, at the cost of a significant drop in fitness[95]. RF-1 has also been deleted in an *E. coli* MDS42$^{prfB\mathrm{T246A}}$ strain where the frameshift-mediated auto-regulatory mechanism of RF-2 had been abolished, presumably resulting in high expression levels[94]. An A293E mutation with unknown mechanistic consequences was required to rescue growth, but the overall fitness of the resulting strain in comparison to the wild-type MDS42 progenitor is unclear. Some mutations in RF-2 are known to shift its specificity towards TAG[93]; low levels of RF-2 mediated termination at TAG may be partly responsible for the compensatory effect of A293E, but this strain still displayed markedly improved non-canonical amino acid incorporation at multiple amber codons within a single mRNA message.

Several efforts have aimed at removing a subset[97,98] (up to 95) or all[99,100] (321) annotated amber codons from the *E. coli* genome through site-directed mutagenesis with ssDNA oligonucleotides. The resulting strains allowed deletion of RF-1. 321 amber codons were removed in the genomes of separate strains[99], which were then combined into a single genome by hierarchical directed conjugation[100]. The strategy for removing 321 amber codons employed a DNA-repair deficient strain (Δ*mutS*[101,102]) which led to the accumulation of a comparable number of off-target mutations (355). While RF-1 knockout itself was not deleterious, these mutations increased the doubling time with respect to the parental *E. coli* MG1655 by approximately 60%. Despite this fact, the strain performed well in incorporation of multiple unnatural amino acids in response to amber codons, and has been shown to incorporate up to 30 *p*-azido-L-phenylalanine residues into an elastin-like protein fusion[103]. Recently, the growth of this strain has been improved by adaptive laboratory evolution[104], which may increase its usefulness for genetic code expansion. Other studies have shown that removing amber in only 7 essential genes is sufficient to permit RF-1 deletion[97]. Removal of 95 instances of amber in a non-repair deficient strain resulted the accumulation of only 9 off-target mutations[98]. This allowed for removing RF-1 with no fitness impairment.

In eukaryotes, eRF-1 terminates translation at all three stop codons. One study engineered eRF-1 to decrease its affinity for the amber stop codon in HEK293T cells. Together with optimization of tRNA expression levels, this resulted in a several-fold increase in yield when expressing proteins containing in-frame amber codons[105]. A recent study has attempted to achieve selective stop codon suppression in only a subset of cellular mRNAs, by co-localizing target transcripts and orthogonal aminoacyl-tRNA synthetase/tRNA pairs in defined, phase-separated regions within the cell[106]. This strategy fuses translational components to protein domains with a natural tendency to aggregate (FUS, EWSR1 and SPD5), as well as truncated kinesins that direct the aggregates to the cell poles. In the design, target mRNA molecules are recruited to these proteinaceous aggregates via an MS2 tag in their 3' UTR, which binds a major capsid protein (MCP)-EWSR1 fusion. This allows for sequestering the transcript of interest, bearing a stop codon, away from the cytoplasm, and promotes local stop codon suppression in the aggregates. However, the maximum enrichment achieved with this system was approximately 10-fold higher suppression of MS2-tagged mRNAs with respect to untagged control transcripts, and the overall translation efficiency dropped significantly when compared to 'normal' cytoplasmic stop codon suppression. Moreover, the presence of an MS2 loop in the 3'UTR of the transcript of interest may selectively increase its cytoplasmic stability with respect to the untagged control, and this may partly account for the increased rates of suppression observed. Nevertheless, the spatial confinement of translational components is a conceptually elegant approach for achieving orthogonal translation.

1.2.3 Genetic code expansion beyond the amber stop codon

Encoding two or more distinct non-canonical amino acids within a single polypeptide requires i) multiple aminoacyl-tRNA synthetase/tRNA pairs, which are orthogonal with respect to the endogenous cellular machinery and to each other, and ii) multiple blank spaces in the genetic code to accommodate the additional building blocks, so that each space encodes amino acid insertion by only one pair.

Some imported aminoacyl-tRNA synthetase/tRNA pairs are naturally mutually orthogonal; this is the case for the *M. jannaschii* TyrRS/tRNA^{Tyr} and the *M. barkeri* PylRS/tRNA^{Pyl} pairs, which have been combined successfully in cells to encode pairs of non-canonical amino acids[107–111]. Directed evolution as also been used successfully to engineer orthogonality, generating mutually orthogonal pairs starting from a single progenitor[112]. Recent and ongoing efforts aim to discover and evolve more mutually orthogonal aminoacyl-tRNA synthetase/tRNA pairs, with a view to increasing our capacity to independently incorporate multiple non-canonical amino acids[113].

The problem of available space in the genetic code is exemplified well by a recent study, which aimed at incorporating 3 distinct non-canonical amino acids into a single protein, using suppression at all three stop codons in *E. coli*[114]. The study employed derivatives of the *M. jannaschii* TyrRS/tRNA$^{Tyr}{}_{CUA}$ and PylRS/tRNA$^{Pyl}{}_{UUA}$ pairs to incorporate O-methyl-L-tyrosine into TAG and cyclopropene-lysine into TAA, respectively. The third pair, derived from *E. coli*, had been generated previously; the yeast orthogonal tryptophanyl-tRNA synthetase/tryptophanyl-tRNA pair was used to functionally replace the *E. coli* copy in the genome, and this allowed subsequent re-introduction of mutants of the *E. coli* pair and their engineering towards incorporation of several tyrosine analogues[115]. A TGA-decoding variant of this *E. coli* tryptophanyl-tRNA synthetase/tryptophanyl-tRNA pair was used to incorporate 5-hydroxytryptophan. Co-expression of all three pairs upon supplementation with their cognate amino acids achieved triplet incorporation. However, competition between release factors and orthogonal suppressors at stop codons resulted in relatively low expression yields. Moreover, the cell's ability to terminate translation was impaired, which forced researchers to use a sfGFP reporter followed by a TEV protease recognition site and multiple stop codons; only by employing TEV protease could they precisely control the terminal sequence of their protein of interest.

In order to achieve clean codon reassignments and higher efficiencies of incorporation, additional truly blank spaces in the genetic code need to be created. Strategies to do so are discussed in the next section.

1.3 – Increasing the information encoded in DNA

The encoded ribosomal synthesis of biopolymers bearing unnatural monomers requires the specific assignment of codons, beyond the amber stop codon, to unique monomers. This is challenging because all DNA triplets in the genetic code already encode natural amino acids or termination signals. Strategies for creating blank codons for their assignment into unnatural monomers have included i) the development of codons bearing more than 3 bases, ii) the expansion of the cell's DNA alphabet, and iii) synonymous codon compression.

1.3.1 Quadruplet codons

The triplet nature of the genetic code dictates that there are 64 codons available for encoding translation. Theoretically, a quadruplet genetic code would provide up to 256 (4^4) spaces in the code, and one can imagine that the ability to reassign these to unnatural building blocks could greatly increase the information coding capacity of DNA. Whether a quadruplet code would achieve effective discrimination between codons that only differ at the 4th base is unknown[116–119], and the route to its achievement is unclear. Even if it were technically possible to generate a 256 codon code, the currently available set of mutually orthogonal aminoacyl-tRNA synthetase/tRNA pairs is limited, and would prevent its efficient utilization. Nevertheless, the last two decades have seen substantial advances in the incorporation of non-canonical amino acids in response to quadruplet codons, and these have been shown to allow for differential decoding with respect to triplet codons.

Several natural and engineered quadruplet suppressors exist, which can incorporate amino acids in response to four-base codons[116–121]. Suppression of amber-derived quadruplet codons is enhanced in strains mutagenized to lack the amber codon and RF-1[100], when using engineered quadruplet decoders[122]. The natural ribosome, however, is rather inefficient at using quadruplet codons. Work on this area has benefited from the creation of orthogonal ribosomes[87,88], as discussed in section 1.2.2. Evolved versions of these ribosomes can read quadruplet codons more efficiently than natural ribosomes by virtue of mutations near the decoding centre in the 16S RNA, while still permitting triplet decoding at high fidelity[108]. The efficiency of quadruplet incorporation by such orthogonal ribosomes has been further increased by using evolved of pyrrolysyl-tRNAs, and their corresponding synthetases[111]; the combination of two mutually orthogonal quadruplet-decoding tRNA/synthetase pairs enabled the site-specific incorporation of two fluorophores for FRET studies. These improved quadruplet-decoding ribosomes still maintain the ability to decode both triplets and quadruplets, although the yields of protein synthesis using the orthogonal system are generally lower than wild-type.

Further evolution and optimization of orthogonal ribosomes towards exclusively decoding quadruplet codons may culminate in the establishment of two parallel genetic codes within the cell, where one operates in triplet and the other in quadruplet.

1.3.2 Expanding the cell's DNA alphabet

A complementary approach for increasing the coding capacity of the cell consists of establishing a third pair of bases into DNA, in addition to A-T and C-G, effectively expanding the genetic alphabet. Several groups have identified and developed pairs of synthetic nucleotides that base-pair through hydrogen-bonding geometries distinct from Watson-Crick pairing[123,124], or by hydrophobic packing and shape complementarity[125].

The stable replication of an unnatural base pair (UBP) within the cell, together with the development of corresponding translational machinery that reads it uniquely, may allow for the specific assignment of the unnatural codon to an unnatural amino acid. It has been known for over 20 years that several of these artificial pairs can pack into a DNA helix, and are substrates for a number of polymerases in DNA replication and transcription *in vitro* with increasingly refined specificity and retention rates[126–129].

Almost 30 years ago, the *in vitro* ribosomal incorporation of an unnatural amino acid into a short peptide, via a codon-anticodon pairing which involved an unnatural nucleobase, was achieved[130]. The system employed a eukaryotic lysate, a chemically synthesised mRNA and tRNA, and a chemically aminoacylated suppressor tRNA bearing isoG in its anticodon. The performance of *in vitro* unnatural base pair (UBP) systems has continued to be refined; the latest example is Hachimoji DNA, a genetic system formed of 8 different bases amenable for *in vitro* transcription by T7 RNA polymerase into corresponding RNA polymers, which yield functional aptamers[131]. However, progress towards the implementation of unnatural base pairs *in vivo* has only come in the last few years, and their success has been limited to the family of base pairs based on hydrophobic stacking which were shown to be replicated and transcribed well *in vitro*[126,132].

Because natural bacteria lack the pathways for the biosynthesis of the unnatural nucleotides, work has focused on making these available inside the cell. The most successful strategy has overexpressed a nucleotide triphosphate transporter from the algae *Phaeodactylum tricornutum* in *E. coli*, to mediate entry of unnatural hydrophobic nucleotide triphosphates d5SICS and dNaM, and their derivatives, into the cytosol[133]. The unnatural triphosphates were sufficiently stable to be substrates for replication inside the cell. A ColE1 plasmid containing the UBP was constructed *in vitro*, and transformed into *E. coli* cells expressing the algal transporter in medium supplemented with the unnatural nucleotide triphosphates. The authors harvested the plasmid bearing the UBP after 24 doublings, and estimated a retention rate of 99.4% per doubling, with a tendency to be replaced by the tA/tT pair. This rate of error was significantly higher than that of natural DNA replication, and the strain bearing the transporter and UBP plasmid grew poorly.

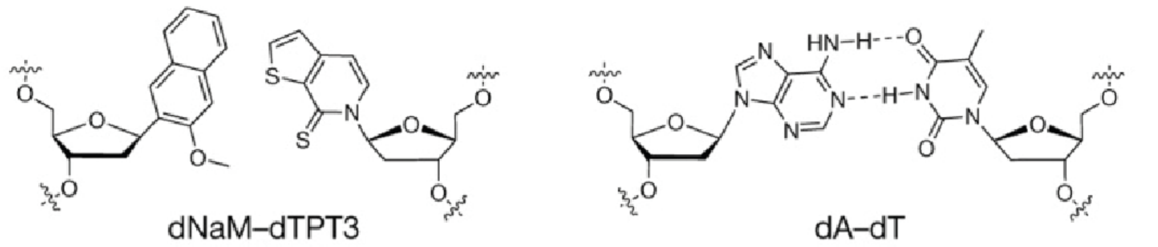


Figure 1.3 – Example of an unnatural base pair. The base pair on the left (dNaM-dTPT3) was used to encode non-canonical amino acids *in vivo,* as described in this section. On the right is a natural dA-dT pair. Adapted from ref[134].

Both the retention rate and the fitness of the host strain were improved in a more recent report[135]. Codon optimization, removal of an unnecessary N-terminal peptide and chromosomal integration of the nucleotide transporter alleviated the toxic effects derived from its overexpression in *E. coli*. The resulting strain showed comparable retention rates of the d5SICS-dNaM pair at various copy numbers and locations within the UBP plasmid, as long as the sequence context remained unchanged. However, altering the identity of the bases immediately upstream and downstream of the UBP yielded highly variable retention rates. Retention rates were generally improved when using dTPT3 (a sulphur derivative of d5SICS) with dNaM, but some sequence combinations still abolished UBP maintenance. While not intrinsically improving the retention capacity of the system, the Cas9-mediated cleavage of the most frequent revertants allowed for selective depletion of mutant plasmids, and helped maintain a homogeneous population of plasmids bearing the UBP.

This optimized strain has been used for incorporating a non-canonical amino acid in response to an unnatural base pair[134]. A variant of sfGFP containing dNaM in the 2nd base of the codon for position 151 could be transcribed *in vivo* by T7 RNA polymerase. The message was effectively decoded by derivatives of a seryl-tRNA, a *M. mazei* pyrrolysyl-tRNA and an evolved *M. jannaschii* tyrosyl-tRNA where the 2nd postion in the anticodon was dTPT3. These directed the incorporation of serine, propargyl-L-lysine (PrK) and *p*-azido-L-phenylalanine (pAzPhe) respectively, at levels comparable to wild-type sfGFP production in the case of serine and around 65% in the case of PrK and pAzPhe. The fidelity of amino acid incorporation ranged from 93% to 99%, though it is unclear what fraction of mis-incorporation results from reading by alternative tRNAs, mis-charging of the cognate tRNAs with other amino acids, or loss of the UBP in replication or transcription. Nevertheless, it is remarkable that UBPs based on shape complementarity and hydrophobic interactions can mediate efficient decoding in the complex ribosomal environment.

Constructing orthogonal replicational, transcriptional and translational pathways based on unnatural base pairs will require systems that permit robust growth across different culture media, and can maintain high retention rates in a wide variety of sequence contexts. As well as continuing to refine and identify optimal synthetic base pairs, the engineering of DNA polymerases for improved operation with UBPs may further enhance their maintenance, but this will necessarily require a balance between the improvement of polymerases for UBP work and the maintenance of replication fidelity of the natural bases. As with the development and optimization of orthogonal ribosomes for quadruplet decoding, addressing this problem could benefit from a similar approach, with orthogonal polymerases assigned exclusively to particular UBP-containing replicons. Orthogonal replication systems have been reported in yeast[136,137]. However, these are not transferrable to organisms such as *E. coli*, where UBPs have been shown to be functional. The development of polymerase-template pairs that confine the polymerase to only one DNA template within the cell may enable polymerase evolution without compromising cell viability by mis-replication of the genome.

1.4 – Synonymous codon compression

In addition to expanding the cell's DNA alphabet, another strategy for increasing the information encoded in the genetic code is the compression of synonymous codons. 18 out of the 20 canonical amino acids are encoded by up so six synonymous codons. Hence, the genome-wide replacement of a subset of codons by their synonyms, followed by the deletion of their cognate decoding elements, may liberate these codons for reassignment into unnatural monomers. This reassignment model would be conceptually analogous to the 'codon capture' theory of naturally occurring reassignments[21], and may enable the expansion of the cell's repertoire of genetically encoded building blocks and facilitate the synthesis of non-canonical biopolymers.

Site-directed mutagenesis has enabled the creation of *E. coli* strains where up to 321 instances of the amber codon have been removed[99,100,104] (discussed in 1.2.2). Sense codons, however, are 1 to 2 orders of magnitude more frequent in genomes than amber codons, and their removal by site-directed mutagenesis-based strategies becomes impractical. This, together with the decreasing costs of DNA synthesis, has resulted in a number of efforts to achieve synonymous codon compression by whole-genome synthesis. Re-writing entire bacterial genomes with compressed genetic codes presents two fundamental challenges. First, synonymous codon choice in natural genomes is not random and, while we have some understanding of which biological factors dictate such choices, finding optimal synonymous replacements for a given codon remains challenging. Second, realizing synthetic DNA designs *in vivo* requires the development of new technologies for the assembly, manipulation and implementation of synthetic DNA at the genome-scale, as well as the troubleshooting of deleterious DNA designs. Because testing hypotheses on optimal genome-wide codon choice requires the ability to manipulate large sections of natural genomes, these two problems are somewhat entangled.

1.4.1 Consequences of synonymous codon choice

It has long been known that synonymous codon choice can influence gene expression and cell fitness on a number of levels[138]. Synonymous codons have been implicated in the formation of mRNA secondary structures that modulate translation initiation[139–141], mRNA decay[142–144], translational speed[145] and co-translational folding[146,147].

The literature is populated with several alternative (and sometimes conflicting) explanations for the observed biases in codon usage. For example, some of the work on synonymous codon choice correlates 'optimal' codons with increased protein synthesis; this is associated to increased local translation speeds and thought to reflect relative tRNA abundance[148–152]. In line with this, the N-terminal region of most open reading frames is enriched for codons considered to be suboptimal[153]. A model has been proposed in which these slow down translation towards the beginning of the open reading frame, in order to avoid ribosomal traffic jams further down the transcript[154]. This is in contraposition to other studies which suggest that elongation rates do not vary substantially, and translation initiation is the rate-limiting factor[139–141].

Due to the high number of variables involved, a comprehensive understanding of the pressures driving codon choice is lacking, and the development of models that integrate the different factors to explain variations in gene expression remains a key question. In this light, studies that systematically explore multiple variables, aiming to experimentally dissect the relative contributions of each factor to translational efficiency, are of particular interest.

One study focused on codon choice determinants towards the 5' of the open reading frame[155]. The authors created a library of GFP variants consisting of GFP fused at its N-terminus to the first 11 codons of 137 different endogenous *E. coli* genes. For each 11aa-GFP fusion, 13 synonymous variants across the 11aa peptide were created, and each of these variants was further combined with promoters and RBSs of varying strengths, resulting in the synthesis of over 14,000 constructs. This was followed by measurements of DNA, RNA and protein levels in all regimes. The results showed that the presence of rare codons at the N-terminus is correlated with increased protein expression, and the analysis suggested that this is primarily due to the reduced formation of RNA structures by rare codons, which are more A/T rich.

More recently, a larger study examined 8 factors thought to influence protein expression levels, including mRNA structure and codon adaptation in different parts of the transcript, A/T content and amino acid identity[156]. The authors conceived what they term a 'full factorial design' – they split each relevant factor into 2 or 3 parameter ranges, and then implemented them in a synthetic construct fused to GFP in all possible combinations. For each combination, the authors employed a property-scoring algorithm to independently design multiple DNA sequences in the same parameter range, and then generated two mutational replicates of each variant, while aiming to maintain the same properties. The result was a library of over 244,000 different sequences containing all possible 1,458 parameter combinations in 168 mutational replicates. The synthetic reporter construct was embedded into a bicistron; the reporter's Shine-Dalgarno overlapped with a leader peptide which contained an early amber codon. Providing an amber suppressor results in translation of the leader peptide with, the authors hypothesised, concomitant ribosome-mediated unwinding of RNA structures around the reporter's Shine-Dalgarno. Because this is thought to facilitate initiation, this allowed for testing translational efficiency in two different initiation regimes. *E. coli* cells containing the library of GFP variants were sorted into bins of comparable fluorescence, followed by DNA sequencing and back-tracing of the constructs to the predicted sequence properties for analysis. The results showed that mRNA structures are, by far, the dominant factor governing translational efficiency. Only when mRNA structures around the Shine-Dalgarno were attenuated via addition of an amber suppressor, the effect of other metrics (notably codon adaptation[149]) became significant.

Importantly, analysis of this large dataset showed that only about half of the observed variance could be explained by variation in the computed parameters for the different DNA sequences. This underscores the difficulties in predicting biological features from the DNA sequence alone. Even very closely related sequences (1 to 4 nucleotide differences) with similar computationally-derived parameter ranges showed significant divergence in protein expression. The challenge of predicting biological features is particularly limiting for RNA structures, whose dynamics *in vivo* are known to differ substantially from those *in vitro*[157].

It is thought that synonymous codon variants that allow for efficient translation initiation, reducing the cost of protein synthesis, are selected for[142,158]. However, codon choice can affect fitness though mechanisms independent of translation. A recent study analysed a library of synonymous GFP variants and found that a subset were toxic[159]. In one instance, disruption of the reading frame by the introduction of a premature stop codon did not alleviate the toxic effect, indicating it originated at the RNA level. Notably, no significant RNA structure was predicted for this transcript and the underlying reason for toxicity is unclear. In other cases, toxicity may be explained by the accumulation of highly structured transcripts that hamper RNAse-mediated decay[156].

Synonymous codon choice remains a controversial and highly active research field, and the research outcomes will necessarily influence efforts aimed at carrying out genome-wide synonymous codon substitutions. Despite the challenges and unknowns outlined here, several efforts have aimed at creating recoded genomes; these are discussed in the following section.

1.5 – Creating synthetic recoded genomes

As stated above, probing synonymous codon choice genome-wide is intrinsically linked to the development of technologies to perform large-scale changes in the DNA of living cells. Our ability to manipulate genomes has been greatly facilitated by progress in the synthesis and assembly of DNA sequences.

1.5.1 Genome synthesis in context

Advances in the chemical synthesis of RNA/DNA nucleosides and nucleotides in the early 1950's rapidly led to the chemical synthesis of dithymidine monophosphate[160]. 15 years later, Agarwal *et al.* reported the chemical synthesis of the first gene, a 76 nucleotide alanine-tRNA from *E. coli*[161]. The dsDNA molecule was constructed from short segments of 6-12 polynucleotides, which contained 4-5 nucleotides of complementarity to adjacent segments. Individual polynucleotide segments were 5′

phosphorylated with T4 polynucleotide kinase, followed by annealing of the phosphorylated segments and ligation with T4 ligase. Albeit at a very small scale, this work laid the conceptual foundation for most of the subsequent work on DNA assembly from synthetic oligomers. A similar approach was used in 1977 to synthesise the gene for human somatostatin, 14 amino acids long[162], with restriction overhangs which facilitated its cloning and propagation in *E. coli*.

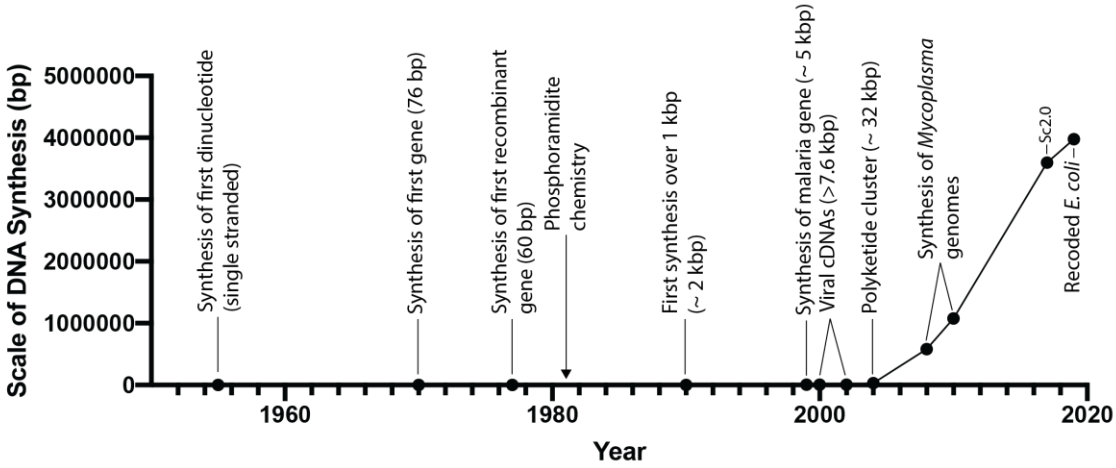


Figure 1.4 – Timeline of DNA synthesis

In 1981 Caruthers and co-workers reported the use of phosphoramidite chemistry to synthesise polynucleotides on a solid support[163]. The speed of the reaction and ease of storage of the starting materials constituted a major advance over previous chemistries. This technology still forms the basis of commercial DNA synthesis pipelines to this day, and has been a major contributor to the rapid increases in synthesis scale during subsequent decades. In the 1990's and early 2000's, commercial DNA synthesis had entered the kilobase range[164,165] and several efforts completed the synthesis of viral cDNAs over 7 kbp[166,167]. The cDNA of the poliovirus genome (7.6 kbp) could be transcribed *in vitro*, and was encapsulated into infectious poliovirus particles upon incubation of the synthetic transcript with a HeLa cell-free extract[167].

In 2008 Craig Venter, Dan Gibson and co-workers reported the design, chemical synthesis and assembly of a 582 kbp minimal *Mycoplasma genitalium* genome[168]. The authors obtained 5-7 kbp synthetic DNA fragments from vendors, and assembled them into progressively larger segments in the form of BACs by enzymatic *in vitro* assembly[169] (for smaller fragments), and homologous recombination in *S. cerevisiae*[170] (segments over 200 kbp). While the synthetic genome did not support a viable cell, it was stably propagated in yeast. Only 2 years later, the same group reported the synthesis of a viable *Mycoplasma* genome of almost twice the size (1.08 Mb)[171], assembled similarly to its predecessor. Simultaneous advances in chromosomal transplant in *Mycoplasma* species[172,173] paved the way for the implementation of the synthetic chromosome into a functional cell, via the extraction of intact synthetic chromosomes from yeast in agarose plugs followed by their PEG-mediated transformation into host wild-type *Mycoplasma* cells.

In 2011, the international Sc2.0 consortium announced its ambition to create a synthetic *S. cerevisiae* cell[174]. The features of the synthetic genome would include:

i) the deletion of repetitive sequences, transposable elements and many introns in order to increase genome stability and reduce genome size

ii) the introduction of hundreds of symmetrical loxP sites for inducing genome rearrangements and probing genome stability, plasticity and structure

iii) the recoding of all instances of the TAG stop codon into TAA

iv) the transfer of all tRNAs into a separate, synthetic ‘neochromosome’

In 2017, 6 of the 16 chromosomes had been completed in separate strains[175–180] and work is in progress for the rest. Multiple synthetic chromosomes have been combined into a single yeast strain by mating and sporulation[175]; the ultimate goal of this project is a yeast cell containing only synthetic chromosomes.

The *Mycoplasma* and yeast cases highlight the fact that DNA synthesis can now be performed at scales compatible with genome synthesis. This provides a novel way to address biological questions, and a promising route to the profound engineering of organisms with desired properties. Importantly, the ability to rewrite genomes may provide control over global codon usage, and opens the door to the creation of alternative genetic codes. The next section describes synthetic approaches for the exploration of recoded genomes.

1.5.2 Synthetic recoded genomes

Aside from the biological uncertainties that surround synonymous codon choice (section 1.4.1), from a technical perspective the creation of synthetic recoded genomes requires i) the ability to synthesise DNA at scale and ii) technologies to assemble large synthetic DNA constructs and implement them *in vivo*. As described in the previous section, improvements in the throughput and scale of DNA synthesis, largely brought about by automation, have essentially solved the first problem. However, the assembly of genomes still presents challenges.

Implementing large-scale synthetic DNA designs

Initial genome synthesis efforts in *Mycoplasma* relied on assembling the entire genome from shorter precursors in yeast (*ex vivo*) followed by whole genome transplantation into a target cell[171–173]. However, whole genome transplantation cannot be easily adapted for other organisms which do not share the unique features of *Mycoplasma* (e.g. lack of cell wall, high transformation efficiency). Because most bacterial genomes are several megabases in size, methods that can replace sections of genomes in a stepwise manner are desirable. Pioneering work achieved the cloning of the 3.55 Mb *Synechocystis* PCC6803 genome inside the *B. subtilis* genome, through the stepwise integration of 30 kbp pieces of DNA flanked by two alternating positive selection markers ('inchworming')[181]. Again, this study exploited unique properties of *B. subtilis* (high natural competence, high homologous recombination activity), which limited its transferability. In addition, recoded genomes will introduce an elevated number and high density of changes with respect to the wild-type genome. Methods that provide direct feedback on synthetic sequence design would facilitate the identification of design flaws and help in understanding the consequences of synonymous mutations.

Transformation-associated homologous recombination in *S. cerevisiae* has been widely adopted as a reliable method to assemble large synthetic DNA pieces[182]. From this common intermediate, there are a number of strategies for implementing these constructs into target bacterial cells. In *E. coli*, lambda-red mediated homologous recombination is a well-established method for performing small scale (<10 kbp) manipulations to the genome[183,184], but the efficiency of replacement decreases rapidly with increasing DNA length because of the low efficiency of electroporation of linear DNA. In the recently reported REXER method[185], the synthetic DNA is delivered in a circular form as a bacterial artificial chromosome (BAC), and subsequent *in vivo* Cas9 cleavage generates linear DNA substrates for lambda recombination. This strategy decouples transformation of the DNA from the recombination event, and allows the replacement of over 110 kbp of the genome in a single step.

Another study reported the assembly in yeast of recoded *Salmonella typhimurium* DNA plasmids of about 20 kbp, followed by rolling circle amplification and splitting of the resulting concatemers by restriction enzyme digestion (termed SIRCAs)[186]. The vast amounts of DNA generated by this method helped compensate for the low efficiency of electroporation of the 20 kbp DNA pieces in the *S. typhimurium* host. Similarly to inchworming, both REXER and SIRCAs can be iterated by using two alternating selectable markers for integration of adjacent synthetic genome segments; the use of double positive-negative selection cassettes enables counter-selection against the previous maker, increasing the efficiency of the process. After each step, analysis of internal recombination-mediated crossovers between the wild-type and synthetic DNA can provide information on disallowed synthetic DNA designs. The different groups that form the Sc2.0 consortium carry out replacement of the natural chromosome by its synthetic counterpart in an analogous fashion, through variations of a general strategy they term SwAP-In[175]. Like REXER and SIRCAs, SwAP-In uses two alternating selectable markers to select for integration of adjacent and overlapping synthetic DNA – in yeast, this is facilitated by the high levels of endogenous homologous recombination.

Genome recoding

Several efforts have attempted to construct recoded prokaryotic genomes, beyond removal of the amber stop codon[99,100] described in Section 1.2.2. As exposed in Section 1.4, synonymous codon substitutions can have significant effects on biological function, and when performed on a genomic scale these can result lethal. Genome recoding efforts have adopted different strategies to define the rules for synonymous codon substitution.

In 2016, a minimal *Mycoplasma* genome (531 kbp, 473 genes) was synthesized[187]. This genome was obtained through several iterations of the synthesis and transplant workflow that had previously led to the creation of the 1.08 Mb synthetic *Mycoplasma*[171] (Section 1.5.1). In addition to reorganizing a significant portion of the genome based on gene ontology annotations, the study recoded a relatively small (5 kbp) section of the genome encompassing 3 essential genes, aiming to modify codon usage based on codon adaptation metrics. In this case, recoding did not seem to affect cell viability.

One study described the design and testing of an *E. coli* genome with 57 codons[188]. This work employed an algorithm that screens for target codons in protein-coding genes, and tries to identify the best synonymous replacements at each instance as those that minimally disrupt predicted local features (e.g. predicted secondary structure, ribosomal binding sites and codon adaptation). The authors divided their genome design into 87 ~50 kbp segments. 55 of these segments were assembled into low-copy plasmids and transformed into *E. coli*; 44 segments could support deletion of the corresponding 50 kbp wild-type genomic segment, while the rest were lethal. 10 of the successful segments could be integrated into their corresponding genomic loci at single copy in separate strains, albeit with a number of codon reversions and deletions. One other segment failed.

Other efforts have adopted more empirical approaches for identifying viable synonymous codon substitutions. In a previous report, our group identified a ~17 kbp region of the *E. coli* genome rich in essential genes on which to test a variety of serine, leucine and alanine recoding schemes, where one codon is systematically replaced by one synonym[185]. Genome replacements were conducted by REXER, as described above. Some recoding schemes emerged as promising, whereas others were non-tolerated. In one instance, a deleterious recoding scheme was rescued across the entire 17 kbp by a single mutation in a non-tolerated codon. Recoding efforts in *Salmonella* have used SIRCAS to replace ~200 kbp of the genome with a synthetic version, characterized by a defined recoding scheme where two leucine codons are systematically replaced by two synonyms[186].

Another recent report describes the design and assembly *ex vivo* of a minimal >700 kbp *Caulobacter ethensis* genome[189]. The design removes two leucine codons, but further scrambles >50% of all codons to their synonyms, in an attempt to strip away any high-order non-coding biological features from protein-coding genes and achieve a simpler, more modular genome. Conjugation of plasmids containing >20 kbp synthetic fragments into a wild-type *Caulobacter* host followed by transposon mutagenesis suggested that 80% of the subset of synthetic genes tested could functionally replace the truncated wild-type counterparts. However, the multi-copy format of the synthetic constructs and the lack of fitness data difficult the interpretation of the results. In general, strategies to assess recoding based on the functionality of individual genes may hamper the assembly of viable recoded genomes due to co-lethality and cumulative fitness defects.

These studies exemplify the wide-spread interest for synthetic recoded genomes. The ability to define genome sequences in this manner may not only prove useful for expanding the genetic code with additional monomers, but also allow us to modulate the evolvability of organisms by constraining the mutational space accessible by nucleotide substitutions in an organism's DNA, which may be relevant in biocontainment[190]. Additionally, re-wired genetic codes would constitute a language barrier between natural and engineered organisms. This may be important in biocontainment by preventing horizontal gene transfer[191,192], and may confer resistance to viral infections, with obvious positive implications for biotechnological industrial processes[193]. These possibilities have been supported by proof-of-principle experiments in the amber-free *E. coli* strain[100].

1.5.3 Assembling synthetic genomes

Yeast is often the platform of choice for assembling large synthetic DNA constructs, as exemplified by the synthesis projects described above. While genome-sized DNA constructs can be transferred directly from yeast to *Mycoplasma*, other organisms are not amenable to this technique and require the gradual implementation of sub-sections of the genome inside the host; these constitute assembly intermediates. For logistical reasons, it is more efficient to generate multiple of these in parallel, but somehow the synthetic sections from each intermediate need to be combined into a single synthetic genome.

This issue may be resolved by leveraging unique features of particular organisms. In yeast, for instance, multiple synthetic chromosomes originating from distinct strains have been combined into a single cell by mating and sporulation[175]. Here, investigators increase synthetic transfer efficiency during mating by selectively inactivating the centromeres of native chromosomes, promoting consolidation of the synthetic counterpart. However, the biological processes that underlie this methodology do not naturally occur in prokaryotes, which have accumulated most of the attention in the genome synthesis field.

Directed conjugation has long been used to transfer genetic material between bacterial strains. The construction of the amber-free *E. coli* strain relied on the use of successive rounds of directed conjugation to merge the genomes of different strains, each with a subset of the amber codons removed[99,100]. For this, *oriT* sequences were integrated in the genome immediately upstream of synthetic sections, to mark the beginning and directionality of genome transfer. Appropriately placed positive and negative selection markers in donor and recipient strains allowed for selecting conjugants where the synthetic region of interest had been transferred.

While conjugation allows for efficient directional transfer, it requires extensive homologies between donor and recipient genomes, and it does not allow for much flexibility regarding the site of implementation of synthetic DNA in the recipient genome. In general, the prokaryotic genome assembly field may benefit from technologies that allow for the transfer of defined DNA sequences between strains, and endow scientists with control over the site of integration, as well as the orientation. Such methods may facilitate the transition from assembly intermediates to fully synthetic genomes.

1.6 – Conclusion and thesis goals

Organisms with recoded genomes have vast potential for genetic code expansion, with implications in fundamental biological research, therapeutic strategies, biocontainment and the biotechnological industry. Moreover, recoding genomes may provide insights into the malleability of the genetic code, and increase our understanding of the underlying biological processes through a bottom-up approach.

Despite recent progress in understanding synonymous codon usage and our increased ability to synthesize and assemble genomes, there are no precedents for genome-wide sense codon compression. A number of technologies have been developed for replacing large sections of natural genomes with synthetic counterparts. Among these, REXER and its iteration (GENESIS), developed in our group, have proved to be particularly powerful, achieving replacement of over 100 kbp of DNA in a single step. This thesis aimed to implement and expand these technologies for genome engineering to create a synthetic *E. coli* genome with reprogrammed decoding rules.

The first part of this thesis (**Chapter 2**) describes tests towards validating the suitability of REXER and GENESIS for carrying out whole genome replacement. The goal of this Chapter was to demonstrate the robustness of the techniques, lay out a strategy for genome replacement, identify potential bottlenecks and create strategies to accelerate the process.

In the second part of this thesis (**Chapter 3**), the replacement strategy outlined in Chapter 2 is applied to the synthesis and assembly of a recoded *E. coli* genome. The resulting strain constitutes a chassis for probing sense codon reassignment into unnatural monomers, using genetic code expansion technologies described in Section 1.2.

The third part of this thesis (**Chapter 4**) describes novel genome engineering technologies for carrying out programmed structural rearrangements to the genome of *E. coli* by splitting it into diverse pairs of chromosomes, which are substrates for further engineering. The suite of technologies presented here may be useful in future synthetic genome assemblies.

Chapter 2 – Inter-strain conversion in *E. coli* by stepwise genome replacement

Note: I performed all experiments described in this Chapter, except when explicitly indicated at the beginning of a section.

2.1 – Towards a technical demonstration of stepwise genome replacement

As described throughout Chapter 1, DNA synthesis can now be carried out at a scale compatible with whole genome synthesis, but the technological challenge of implementing the synthetic DNA inside a living cell remains. Pioneering genome synthesis efforts in *Mycoplasma* achieved synthetic organisms via transplantation of whole synthetic genomes directly from yeast into the target cells[171–173]. However, this was facilitated by the fact that *Mycoplasma* lacks a cell wall and can uptake megabase-scale pieces of DNA by transformation. Whole genome transplant is not applicable in organisms that do not fit these requirements, or whose complete genomes cannot be cloned in yeast. Another caveat of this approach is that a single deleterious sequence in the synthetic design can result in failure of the transplant, providing little feedback as to what went wrong. In the *Mycoplasma* work, for instance, identifying a synthesis error in essential gene *dnaA* proved quite challenging[171]. In this respect, strategies for replacing the genome in discrete steps, rather than all at once, could allow for narrowing down problematic regions and make troubleshooting more tractable. With this in mind, previous work by our group has focused on developing technologies that allow for the stepwise replacement of natural bacterial genomes with synthetic versions.

REXER can replace >100 kbp of the genome

Lambda-red mediated recombination is a well-established method for performing integrations and replacements in *E. coli* up to ~10 kbp. In the classical lambda-red workflow, the dsDNA of interest is provided to the cells by electroporation in a linear form, as the lambda machinery requires linear dsDNA as a substrate[183]. However, electroporation of linear DNA in *E. coli* is much less efficient than circular DNA, and this effectively limits the replacement step size that can be achieved at once by lambda-red to 20 kbp, if microgram quantities of template DNA are available[186]. In 2016, the Chin group developed REXER ('Replicon EXcision for enhanced genome Engineering through programmed Recombination'), which allows for replacing up to 120 kbp of the wild-type genome with a synthetic version in a single step[185]. It does this by delivering the DNA of interest to *E. coli* in a circular BAC form. Providing these cells with a spacer plasmid, which encodes gRNAs for Cas9 cleavage, leads to the Cas9-mediated generation of a linear synthetic DNA insert *in vivo*. This linearized DNA is then a substrate for lambda-red recombination (**Figure 2.1**). Because REXER decouples the transformation step from the recombination step, the efficiency of recombination is independent from the efficiency of transformation, and this allows for replacing much larger fragments in a single step.

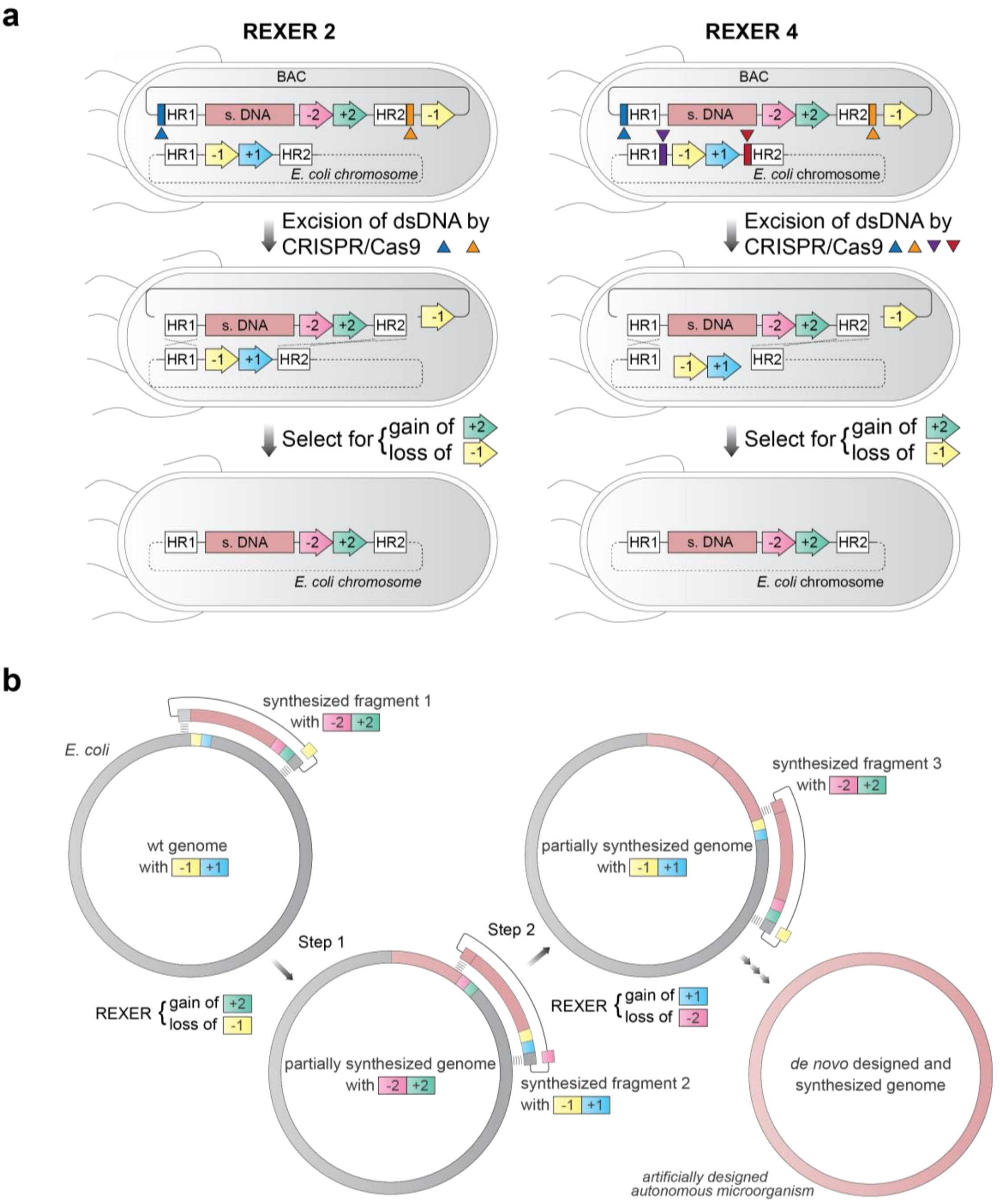

a
REXER 2
REXER 4
BAC
HR1
s. DNA
-2
+2
HR2
-1
+1
E. coli chromosome
Excision of dsDNA by CRISPR/Cas9
Select for {gain of +2 / loss of -1
b
E. coli
synthesized fragment 1 with -2 +2
wt genome with -1 +1
Step 1
REXER {gain of +2 / loss of -1
partially synthesized genome with -2 +2
Step 2
synthesized fragment 2 with -1 +1
REXER {gain of +1 / loss of -2
partially synthesized genome with -1 +1
synthesized fragment 3 with -2 +2
de novo designed and synthesized genome
artificially designed autonomous microorganism

Figure 2.1 - Schematic of REXER and GENESIS. a) In REXER, synthetic DNA is provided to host cells in a BAC, flanked at its 3' end by double selection cassette –2/+2 (pink and green

(continued from previous page) arrows, respectively). The host genome contains a -1/+1 (yellow and blue arrows, respectively) double selection cassette at the beginning of the region to be replaced. Another copy of -1 is present in the BAC vector. Cas9 generates two cuts in the BAC in the case of REXER2 (blue and orange triangles), or two additional cuts in the genome in the case of REXER4 (purple and red triangles). Cas9 cleavage exposes homology arms HR1 and HR2 as linear ends in the BAC (REXER2) or in both the BAC and genome (REXER4), which are substrates for lambda-red recombination. Selection for the gain of +2 and the loss of -1 identifies clones where the genomic region has been replaced by synthetic DNA. **b)** Iterative REXER steps can replace adjacent regions of the genome, by iterating double selection markers -1/+1 and -2/+2. At each step, genetic selection for the loss of the negative marker in the genome and the gain of the positive marker in the BAC identifies successful replacements. This iteration is called GENESIS, and is facilitated by the fact that the genome resulting from one REXER step is a direct substrate for the next step.

REXER uses two distinct positive-negative double selection cassettes. These can be considered -1/+1 and -2/+2, and are coupled to the wild-type genome and synthetic BAC, respectively. After recombination across a region of interest, selection for -1 and +2 identifies clones that have lost the wild-type genomic DNA, and have integrated the synthetic version. In the resulting cell, the positive-negative double selection cassette previously in the BAC (-2/+2) is now integrated in the genome downstream of the replaced region (**Figure 2.1a**).

Two key features make REXER a promising tool for attempting whole genome replacement. Firstly, REXER can be iterated, in a process termed GENESIS ('GENomE Stepwise Interchange Synthesis', **Figure 2.1b**). After one REXER step, the product genome contains a positive-negative selection cassette immediately downstream of the replaced section, and hence serves as a direct template for a subsequent REXER step to replace the adjacent region. Using two alternating positive-negative selection markers facilitates iteration. Secondly, REXER can provide feedback on the viability of the synthetic sequences. The clones resulting from REXER may present chimeric genomes, resulting from crossovers between synthetic and wild-type DNA across a region of interest. Such crossovers may reflect selective pressures for eliminating particular synthetic DNA sequences. The genotypes of a post-REXER population can be analysed by deep sequencing, and compiling the patterns of crossover across the population can help identify deleterious sequences at high resolution[185].

Testing the feasibility of genome replacement

Section 1.4 described the challenges in identifying optimal codon substitutions genome-wide. With views of creating a synthetic recoded *E. coli* genome, previous work by the Chin group has used REXER to try and identify defined synonymous codon compression schemes for serine, leucine and alanine, where one codon is systematically replaced by one synonym. Several synonymous versions of a 17 kbp region of the *E. coli* genome comprising a cell division operon, rich in essential genes, were synthesized and used to replace the corresponding wild-type section. Some of these recoding schemes were well tolerated, whereas others were catastrophic and entirely rejected by the cell. In one instance, a single nucleotide substitution rescued viability of the entire 17 kbp stretch[185].

These results highlighted the feasibility of using defined synonymous codon compression schemes for creating recoded genomes, and constituted a demonstration of our ability to identify synthetic deleterious sequences and fix them. With this, our group set out to devise a strategy for the synthesis and assembly of a recoded *E. coli* genome. When choosing a reference strain for synthesis, *E. coli* MDS42 appeared to be the best candidate, due to its good genomic stability conferred by the deletion of transposable and insertion elements[194]. This strain is a genome-reduced version of *E. coli* W3110, and at 3.97 Mb its genome would be cheaper to synthesize compared to most other *E. coli* strains (usually over 4.5 Mb).

In previous work, ~220 kbp of the wild-type MDS42 genome were replaced with synthetic sequences that were almost identical to the wild-type, except for the presence of several watermarks that served to verify replacement. Whether a complete genome replacement could be carried out with a synthetic recoded genome was unknown, and we anticipated that we would encounter a number of biological complications derived from the introduction of deleterious sequences. Aside from sequence-dependent hurdles, before initiating an expensive genome synthesis we wanted to have a technical demonstration that REXER is functional throughout the genome, and that it can be iterated multiple times in order to mediate full genome replacement. Moreover, while REXER had been successful in replacing up to 120 kbp of DNA in a single step, its size limitations were not thoroughly tested; the ability to replace larger regions per step may accelerate the replacement process.

We reasoned that using REXER to replace the genome of one strain of *E. coli* with the genome of another 'donor' strain would constitute a suitable technical demonstration of stepwise genome replacement, because we knew that both the start and end genome sequences were viable. We envisioned that in this demonstration we could avoid the DNA synthesis costs by cloning large portions of the 'donor' genome into BACs in a configuration suitable for REXER, and subsequently using these for replacing the genome of the 'recipient' strain (**Figure 2.2a**). We chose *E. coli* DGF-327[195] as a donor. This strain has a 3.27 Mb reduced genome, derived from a series of deletions in its parental strain *E. coli* MG1655. Because some regions of its genome are missing with respect to MDS42, these loci provide convenient watermarks for tracking replacement at each step.

A plan for stepwise genome replacement in MDS42

Replacement had been demonstrated at 120 kbp steps, and the *E. coli* MDS42 genome is around 4 Mb. Consequently, we envisioned a replacement strategy based on ~40 iterations of REXER. The MDS42 genome sequence was divided into fragments of 90-135 kbp. Each of them corresponded to a hypothetical, minimal REXER step. If, however, we found that REXER could mediate larger replacements, we retained the flexibility to replace multiple fragments per step. We term the boundaries between these fragments 'landing sites'; they are labelled L00-L36 in **Figure 2.2b**. Throughout the replacement process, landing sites would be the sites of integration of the REXER double selection cassettes. Negative and positive selection must be functional at these loci, and integration of the markers at these sites must preserve cell viability. Landing sites were carefully designed to be in intergenic regions between non-essential genes, trying not to disturb any known biological features in order to minimise functional disruption.

DGF-327 and MDS42 are both products of genome reduction efforts. Because the genome of DGF-327 is smaller, it shares 25 of the 37 landing site sequences designed in MDS42. DGF-327 contains a large inversion of ~700 kbp with respect to MDS42, owing to the architecture of their respective parental strains (**Figure 2.2b**). This inversion occurred by recombination between ribosomal RNA operons D and E. In order to carry out stepwise replacement, we planned to convert the recipient MDS42 strain into a DGF-327-like configuration by 'reverting' the inversion using asymmetric Cre-LoxP-mediated recombination[196,197].

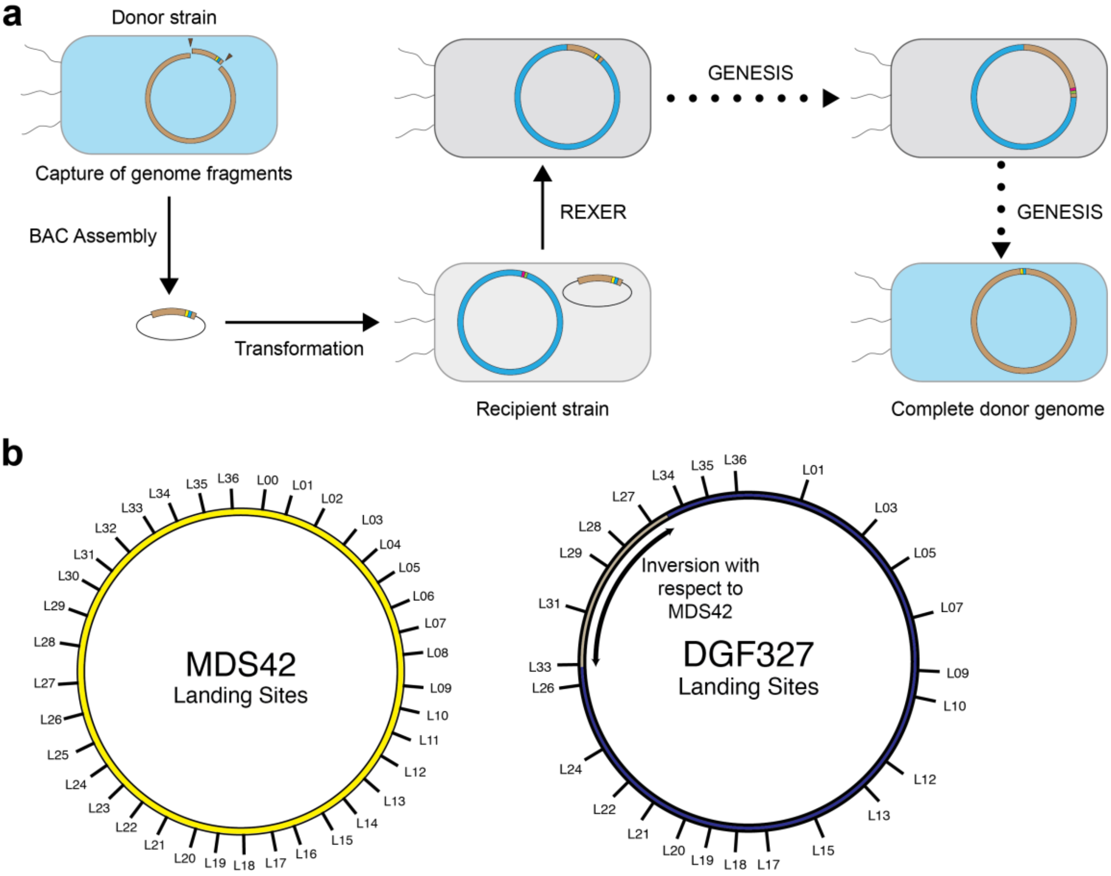


Figure 2.2 – Stepwise genome conversion. a) Schematic of the strain conversion strategy. Large sections of the donor genome are cloned into BACs. These are then transformed into the recipient strain, where the donor section in the BAC replaces the corresponding recipient section via REXER. Iteration of this process (GENESIS) yields hybrid strains that are progressively more 'donor-like', until complete replacement is achieved. **b)** Diagram of the 37 landing sites designed for genome replacement in the recipient *E. coli* MDS42 genome (left), ~100 kbp apart. 25 of these are also present in donor strain *E. coli* DGF-327 (right). The two strains differ in their architecture by an inversion of ~700 kbp, indicated by the double-headed arrow.

We planned to begin multiple parallel REXER-mediated replacements in distinct strains, starting from different genomic coordinates. We envisioned that we could merge these MDS42/DGF-327 chimeras into a single, fully DGF-327 strain, by directed conjugation[99] or by chromosomal transplant, a technique that was, at the time, in development, and which is described in Chapter 4. **Figure 2.3** depicts an overall workflow for the genome replacement, from capture of DGF-327 genomic segments into BACs through their to verification and establishment into an MDS42 strain. The details of the workflow are explained in the following sections.

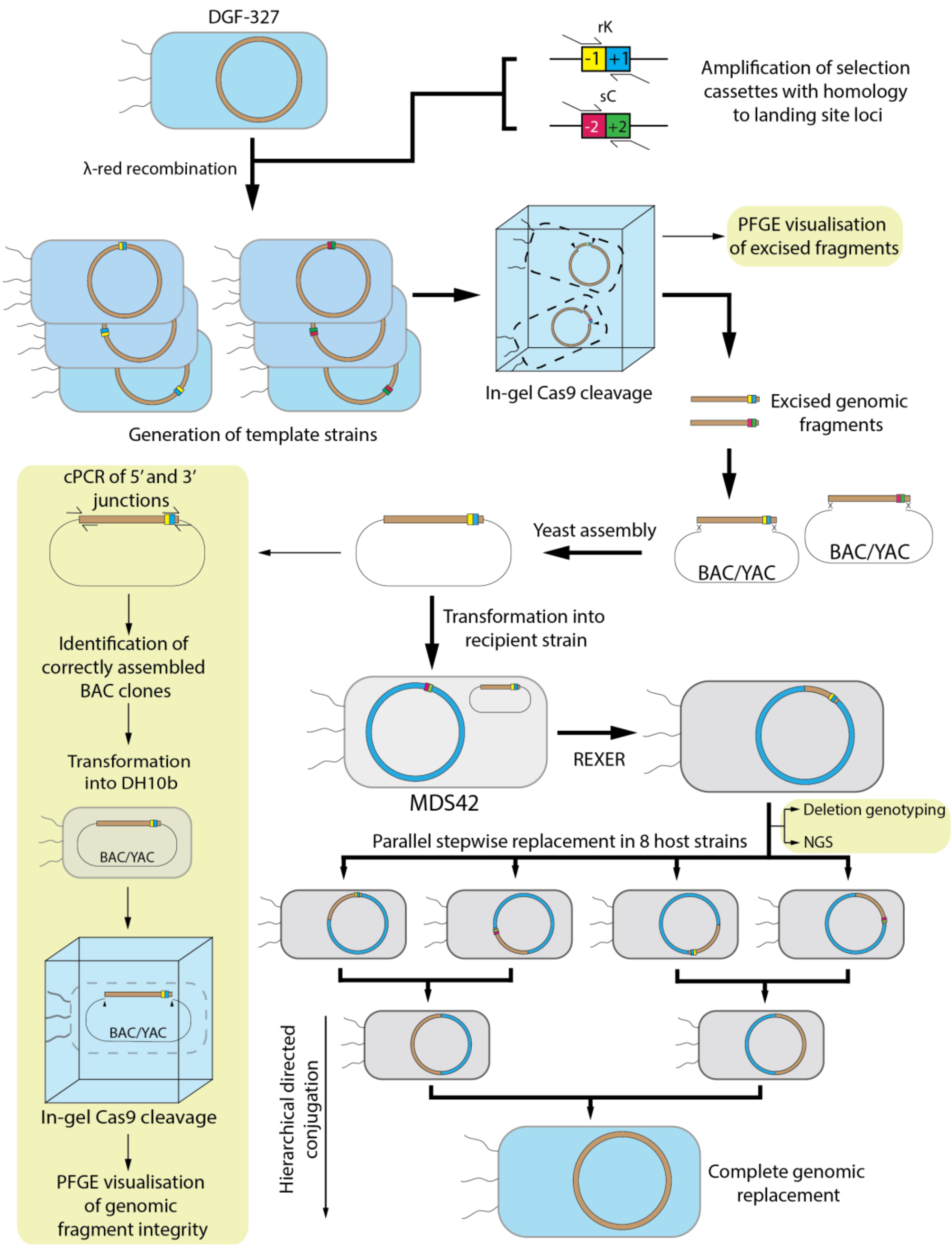

DGF-327
rK
-1
+1
sC
-2
+2
Amplification of selection cassettes with homology to landing site loci
λ-red recombination
PFGE visualisation of excised fragments
In-gel Cas9 cleavage
Generation of template strains
Excised genomic fragments
cPCR of 5′ and 3′ junctions
Yeast assembly
BAC/YAC
BAC/YAC
Identification of correctly assembled BAC clones
Transformation into recipient strain
Transformation into DH10b
REXER
MDS42
BAC/YAC
Deletion genotyping
NGS
Parallel stepwise replacement in 8 host strains
BAC/YAC
In-gel Cas9 cleavage
Hierarchical directed conjugation
PFGE visualisation of genomic fragment integrity
Complete genomic replacement

Figure 2.3 – Experimental flowchart for stepwise genome replacement. Verifications and quality control steps are represented by green boxes and thin lines.

2.2 – Cloning large sections of the *E. coli* genome into REXER BACs

We devised a strategy for constructing BACs containing defined fragments of the donor DGF-327 genome, which would be substrates for REXER in an inter-strain stepwise genome replacement. Transformation-associated recombination (TAR) has been extensively used to isolate large sections of prokaryotic and eukaryotic genomes into yeast artificial chromosomes (YACs), in both linear[198,199] and circular forms[170,200–202]. The technique exploits the high intrinsic activity of the homologous recombination machinery in *S. cerevisiae;* upon co-transformation with a vector, it can stitch together multiple DNA fragments with overlapping homologous ends into a single, independently replicating unit. TAR has been useful in the cloning of large genes and gene clusters[203,204], the construction of genomic DNA YACs for genome mapping and sequencing[205–208], and the assembly of synthetic *Mycoplasma* genomes[168,171,187,209].

The efficiency of capture of genome fragments into YACs is greatly increased when the homology arms lie near double-stranded breaks. Usually, target genomes are digested with restriction endonucleases, and the ends of the YAC vectors are designed to be homologous to the restriction fragment ends[208,210–212]. In our case, constructing BACs with this approach would confine the REXER step sizes to the restriction patterns of the genome, and would be incompatible with the general replacement strategy outlined above. Generating the desired DNA double stranded breaks with CRISPR/Cas9 seemed like a much more attractive option - this approach has been previously described for generating defined YACs of 55 kbp from the human genome at high efficiency[213].

The isolation of DNA molecules larger than 100 kbp is challenging because DNA molecules become much more prone to shearing with increasing size. A useful approach for overcoming this limitation is embedding the starting DNA in agarose plugs, which provides a protective matrix from liquid shear forces and minimises DNA breakage[173,211,212,214]. Recently, CRISPR/Cas9 was used to isolate large regions of the *E. coli* chromosome by *in vitro* in-gel cleavage (termed CATCH[215]), and the resulting fragments were cloned into BACs of up to 150 kbp by Gibson assembly. Although Gibson assembly has been used once to clone a BAC of 300 kbp in *E. coli* [169], the authors of CATCH failed to clone BACs of 200 kbp. Since the efficiency of electroporation in *E. coli* drops greatly with increasing DNA length, we hypothesised that the low abundance of the assembled circular BAC in the Gibson assembly mix may be insufficient for robust electroporation of DNA pieces >200 kbp.

We reasoned that by combining CRISPR/Cas9 cleavage of agarose-embedded DNA with TAR, and purifying the DNA from yeast clones that had correctly assembled the BACs, we may be able to obtain homogeneous BAC samples at higher concentrations and more efficiently. This may allow us to generate BACs in excess of 200 kbp and transform them into *E. coli* more reliably, which would allow us to probe the size limits of a REXER step. Additionally, we expected TAR to generate fewer errors than Gibson assembly at the insert-vector junctions; this was important, as downstream these would serve as homology arms during REXER.

Testing the workflow for capturing genomic regions into BACs for REXER

The workflow we devised for assembling the donor genome into BACs for REXER is illustrated in **Figure 2.4**. First, donor *E. coli* DGF-327 cells are embedded into low-melting point agarose plugs. This is followed by in-gel cell lysis with lysozyme and proteinase K. Cas9 can then cleave the exposed chromosomal DNA in-gel, at the sites defined by custom sgRNAs. The cleaved genomes are subsequently released from the agarose matrix by melting of the plugs and digestion with β-agarase. The agarose breakdown products are then depleted from the mixture by drop dialysis, and the resulting DNA solution is co-transformed in *S. cerevisiae* with a BAC/YAC vector containing homology to both ends of the excised genomic fragment. Total DNA is then extracted from the yeast clones containing a correctly assembled BAC/YAC, and transformed into recipient *E. coli* MDS42 cells for REXER. The BAC/YAC vectors are equipped with both bacterial and yeast replication and segregation elements, but throughout this dissertation they are referred to as ‘BACs’ for simplicity.

We set out to test the assembly workflow. For this test, we used a luminescent derivative of *E. coli* MDS42 (previously generated by our group[185]) as a starting genome. Our target for excision was a 100 kbp section that contained, interspersed, the genes that form the *luxABCDE* operon (**Figure 2.5a**). This operon confers luminescence to the harbouring strain, and facilitates screening for complete assemblies. The section of interest was flanked at its 3′ end by a *sacB-cat* double selection cassette, which confers sensitivity to sucrose and resistance to chloramphenicol, respectively. We designed sgRNAs flanking this region (**Appendix A.3**), including the *sacB-cat* cassette, and performed in-gel digestion with Cas9. As a control, we treated separate agarose plugs with restriction enzyme FseI, which cuts the genome twice approximately 330 kbp apart. After cleavage, we analysed the plugs by pulse-field gel electrophoresis (PFGE) and confirmed the presence of bands corresponding to the expected cleavage products (**Figure 2.5b**).

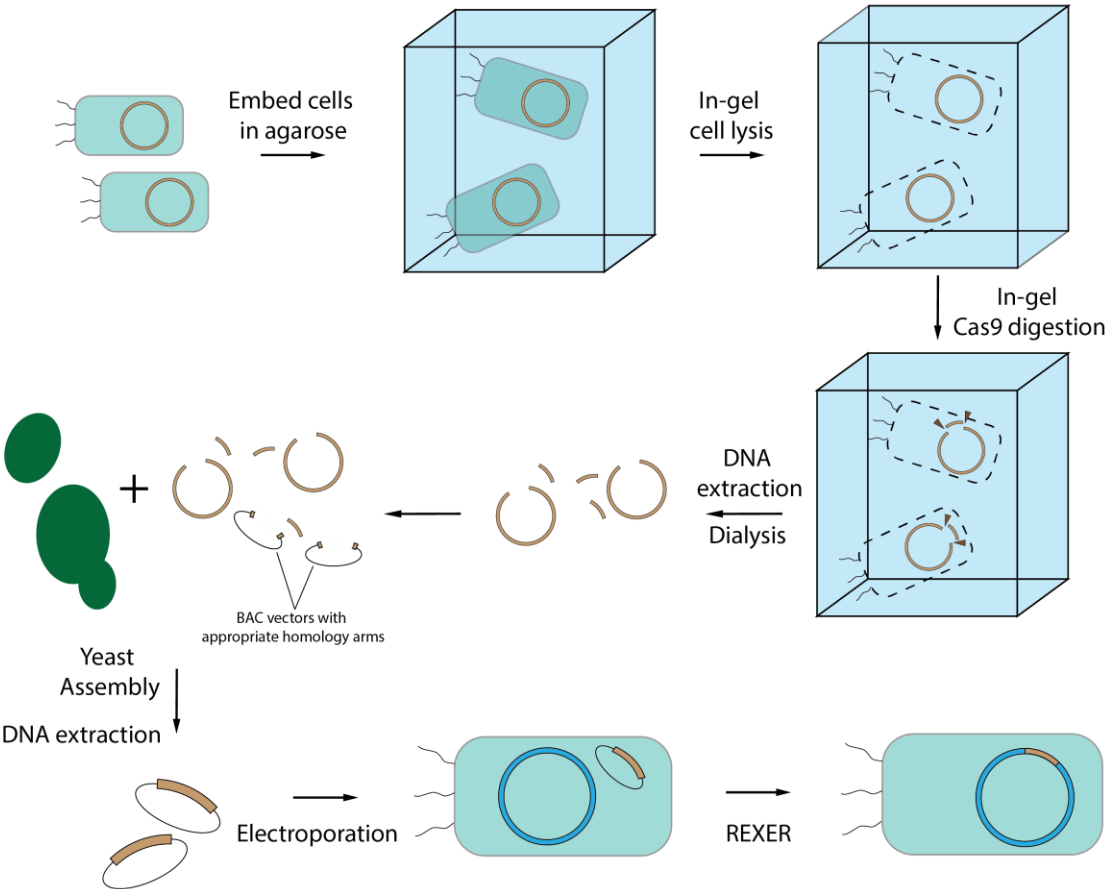


Figure 2.4 - Workflow for the capture of donor genome fragments into REXER BACs.

This group had previously generated two analogous versions of a shuttle BAC/YAC vector for REXER, containing i) the BAC replication and segregation machinery, ii) a yeast centromere (*CEN6*) fused to an autonomously replicating sequence (ARS), iii) a *URA3* auxotrophy marker and iv) a copy of either *sacB* or wild-type *rpsL* for counter-selection during REXER. We amplified the *rpsL* version of this vector with primer overhangs containing approximately 80 bp of homology to either end of the cleavage product, and transformed the product into *S. cerevisiae* spheroplasts together with the cleaved genomic DNA, as in ref[216].

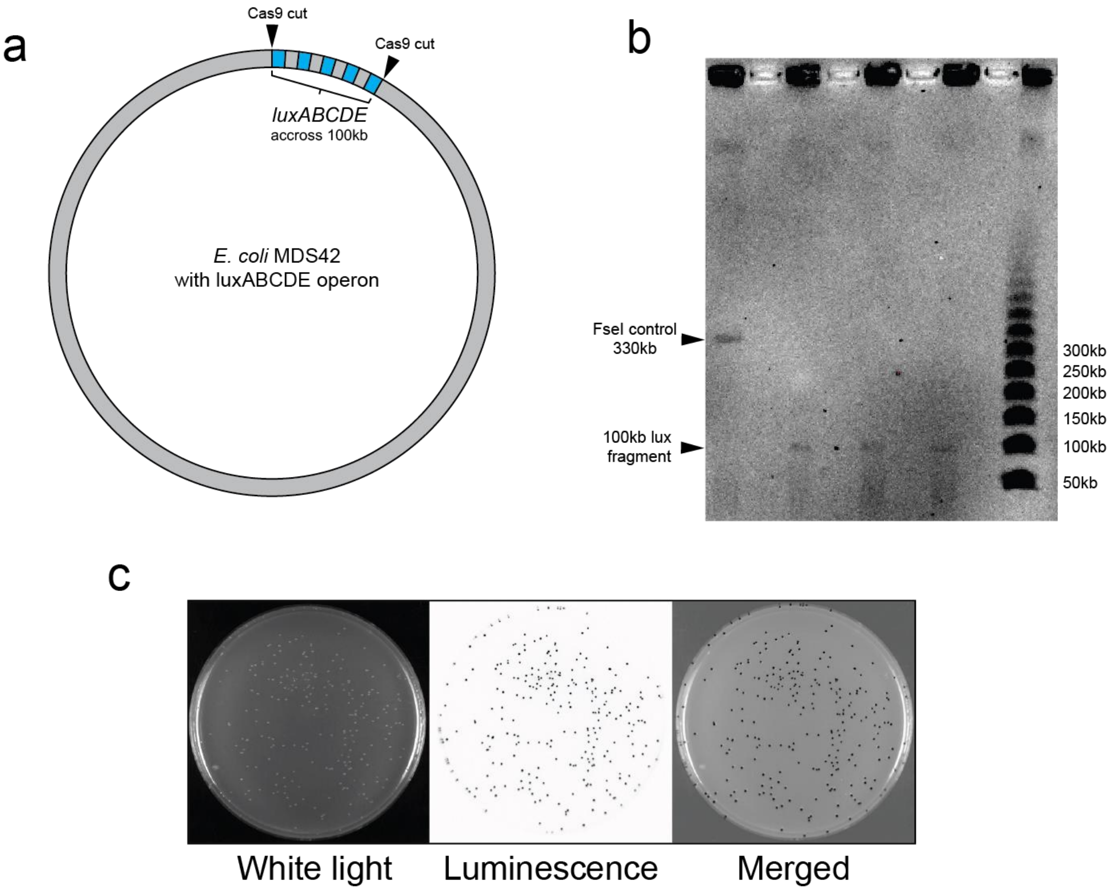


Figure 2.5 - Demonstration of BAC assembly from excised fragments. a) A ~100 kbp region of the genome of a luminescent derivative of *E. coli* MDS42, containing the *luxABCDE* operon, is targeted for excision by Cas9. **b)** Pulse-field gel electrophoresis of 3 post-digest agarose plugs shows a band migrating at approximately 100 kbp, corresponding to the size of the targeted excision. On the left is a control of the same batch of plugs, digested with FseI (expected band was ~330 kbp). The ladder is New England Biolabs Lambda PFG Ladder. **c)** A total DNA extract from a positive clone from yeast assembly was electroporated into wild-type *E. coli* MDS42. Over 99% of the colonies were luminescent, consistent with the presence of a BAC containing the intact 100 kbp harbouring *luxABCDE*.

In order to identify yeast clones bearing correctly assembled BACs, we performed colony PCR across the 5´and 3´ BAC/YAC vector-insert junctions. We screened 87 clones, of which 35 (approx. 40%) contained the expected products at both ends (for representative PCR products see **Appendix A.2**). Total DNA was purified from one of the positive clones, and transformed into wild-type *E. coli* MDS42 by electroporation. The transformation yielded over 200 colonies, of which >99% were luminescent

(**Figure 2.5c**). This demonstrated that the cells contained a BAC harbouring all 5 genes required for luminescence, suggesting that our target region had been assembled intact into a BAC. These results served to validate our workflow for assembling sections of the *E. coli* genome into BACs, in a form ready for REXER. Of note, parallel attempts to generate the same BAC by Gibson assembly using the same cleaved DNA and BAC/YAC vector failed to yield any transformants in our hands.

2.3 – Generating genomic templates for assembly

Note: Integration of selection cassettes at DGF-327 landing sites was performed jointly with Christopher Wan and Yonka Christova.

The iteration of REXER requires the alternation of two distinct double selection cassettes (-1/+1 and -2/+2). At each step, one will be on the genome and the other will be in the BAC together with the DNA of interest. As described above, in our replacement strategy 'landing sites' mark the sequences of recombination during REXER, as well as the sites of integration of selection markers throughout the replacement process, and DGF-327 contains 25. We reasoned that if we had an array of DGF-327 strains, each containing a selection cassette in a given landing site, we could simultaneously i) verify that they were functional selection cassette locations for REXER and ii) use their genomes as templates for excision of the relevant fragments, effectively coupling the DNA of interest to the REXER selection cassette. The two double selection cassettes used here are:

- *rpsL-kanR* (**rK**, -1/+1): It confers resistance to kanamycin and, in an *rpsL*K43R genetic background, sensitivity to streptomycin. While *rpsL*K43R confers streptomycin resistance, the wild-type copy of *rpsL* in the cassette is dominant negative.
- *sacB-cat* (**sC**, -2/+2): It confers resistance to chloramphenicol and sensitivity to sucrose.

We set out to generate templates for genomic excision of DGF-327 DNA by systematically integrating double selection cassettes at these 25 positions using classical lambda-red recombination[183,184]. The iterative nature of REXER dictates that in theory only one of the two selection cassettes would be required at each landing site. However, with views of retaining flexibility in the choice of our step sizes, we decided to generate 50 strains, each harbouring one of the selection cassettes at a particular landing site. This provided a wider variety of templates that would allow us to rapidly adapt should we decide to increase our step size and 'step over' a particular landing site, or conversely shorten any REXER steps that may prove problematic.

We began by introducing the K43R mutation in *rpsL* in the DGF-327 genome by recombineering with a ssDNA oligo, in order to render it resistant to streptomycin[102]. Next, we amplified the selection cassettes by PCR with primer overhangs containing 50-100 bp of homology to their target landing sites, and integrated them individually and separately in DGF-327$^{\textit{rpsL}\text{K43R}}$ by classical lambda-red recombination. We re-streaked the resulting colonies to reduce the background derived from dead cells, and verified integration of the cassettes at their designated landing sites by colony PCR using primers flanking the sites of integration (**Appendix A.1**). We performed Sanger sequencing of the integration sites to verify that no mutations had been introduced in the homology regions during recombination. In order to assess the selection phenotypes of the resulting strains, we resuspended the recombinant colonies in water and stamped them in LB agar plates supplemented with the relevant selective agents (**Appendix A.1**). All strains exhibited the desired selection phenotypes: resistance to kanamycin and sensitivity to streptomycin (rK) or resistance to chloramphenicol and sensitivity to sucrose (sC). From here on, the strains will be named with the selection cassette that they contain plus the landing site position; e.g. DGF-327$^{\textit{rpsL}\text{K43R/rK03}}$ contains an rK cassette at landing site L03.

2.4 – Probing the step size limits of REXER

The largest replacement our group had achieved with REXER was about 120 kbp. Since the assembly of larger BAC/YACs in yeast has been reported (reviewed in ref[182]) and is less likely to be limiting, we envisioned that the ability to perform longer REXERs may accelerate the replacement process.

Having validated a pipeline for assembly of *E. coli* genome fragments into REXER BACs, we set out to probe the size limits of REXER by replacing MDS42 genome segments of varying size by their corresponding DGF-327 fragments. Using the workflow described in the previous sections, we used DGF-327$^{\textit{rpsL}\text{K43R/sC03}}$ as a template strain for excising the region comprising L01-L03 (180 kbp), and DGF-327$^{\text{rpsLK43R/sC05}}$ as a template for excising regions L03-L05 (140 kbp) and L01-L05 (320 kbp) (**Figure 2.6a**). All excision products contained *sacB-cat* cassettes at their 3' ends.

We assembled the three fragments into BACs for REXER by TAR as above. We screened for correct assemblies by performing PCR across the 5' and 3' vector-insert junctions. 16/24 clones (75%) were positive for the 140 kbp fragment, while both 180 kbp and 320 kbp fragments had 5/32 positive hits (15.6 % - representative examples of positive PCR products in **Appendix A.2**).

REXER of both L01-L03 (180 kbp) and L01-L05 (320 kbp) requires an MDS42 recipient strain with an rK selection cassette at L01, while L03-L05 (140 kbp) requires an rK at L03. We prepared MDS42$^{\textit{rpsL}\text{K43R/rK01}}$ and MDS42$^{\textit{rpsL}\text{K43R/rK03}}$ by classical lambda-red recombination, as above. We extracted total DNA from one clone of each assembly, and electroporated the purified DNA into the corresponding MDS42 derivatives. We obtained transformants for the 180 kbp BAC in MDS42$^{\textit{rpsL}\text{K43R/rK01}}$ and the 140 kbp BAC in MDS42$^{\textit{rpsL}\text{K43R/rK03}}$. However, we failed to transform any of 5 different yeast DNA extracts of the 320 kbp BAC into MDS42$^{\textit{rpsL}\text{K43R/rK01}}$. This was unsurprising as we knew that 320 kbp is near the upper size limit for BAC transformation in *E. coli*[217–219].

We proceeded with REXER of 140 kb and 180 kbp in their respective strains, as in the standard protocol. We set out to try both REXER2 and REXER4. Cells were grown in arabinose to induce expression of Cas9 and the lambda-red machinery, and then made electrocompetent. This was followed by electroporation of a pMB1 plasmid encoding the spacers for Cas9 cleavage of the BAC, in REXER2, or both the BAC and the host genome, in REXER4. After recovery, cells were spread on selective medium containing chloramphenicol and streptomycin. For the 140 kbp and 180 kbp replacements, REXER2 yielded ~1.1x10^5 and ~3000 CFUs, respectively, while REXER4 yielded ~1.6x10^5 and ~4x10^5 CFUs. 8 to 16 of the resulting colonies were streaked to reduce any background.

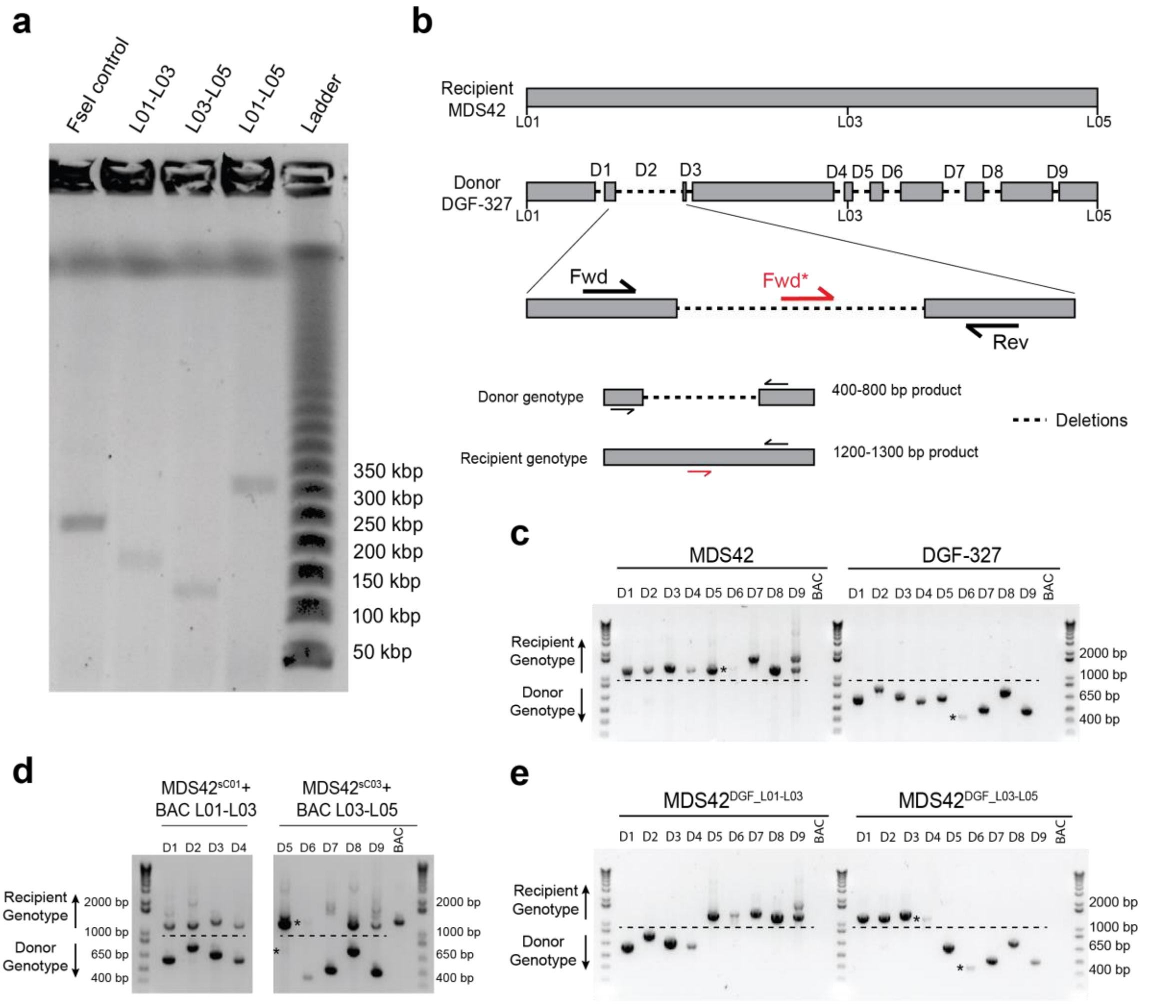


Figure 2.6 – Probing REXER step size limits. a) Pulse-field gel electrophoresis of DGF-327 plugs after Cas9 cleavage of fragments L01-L03 (~180 kbp), L03-L05 (~140 kbp) and L01-L05 (~320 kbp). All excised fragments migrate accordingly with their expected size. On the left is a control digest of wild-type DGF-327 plugs with FseI, which cuts twice ~246 kbp apart. **b)** Schematic of multiplex PCR genotyping. On top is a comparison of the region L01-L05 in strains MDS42 and DGF-327. The regions that have been deleted in DGF-327 (donor) but prevail in MDS42 (recipient) are indicated by a dotted line. Deletions are labelled D1 to D9. At each deletion, PCR is performed with 3 primers; Fwd and Rev are placed outside the deletion region, and yield a product of 400-800 bp only in a DGF-327 genotype. If the genotype at a deletion site is MDS42, then PCR occurs between Fwd* and Rev, yielding a product of 1200-1300 bp. PCR products between Fwd and Rev in an MDS42 genotype do not form due to excessive length. **c)** PCR products of MDS42 and DGF-327 controls at each of the deletion sites. PCR products above the dashed line correspond to recipient genotype (MDS42), whereas products below the line correspond to donor genotype (DGF-327). **d)** Deletion profiles of pre-REXER MDS42 strains harbouring BACs that contain DGF-327 regions L01-L03 (left) and L03-L05 (right). As expected, products for both the donor and recipient genotypes are present. Faint PCR products are indicated by asterisks. 'BAC' PCR targets *URA3*. **e)** Post-REXER strains MDS42$^{DGF\_L01\text{-}L03}$ and MDS42$^{DGF\_L03\text{-}L05}$ show deletion profiles consistent

(continued from previous page) with replacement of the corresponding sections of their genomes by donor DGF-327 DNA.

DGF-327 has a number of genome deletions compared to MDS42, and we leveraged these to evaluate the success of the replacement. We performed multiplex colony PCR at the deletion junctions (illustrated in **Figure 2.6b**), and designed primers for generating differential strain-dependent products. In this way, at each deletion locus a local DGF-327 genotype yields a 400-800 bp PCR product, whereas an MDS42 genotype yields a 1200-1300 bp product (**Figure 2.6c**). In the region spanning L01-L05, there are 9 deletions in DGF-327 with respect to MDS42 – these are labelled D1-D9 (**Figure 2.6b**). D1-D4 are in the L01-L03 region (180 kbp replacement), whereas D5-D9 are in L03-L05 region (140 kbp). Upon transformation of both BACs in the relevant MDS42 strains, we observed both the large and small PCR products at each of the junctions, consistent with the simultaneous presence of both genotypes prior to REXER (**Figure 2.6d**).

Both REXER2 experiments yielded colonies containing DGF-327 genotype at all relevant junctions. 1 out of 8 clones resulting from the 180 kbp replacement were positive across D1-D4, and the same was true for 1 out of 5 clones for 140 kbp across D5-D9. The deletion genotypes of the positive clones are shown in **Figure 2.6e**. However, none of the clones we screened for the REXER4 experiments (16 for 180 kbp and 5 for 140 kbp) exhibited DGF-327 genotype across all relevant junctions; instead, they were a mixture of DGF-327 and MDS42. Although the sample sizes are small, these results are consistent with previous observations that REXER4 produces more CFUs than REXER2, but also more crossovers[185]. From these results, it seems likely that REXER may be able to mediate replacement of over 200 kbp in a single step, but either the delivery or stable propagation of larger BACs constitutes the major bottleneck.

2.5 – Stepwise genome replacement

Iterating REXER

At this point, we had replaced the MDS42 genome with DGF-327 DNA across L01-L03, and the resulting MDS42/DGF-327 chimeric strain (termed MDS42$^{\mathrm{DGF\_L01\text{-}L03}}$)

contained a *sacB-cat* cassette at landing site L03. Knowing that we could only reproducibly transform REXER BACs under 200 kbp, we prepared BACs for the following REXER steps. These were L03-L05 (140 kbp), using DGF-327$^{\textit{rpsL}\text{K43R/rK05}}$ as a template, and L05-L07 (159 kbp), using DGF-327$^{\textit{rpsL}\text{K43R/sC07}}$ as a template. As above, we screened for correct yeast assemblies by colony PCR at the 5' and 3' insert-vector junctions (representative examples in **Appendix A.2**).

In order to assess the integrity of the BACs and rule out large deletions or mis-assemblies, we extracted total DNA from one PCR-positive yeast clone for each assembly and transformed the BACs into *E. coli* DH10b. We then embedded these strains into agarose plugs and performed in-gel Cas9 digestion with sgRNAs targeting the BAC vector-insert junctions. We proceeded to verify whether the linearized inserts would migrate according to their expected sizes in PFGE, and we found that this was the case (**Figure 2.7a**).

In order to proceed with replacement of L03-L05 in MDS42$^{\text{DGF\_L01-L03}}$, we first had to remove the spacer plasmid that remained from the previous REXER step. Passaging the cells and screening for loss of the plasmid turned out to be a very time-consuming task, and after several days we failed to isolate a plasmid-free clone. We realised that doing this after each REXER iteration would delay the replacement process, and the extra passaging time may increase the chances of acquiring unwanted mutations. We reasoned that we may be able to provide the spacers necessary for REXER in a plasmid-free form, by electroporating them as linear dsDNA. Their inability to be replicated in *E. coli* would block their propagation through cell divisions, and would remove the need to passage the cells after every REXER step. However, linear DNA is electroporated less efficiently than plasmid DNA and is a substrate for exonucleases, so it was unclear whether the linear spacer cassette would be sufficiently abundant and stable *in vivo* to enable robust transcription of the crRNAs.

We used the plasmid encoding spacers for REXER of L01-L03 as a template to amplify a linear spacer cassette (**Figure 2.8**). We then repeated the REXER of L01-L03 in MDS42 exactly as above, but electroporating several micrograms of linear spacer PCR product instead of the plasmid. Among the resulting colonies, we observed replacement across L01-L03 in 3 out of 8 clones (**Figure 2.7c**). None of the resulting clones could grow on ampicillin (**Figure 2.7f**), indicating that the crRNAs were not being transcribed from background spacer plasmid but rather from the linear cassette.

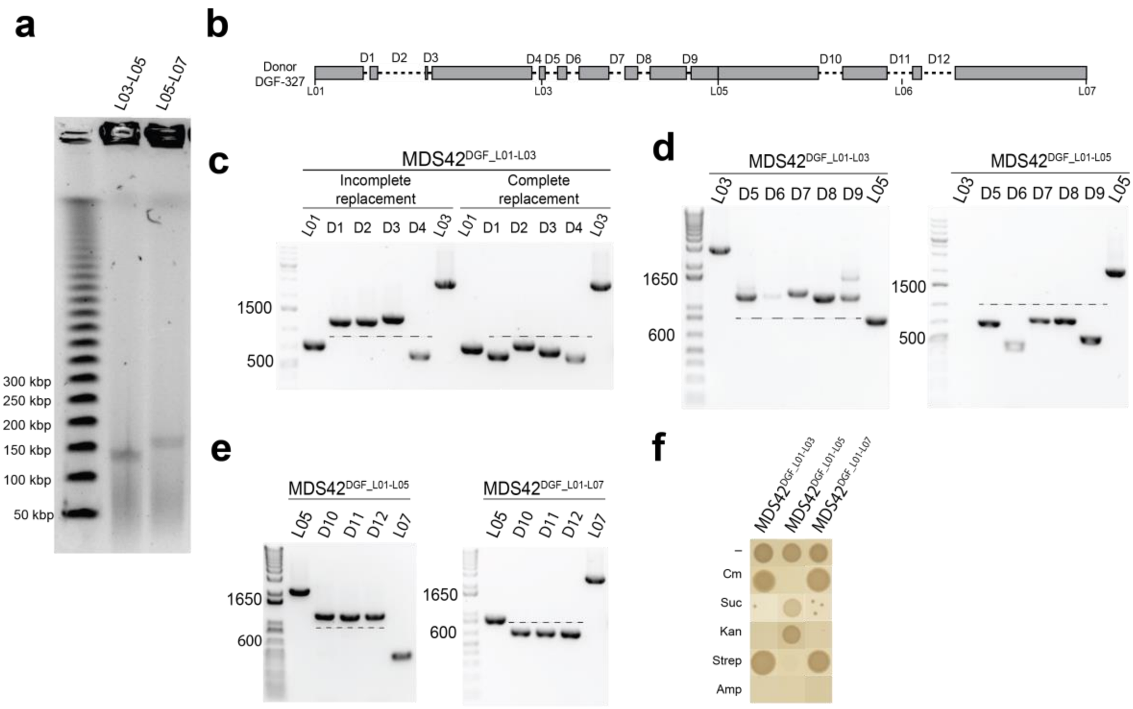


Figure 2.7 – Iterative REXER replaces 664 kbp of the *E. coli* MDS42 genome. a) Pulse-field gel electrophoresis of linearized BAC inserts for genomic regions L03-L05 and L05-L07. Both migrate according to their expected size. **b)** Deletion map of the region spanning L01-L07. Dotted areas are deleted in DGF-327 with respect to MDS42. Landing site positions and deletions D1-D12 are indicated. **c)** Deletion genotype after REXER of L01-L03, as in **Figure 2.6e**, but with linear spaces. The left shows an example of incomplete replacement, where D1-D3 have retained MDS42 genotype. On the right is a complete replacement. PCRs flanking the replacement boundaries L01 and L03 are consistent with wild-type genotype and the presence of *sacB-cat*, respectively. **d)** Deletion genotype before (left) and after (right) REXER of L03-L05. After REXER, DGF-327 genotype is observed at junctions D5-D9. The product size at L05 after REXER increases due to the presence of an *rpsL-kanR* cassette. L03 doesn't yield a product because the specific primer used here binds within a DGF-327 deleted region. **e)** Deletion genotype before (left) and after (right) REXER of L05-L07, analogous to d). The product at L07 increases in size, consistent with the gain of *sacB-cat*. Conversely, at L05 it decreases due to the loss of *rpsL-kanR*. **f)** Selective growth assay of the REXER intermediates leading to MDS42$^{DGF\_L01\text{-}L07}$. Strains containing *sacB-cat* (MDS42$^{DGF\_L01\text{-}L03}$ and MDS42$^{DGF\_L01\text{-}L07}$) are sensitized to 7.5% sucrose (Suc) and resistant to 20 µg/mL chloramphenicol (Cm). Conversely, MDS42$^{DGF\_L01\text{-}L05}$ contains *rpsL-kanR* and is sensitive to 100 µg/mL streptomycin (Strep) but resistant to 50 µg/mL kanamycin (Kan). None of the strains is resistant to 50 µg/mL ampicillin (Amp), indicating the absence of spacer plasmid. `-´ indicates plain LB agar.

We transformed the BAC containing the DGF-327 fragment L03-L05 and carried out REXER, again with linear spacers. 7 out of 8 of the clones screened had a full DGF-327 genotype across all relevant deletions (**Figure 2.7d**). The resulting strain, MDS42$^{\mathrm{DGF\_L01\text{-}L05}}$, was a direct substrate for the third round of REXER, replacing L05-L07. There are 3 deletions in L05-L07 that differentiate DGF-327 from MDS242 (**Figure 2.7b**), and 4 out of 8 clones had DGF-327 genotype at all 3 deletions (**Figure 2.7e**). After 3 steps, we sequenced the genome of the resulting MDS42$^{\mathrm{DGF\_L01\text{-}L07}}$ strain and verified that the region from L01-L07 had been replaced as expected. These results gave us confidence that REXER can be iterated rapidly multiple times, and that linear spacer cassettes are efficient enough to mediate robust cleavage.

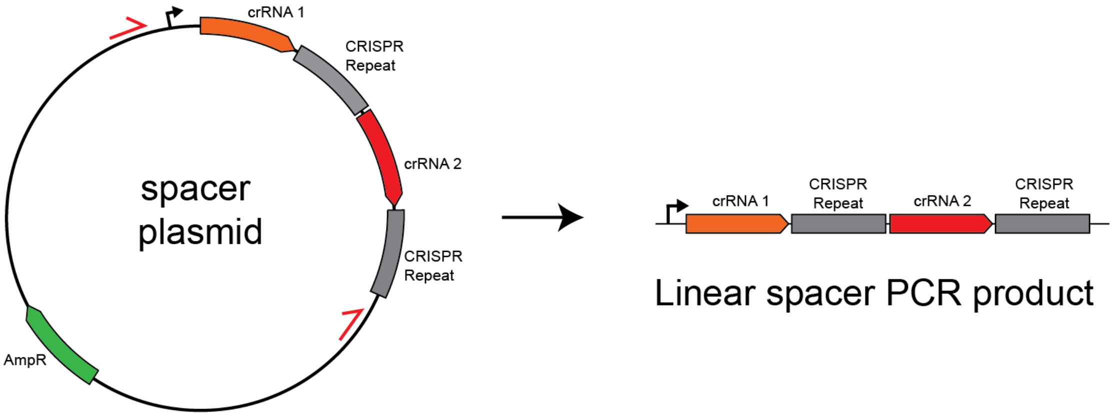


Figure 2.8 - Linear spacer generation. Spacer sequences for Cas9 cleavage (labelled crRNA1/2) are cloned in arrays that mimic the natural CRISPR architecture, in between CRISPR repeats. In the original REXER protocol, these are electroporated in a plasmid form. Here, to facilitate iteration, we generate linear dsDNA spacer arrays by PCR amplification from the spacer plasmid (primers shown in red).

Constructing a library of DGF-327 fragments

In order to continue replacement, we utilized the landing sites specified in Section 2.2 to define the minimal number of steps necessary to achieve complete replacement, aiming to keep the fragments below 200 kbp to minimise electroporation complications. We defined the steps under two assumptions:

i) we would initiate multiple parallel replacements from different points of the MDS42 genome and combine the resulting chimeric genomes by directed conjugation into a single DGF-327 strain

ii) the replacement steps encompassing the 700 kbp inversion would be performed in an MDS42 host in which the region had been flipped into a DGF-327-like configuration

The envisioned replacement consists of 20 steps, ranging from 70 to 211 kbp (**Figure 2.9a).** We set out to assemble BACs for L07-L22, as described above. The size of each step, template strain, selection cassette, assembly efficiency and completion status are detailed in **Table 2.1**. Following assembly of the different fragments, we analysed the integrity of each BAC by PFGE as above. We found that, while most DGF-327 fragments were the expected size, the BAC harbouring fragment L19-L21 appeared to contain a 100 kbp truncation (**Figure 2.9b**).

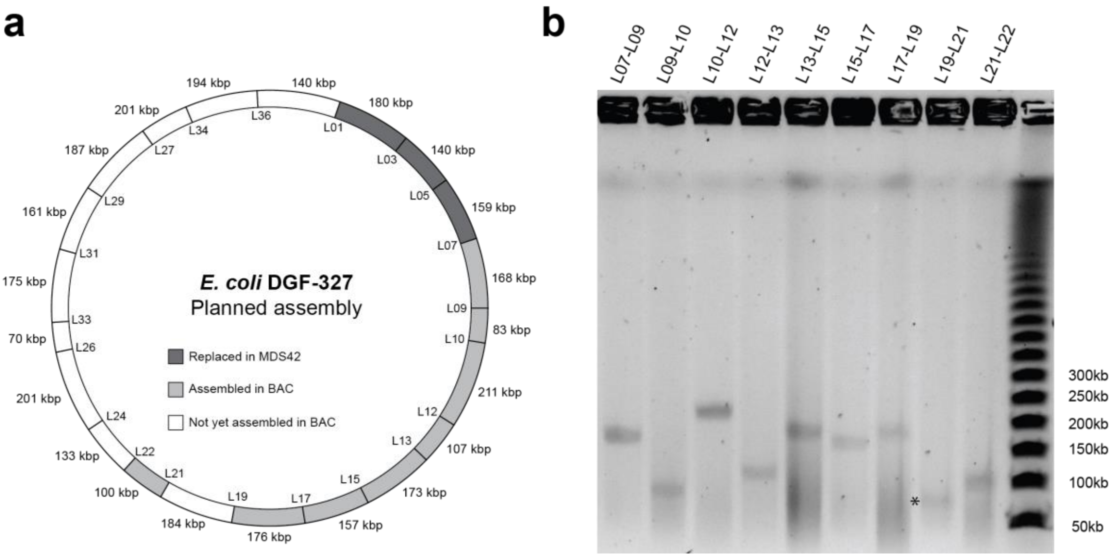


Figure 2.9 - Constructing a library of DGF-327 in REXER BACs. a) The planned steps for replacement are shown in the DGF-327 genome map; the size of each step is indicated. 56% of the DGF-327 genome is cloned into REXER BACs, and ~15% has been implemented in the recipient strain. **a)** Pulse-field gel electrophoresis of *E. coli* DH10b cells harbouring REXER BACs corresponding to the indicated steps. All BAC inserts migrate accordingly to their expected size except L19-L21 (indicated by an asterisk), which appears to contain a ~100 kbp truncation. The ladder is New England Biolabs Lambda PFG.

It is known that chromosome and YAC replication in yeast requires autonomously replicating sequences (ARS), which serve as origins of replication[220,221]. ARS are A/T-rich sequences which bind the Origin Recognition Complex[222], and are thought to support the replication of 120-300 kbp of DNA[223]. When constructing YAC libraries of eukaryotic genomes, this is often not a consideration as ARS-like sequences are naturally present every 20-40 kbp[224]. However, these sequences are rarer in the simpler prokaryotic genomes. There have been reports of failure to clone large YACs from bacterial species[211,212] – the limitations were circumvented by splitting the target fragments or by providing additional ARS sequences, suggesting that, depending on the sequence, some YACs may require the introduction of additional origins to be propagated stably. This notion has been further supported in a recent report, where 11 ARS sequences were insufficient and 16 were necessary to stably replicate a GC-rich 750 kbp synthetic sequence in yeast[189].

Step	Fragment size (kb)	Template strain	Assembly efficiency (cPCR)	Completion
L01-L03	180	MDS42$^{\textit{rpsL}\text{K43R/rK03}}$	5/32	Yes
L03-L05	140	MDS42$^{\textit{rpsL}\text{K43R/sC05}}$	10/16	Yes
L05-L07	159	MDS42$^{\textit{rpsL}\text{K43R/rK07}}$	5/16	Yes
L07-L09	168	MDS42$^{\textit{rpsL}\text{K43R/sC09}}$	8/16	Yes
L09-L10	83	MDS42$^{\textit{rpsL}\text{K43R/rK10}}$	2/16	Yes
L10-L12	211	MDS42$^{\textit{rpsL}\text{K43R/sC12}}$	3/16	Yes
L12-L13	107	MDS42$^{\textit{rpsL}\text{K43R/rK13}}$	6/16	Yes
L13-L15	173	MDS42$^{\textit{rpsL}\text{K43R/sC15}}$	2/16	Yes
L15-L17	157	MDS42$^{\textit{rpsL}\text{K43R/rK17}}$	12/16	Yes
L17-L19	176	MDS42$^{\textit{rpsL}\text{K43R/sC19}}$	5/16	Yes
L19-L21	184	MDS42$^{\textit{rpsL}\text{K43R/rK21}}$	5/16	No
L21-L22	100	MDS42$^{\textit{rpsL}\text{K43R/sC22}}$	16/16	Yes
L22-L24	133	MDS42$^{\textit{rpsL}\text{K43R/rK24}}$	-	No
L24-L26	201	MDS42$^{\textit{rpsL}\text{K43R/sC26}}$	-	No
L26-L33	70	MDS42$^{\textit{rpsL}\text{K43R/rK33}}$	-	No
L33-L31	175	MDS42$^{\textit{rpsL}\text{K43R/sC31}}$	-	No
L31-L29	161	MDS42$^{\textit{rpsL}\text{K43R/rK29}}$	-	No
L29-L27	187	MDS42$^{\textit{rpsL}\text{K43R/sC27}}$	-	No
L27-L25	201	MDS42$^{\textit{rpsL}\text{K43R/rK25}}$	-	No
L35-L00	194	MDS42$^{\textit{rpsL}\text{K43R/sC00}}$	-	No
L00-L01	110	MDS42$^{\textit{rpsL}\text{K43R/rK01}}$	-	No

Table 2.1 – A library of *E. coli* DGF-327 REXER BACs. '-' indicates that the experiment has not been attempted.

It seems possible that the BAC sequence corresponding to L19-L21 is prone to collapse; splitting this step in 2 or providing an extra copy of ARS in the genomic insert may facilitate replacement of this region. However, more yeast clones would need to be analysed to determine whether this is a true instance of instability or simply a stochastic deletion in the assembly process. Fragment L19-L21 contains ribosomal RNA operon G – it is also possible that the truncation may have occurred through recombination with another rRNA operon within the *E. coli* cell after BAC transformation, as these are known to be recombination hotspots in *E. coli* and *Salmonella*[225,226].

In parallel to this work, we began the synthesis of a recoded *E. coli* genome (see Chapter 3). Early successes in the assembly and replacement of recoded fragments eventually led us to put the DGF-327 replacement on hold, and we did not investigate these matters further. Nevertheless, the influences that this work had on our approach to genome recoding are discussed below.

2.6 – Lessons from the DGF-327 replacement and conclusions

The BAC for the 4th consecutive REXER step, encompassing L07-L09 (184 kbp), was assembled correctly and successfully delivered from yeast into *E. coli* DH10b. However, we systematically failed to deliver the same BAC into MDS42$^{\text{DGF\_L01-L07}}$; not from total yeast DNA extracts nor from miniprep purifications of *E. coli* DH10b cells that contained the L07-L09 BAC. Transformation efficiency is known to vary depending on the host strain[227], and DGF-327 has reduced fitness compared to MDS42. It seemed possible that throughout the replacement process, as the DGF-327 genotype became prevalent in the MDS42/DGF-327 chimeric genome, the fitness of the intermediate strain would gradually decrease and BAC electroporation would become increasingly inefficient. Since replacement of the MDS42 genome with a recoded synthetic version was also likely to come at the cost of fitness, we anticipated that BAC electroporation may become limiting.

This led us to incorporate alternative BAC delivery strategies into our workflow for genome synthesis (see Chapter 3– Total synthesis of *Escherichia coli* with a recoded genome), such as conjugation. We constructed an *oriT-apm*^{R} cassette for integration into our BAC vectors. If we succeeded in electroporating a particular BAC into any healthy strain (e.g. *E. coli* DH10b), we could use it as a donor to transfer the BAC to our strain of interest by conjugation, which does not suffer from the same size limitations as electroporation. Additionally, in order to minimise complications derived from BAC delivery and stable assembly in *S. cerevisiae*, we decided not to assemble BACs in excess of 136 kbp, maintaining the replacement strategy outlined in Section 2.1.

DNA inserts over 300 kbp are attainable in eukaryote-derived YACs[182], and entire Mycoplasma genomes over 1 Mb (G/C content ~35%) have been propagated stably in yeast. However, it has been suggested that when cloning YACs from prokaryotic genomes with increasing G/C content, the stability of the YACs decreases more rapidly with increasing insert size[208,210]. One study failed to capture fragments of the *S. elongatus* genome (G/C ~56%) larger than 142 kbp; obtaining larger YACs required multi-step cloning and supplementation of additional ARS sequences[211]. For *E. coli* DGF-327 [~52% G/C], we observed that 11/12 of the attempted yeast assemblies were successful; 10 of these were above 100 kbp and 8 were above 140 kbp. This gave us confidence that there are relatively few sequences in *E. coli* that may suffer from instability in yeast, and by limiting our assemblies to 100-136 kbp we should further decrease the instances of instability.

In summary, here we set out to perform a technical demonstration of stepwise genome replacement using REXER, by converting the genome of *E. coli* MDS42 into the genome-reduced DGF-327 strain. We established a pipeline for generating BACs from defined regions of the *E. coli* genome based on in-gel Cas9 cleavage, and these were direct substrates for REXER. We also showed that, if BAC delivery is successful, REXER is efficient in mediating replacements up to 180 kbp, and that the inability to deliver larger BACs robustly is the major size limitation. While we did not complete the whole genome replacement, we showed that rapid REXER iterations are possible with the use of linear spacer arrays as opposed to plasmid-based spacers. Once substrate BACs were available, we used linear spacers to replace 664 kbp (~15% of the MDS42 genome) in under three weeks. The combination of genome assembly and genome replacement tools shown here provides a framework for the *ex vivo* assembly and transfer of defined chromosomal segments between strains, in a format compatible with their direct integration in the host chromosome.

Chapter 3 – Total synthesis of *Escherichia coli* with a recoded genome

Where appropriate, figures in this chapter are edited and reproduced with permission from their original publication:

Fredens, J.*, Wang, K.*, **de la Torre, D**.*, Funke, L. F. H.*, Robertson, W. E.*, Christova, Y., Chia, T., Schmied, W. H., Dunkelmann, D. L., Béranek, V., Uttamapinant, C., Gonzalez Llamazares, A., Elliott, T. S. & Chin, J. W. Total synthesis of *Escherichia coli* with a recoded genome. *Nature* **569,** 514–518 (2019).

*equal contribution

Note: This project was carried out jointly with Julius Fredens, Kaihang Wang, Louise Funke and Wesley Robertson, with contributions from all other authors in the original publication. My specific contributions and those of the other authors are clarified throughout the text.

3.1 – Introduction

The genetic code defines the relationship between codon sequence and amino acid identity in genetically-encoded protein synthesis and is, with some exceptions[11–13,16,21], effectively universal across all domains of life. 18 out of the 20 canonical amino acids are encoded by up to six synonymous codons. As discussed in Chapter 1, while synonymous codons encode the same amino acid, codon choice can influence gene expression and regulation in multiple ways. These include mRNA folding and stability[138–144], regulation of translational speed[145] and co-translational folding[146,147]. While several studies have attempted to experimentally evaluate and quantify the impact that synonymous codon choice has in protein expression[139,155,156], the biological implications of codon choice are still emerging, and an understanding of the rules that govern codon choice genome-wide is still lacking.

The systematic substitution of certain codons by their synonyms is an attractive approach for probing the malleability of codon choice genome-wide. Indeed, the removal of a subset of codons from the genome would serve to address the question of whether all codons are required to encode protein synthesis in living organisms. Synonymous codon compression may also provide 'blank spaces' in the genetic code – following removal of their cognate decoding elements, the liberated codons may be amenable for reassignment into unnatural building blocks, and may facilitate the genetically-encoded synthesis of non-canonical polymers. Such recoded organisms may also display interesting and useful properties, including biocontainment[192,228,229] and resistance to viruses[100].

Previous recoding efforts in *E. coli* have aimed at removing all[99,100] or a subset[97,98] of the annotated instances of the amber stop codon by site-directed mutagenesis, while introducing comparable numbers of off-target mutations. Albeit with fitness effects ranging from mild to severe, this has enabled the deletion of RF-1, and the reassignment of TAG to natural and unnatural amino acids by means of amber suppression. Because sense codons are often 1 to 2 orders of magnitude more frequent than amber codons, their removal by site-directed mutagenesis is impractical, and genome synthesis constitutes a more attractive strategy.

Several previous efforts have used large-scale DNA synthesis for probing synonymous codon compression, in sections of the genomes of *Mycoplasma*[187], *E. coli*[185,188] and *Salmonella*[186]. Synonymous replacements, however, may alter biological function in unexpected ways, and choosing appropriate synonyms is not trivial. Some approaches have aimed at predicting viable replacements computationally, integrating several variables such as mRNA secondary structures, codon adaptation levels, and maintenance of ribosomal binding sites to identify the best replacements on a case-by-case basis[188,189]. While these studies were highly ambitious and aimed to carry out aggressive recoding, their success in designing viable sequences has been limited; a subset of recoded sequence designs have been shown to complement wild-type function, but they have not achieved a functional genome to date. This may be due to the challenges inherent to the prediction of biological features from the DNA sequence alone, as discussed in Chapter 1.

Other recoding projects, including the one of our group, have adopted more empirical approaches for defining recoding rules. Our group has previously tested a variety of defined synonymous codon compression schemes (where one codon is systematically replaced by one synonym) to probe recoding across a 17 kbp region of the *E. coli* genome, rich in target codons and essential genes[185]. This region is largely formed by genes involved in cell division, and the interaction between them was expected to magnify any deleterious effects and assist in identifying viable schemes. The recoding schemes were defined by identifying the 'closest' replacements for particular serine, leucine and alanine codons based on several metrics, namely **i)** tRNA adaptation index (tAI[230], which correlates codons to the abundance of cognate isoacceptor tRNAs such that genes with higher tAI values have a larger pool of cognate tRNAs available for decoding and are considered translationally more optimal), **ii)** codon adaptation index (cAI[149], which scores the most 'optimal' codons in a given organism as those that occur more frequently in the organism's genome), and **iii)** translational efficiency index (tE[185], a combination of tAI and cAI). Some recoding schemes were well tolerated across the entire region, whereas others were lethal (**Figure 3.1**). In one case, the cause of failure was pinpointed to a single codon, and a single alternative substitution was sufficient to rescue the viability of the scheme. No single metric correlated to the success of the schemes.

As discussed in Chapter 1, DNA synthesis can now be performed at a scale compatible with the creation of entire synthetic genomes. This is exemplified by the synthesis of multiple *Mycoplasma* genomes up to 1.08 Mb[171], as well as multiple synthetic *S. cerevisiae* chromosomes up to 0.99 Mb[175]. Over the last decade, several technologies for the large-scale manipulation of genomes and the implementation of large synthetic DNA designs have been developed[181,185,186,188]. REXER ('Replicon EXcision for enhanced genome Engineering through programmed Recombination') is a particularly powerful approach developed by our group. As demonstrated in previous publications[185] and Chapter 2 of this dissertation, it can replace over 100 kbp the genome in a single step, and can be rapidly iterated in a process we term GENESIS ('GENomE Stepwise Interchange Synthesis'). Importantly, previous work highlighted that analysing the pool of genomes that result from REXER can provide feedback on deleterious DNA designs at high resolution.

Here we set out to use these technologies to create a synthetic, recoded *E. coli* genome, where 2 sense codons and 1 stop codon are replaced by their synonyms genome-wide. In the resulting strain, the genetic code is compressed to 61 codons, and 59 sense codons (as opposed to the usual 61) encode the canonical amino acids, providing a platform for exploring sense codon reassignment *in vivo*.

3.2 – Results

3.2.1 Design of a recoded genome

Note: The design of the synthetic genome described in this section was carried out by Kaihang Wang, Julius Fredens and Tiongsun Chia, and is discussed here for completeness.

In previous work, this group tested a variety of defined recoding schemes for serine, leucine and alanine (**Figure 3.1**). These amino acids were chosen as candidates for codon removal because their corresponding aminoacyl-tRNA synthetases do not utilize the anticodon stem loop of the cognate tRNAs as identity elements for aminoacylation, and this would ultimately facilitate reassignment. In this study, one serine recoding scheme where TCG and TCA are respectively replaced with AGC and AGT emerged as a promising candidate for genome-wide replacement (**Figure 3.1a**).

We decided to design a synthetic recoded genome based on this scheme. Additionally, we aimed to replace TAG stop codons with TAA, as removal of amber codons is viable and has been reported previously[100].

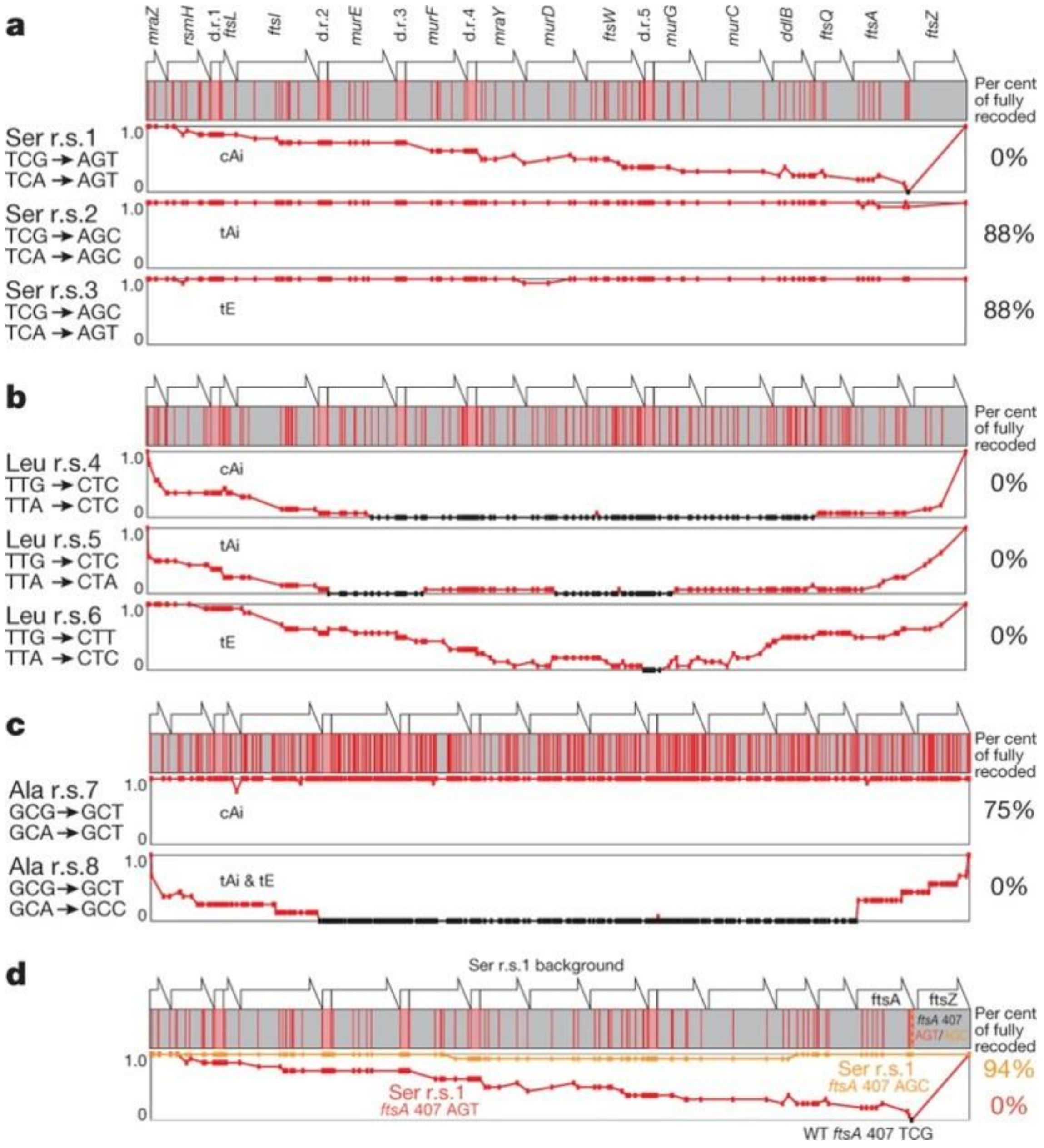


Figure 3.1 - REXER across a 17 kbp region of the *E. coli* genome identifies promising recoding schemes. At the top of each set of graphs is a representation of the genomic region being tested, where genes are drawn as arrows and red lines represent target codons. The graphs below indicate the fraction of clones, out of 16, that were recoded at each codon across the region. Codon positions that were never recoded are indicated in black. The recoding schemes (r.s.) that define each set of recoding rules are indicated on the left of each graph. **a)** Serine recoding schemes r.s.2 and r.s.3 allowed complete recoding, whereas r.s.1 showed a single codon that was never recoded. r.s.3 was chosen for the design of a recoded genome. **b)** All leucine r.s. showed disallowed codon substitutions in this region. **c)** Alanine r.s.7 was well

(continued from previous page) tolerated at all positions, whereas r.s.8 was not tolerated at most positions. **d)** A single T->C substitution in *ftsA* 407 rescues the viability of r.s.1 across the entire region (green line). This figure was adapted with permission from ref[185].

Using the genome of *E. coli* MDS42[194] as a starting point (accession number AP012306.1 as of October 2016), a computational genome sequence was generated where all TCG, TCA and TAG codons in annotated protein-coding genes are replaced with AGC, AGT and TAA respectively (**Figure 3.2a)**. Prokaryotic genomes are usually rather compact, and in *E. coli* many genes overlap; we encountered 91 instances of target codons within overlapping gene sequences. Open reading frames overlapped in either a tail-to-tail (3', 3') or a head-to-tail (5', 3') fashion. If target codons could be replaced without altering the amino acid identity of the overlapping gene, we left the overlap architecture unchanged; this was the case in 12 instances. If substitutions in one open reading frame (ORF) affected the protein sequence encoded in the overlapping ORF, we refactored the overlap architectures according to the following rules:

1. For genes overlapping 3', 3', we duplicated the overlap region and then recoded each ORF individually (**Figure 3.2b**). There were 33 instances of this.
2. For genes overlapping 5' to 3', we duplicated the overlap region plus the 20 nucleotides upstream of the overlap, and provided an in-frame stop codon to terminate translation from the start codon in the overlap region. (**Figure 3.2c**). There were 58 instances of this.

The recoding design relied on the existing annotation of protein-coding genes in our reference genome. In order to minimise the risk of overlooking poorly characterised translated polypeptides, we re-classified 12 pseudogenes as CDS, and recoded them as above. We also encountered three annotated short CDS (*htgA*, *ybbV* and *yzfA*) for which there is no evidence of protein translation[231,232], and which were completely embedded within other CDS that are known to be translated. These three cases were ignored in the design because recoding these was challenging without severely disrupting either the sequence or the local architecture of well-characterised CDS, which may have negative consequences on fitness.

The result of the design process was a genome sequence 3,978,937 bp long, and in which all 18,218 instances of the target codons were substituted by their corresponding synonyms (**Figure 3.2d**).

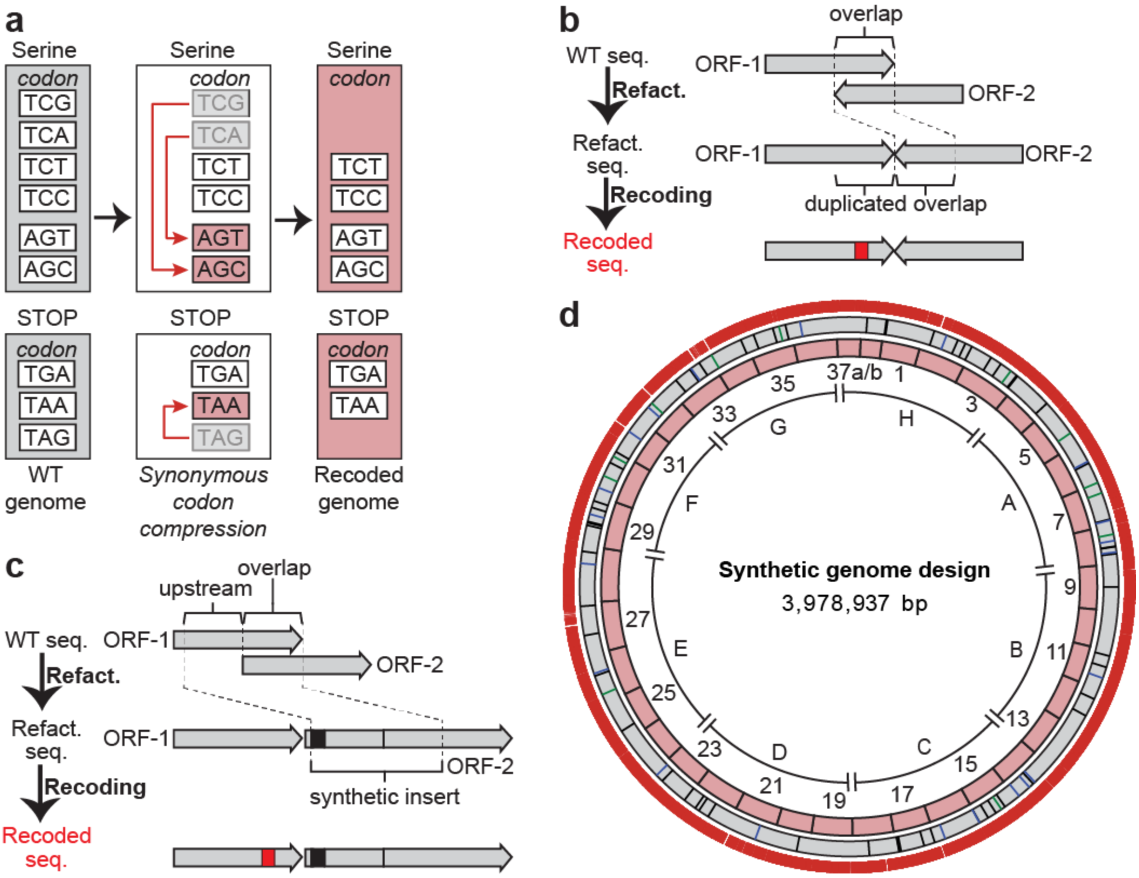


Figure 3.2 - Design of a recoded genome. a) Schematic of the defined recoding scheme used to design a synthetic *E. coli* genome. **b)** Strategy for resolving tail-to-tail overlaps. The refactoring in these cases consists of duplicating the overlap region between ORFs, which can then be recoded independently. The introduction of a recoding event in ORF-1 is shown by a red bar. **c)** Strategy for resolving head-to-tail overlaps. The overlap region is duplicated, together with the 20 bp upstream of the overlap; these form a 'synthetic insert'. In order to prevent translation from the ORF-2 RBS still encoded within ORF-1, an in-frame stop codon is introduced upstream of the synthetic insert (shown as a black bar). The refactored ORFs can then be recoded independently. The red bar in ORF-1 indicates a recoding event. **d)** Schematic of the synthetic genome design. The inner labels A-H correspond to ~ 500 kbp Sections. The pink circle shows the position of the Fragments, labelled 1-37b. The grey circle shows the position of the overlap events, where green lines indicate overlap resolution through silent mutation (12 instances), blue lines indicate tail-to-tail overlaps resolved as in b) (21 instances) and black lines indicate head-to-tail overlaps resolved as in c) (58 instances). In the outermost ring, each codon substitution is indicated by a red line.

3.2.2 Synthesis of recoded sections.

We employed a retro-synthetic approach to genome synthesis. Starting from the designed genome, we divided the sequence into 8 sections of approximately 500 kbp (labelled A-H, **Figure 3.2d**, **Figure 3.3a**). We envisioned that we would synthesise each of these sections independently in parallel strains, and then assemble them into a single genome by conjugation (see 3.2.4). We further disconnected each section into fragments of 97-136 kbp, a size we knew is suitable for performing REXER (Fig **Figure 3.3b**). The boundaries between fragments were placed between non-essential genes, taking care not to disturb any biological features. These boundaries correspond to the 'landing sites' defined in Chapter 2 (**Figure 2.2b, Figure 3.2d**). Each of the fragments was further divided into 9-14 smaller stretches of 6-10 kbp, each overlapping adjacent stretches by 50-80 bp (**Figure 3.3c**). We obtained clonal sequence-verified stretches from a DNA synthesis vendor in plasmid form.

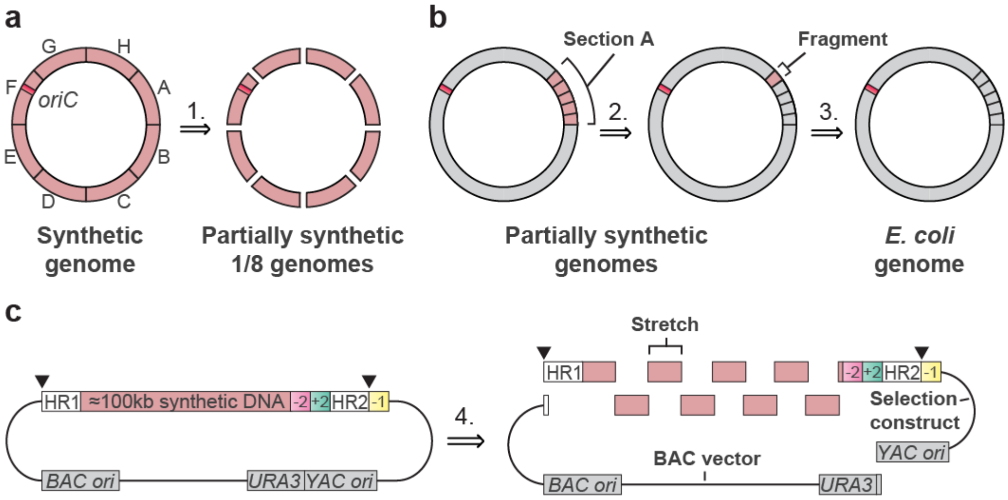


Figure 3.3 - Retrosynthesis of a recoded genome. Overall retrosynthetic approach for gradually building a synthetic genome (on the far left) starting from a wild-type *E. coli* genome precursor (far right). **a)** In Step 1, the synthetic genome is broken down into 8 sections A-H (**Figure 3.2d**), to be synthesised independently. The resulting 1/8th synthetic genomes would be precursors to the assembly of a fully synthetic genome. The position of the origin of replication (*oriC*) is indicated. **b)** In steps 2 and 3, sections are disconnected into ~100 kbp fragments. These are implemented into a wild-type genome to form sections, using REXER and GENESIS. **c)** In order to carry out REXER, fragments need to be contained within REXER BACs. Step 4 shows how BAC designs are broken down into their precursors. In the assembled BACs, the synthetic DNA is coupled to a double selection cassette (here -2/+2), and flanked at both ends by Cas9 cut sites. HR1 and HR2 are the homology regions that mediate recombination during REXER. An additional negative marker (here -1) downstream of HR2 ensures loss of the BAC during REXER.

REXER BAC Assembly

We assembled BACs for REXER, containing synthetic recoded fragments. Each BAC assembly was formed of three main components (**Figure 3.3c**):

- **Synthetic DNA stretches.** These were released from their harbouring plasmids by restriction digestion. The ~80 bp at the 5' end of the first stretch constitute HR1, and contain a Cas9 cut site. During REXER, HR1 mediates recombination between the synthetic and wild-type DNA at the beginning of the relevant fragment.
- **A BAC vector.** This piece contains the BAC replication and segregation components, and a *URA3* marker for selection in yeast. This BAC vector is amplified by PCR with primer overhangs than confer homology to its adjacent components.
- **A REXER selection construct.** This piece contains a double selection marker, flanked at its 5' side with homology to the last synthetic stretch, and at its 3' end by ~80 bp of homology to the genome (HR2) plus a Cas9 cut site. During REXER, HR2 mediates recombination between the synthetic and wild-type DNA at the end of the relevant fragment. The selection construct also contains an additional negative marker and a YAC origin of replication.

BAC assemblies were designed such that the YAC origin of replication and the *URA3* marker came from separate pieces, in order to minimise any background derived from re-circularisation of the BAC/YAC vector. In each assembly, we co-transformed these components into *S. cerevisiae* spheroplasts, and subsequently recovered them in uracil-deficient selective medium. Resulting colonies were screened by colony PCR at stretch-stretch and stretch-BAC vector junctions (**Figure 3.4a**). **Figure 3.4** shows representative examples of the junction genotypes of successful and unsuccessful yeast colonies after assembly of fragments 2, 12 and 18. After identifying BACs that genotyped correctly at all the relevant junctions, we extracted total DNA from the corresponding yeast clones and delivered the BACs into the target *E. coli* cells by electroporation. Once in *E. coli*, the BACs were purified and sequence-verified by next-generation sequencing.

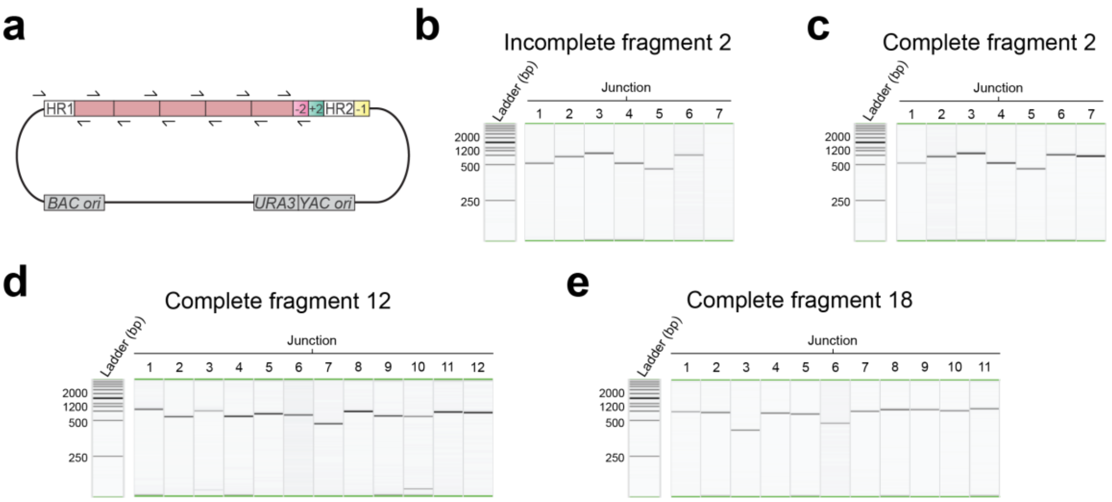


Figure 3.4 - Genotyping assembly of BACs. a) Schematic of an assembled BAC. The pink bar represents a synthetic fragment, where the black lines indicate the junctions between the synthetic stretches. At all or a subset of the junctions, PCR is performed with pairs of primers (black arrows) that bind 200-300 bp away from the junction. **b)** Colony PCR products of a representative incomplete fragment 2 BAC in yeast. PCR products were separated by capillary gel electrophoresis in a QIAxcel Advanced. Products of the expected size were observable for junctions 1 to 6, but not for 7. This may result from failure of the PCR, rather than an absent junction, but we only carried forward clones that were positive at all tested junctions. **c)** Colony PCR products of a representative complete fragment 2 BAC in yeast. 7/7 junctions yielded products of the expected size. **d)** Colony PCR products of a representative complete fragment 12 BAC in yeast, screened at 12 junctions. **e)** Colony PCR products of a representative complete fragment 18 BAC in yeast, screened at 11 junctions.

In order to iterate REXER (GENESIS), BACs harbouring adjacent fragments would need to contain alternating positive-negative cassettes in the selection constructs. Following the original REXER set up, we initially created two classes of BACs with either:

Class I. an *rpsL-kanR* (-1/+1) double selection cassette, followed by a copy of *sacB* (-2) in the BAC backbone

Class II. a *sacB-cat* (-2/+2) double selection cassette, followed by a copy of wild-type *rpsL* (-1) in the BAC backbone

kanR and *cat* confer resistance to kanamycin and chloramphenicol respectively, and *rpsL* and *sacB* confer sensitivity to streptomycin and sucrose.

Note: I performed assembly of BACs for fragments 2, 3, 11, 12, and 18. The remaining assemblies were performed by all other authors in the original publication.

REXER of recoded fragments

As BACs became available from assembly, we started performing REXER experiments on the individual fragments, with a view to promptly identifying any deleterious sequences that we may encounter when building synthetic sections. These were performed in derivatives of *E. coli* MDS42, whose genomes we prepared by introducing double selection cassettes at the beginning of the relevant fragment via lambda-red recombination as in Chapter 2 (**Figure 3.5a**). Lambda-red recombinations and REXERs throughout this Chapter were performed using a 'helper plasmid', which encodes Cas9 and the lambda-red recombination machinery. For each REXER, we can consider the 5' landing site of the fragment to be locus^0, and the 3' to be locus^1. After REXER, we identified colonies that had undergone replacement by assessing the loss and gain of double selection cassettes at locus^0 and locus^1, through colony PCR at these loci (**Figure 3.5b**) and stamping the cells in selective media (**Figure 3.5c**). Clones whose genotypes and selective growth phenotypes were consistent with replacement were carried forward to analysis by next-generation sequencing, which served to identify clones that were completely recoded across a given fragment.

In REXER, the negative marker in the BAC backbone drives loss of the BAC after Cas9 cleavage and recombination. When testing individual fragments, we encountered multiple cases of class I BACs with loss of function in the backbone copy of *sacB*. This precluded selection for loss of the BAC during REXER, and resulted in the presence of a bacterial lawn in the selective agar plates. Sequencing of pre-REXER BACs revealed that this was often due to mutations in the promoter region of *sacB*, accumulated during BAC assembly and delivery, and may reflect a selective pressure for alleviating the toxic effects of *sacB* even in the absence of sucrose. We decided to replace the *sacB* negative marker in class I BACs with a *pheS*-hygR* cassette, which allowed negative selection for loss of the BAC during REXER by providing 4-chloro-phenylalanine in the medium. We directly assembled 8 class I BACs containing *pheS*-hygR* in exchange for *sacB*. In the BACs we had already assembled but that contained a non-functional copy of *sacB*, we replaced it by *pheS*-hygR* via lambda-red mediated recombination; this was the case for BACs harbouring fragments 2, 6, 8, 10, 12 and 16 (I performed replacement for 12 and 16). In class II BACs, where *sacB* is coupled to *cat*, we did not encounter any instances of loss of function, and we did not alter their assembly. We were ultimately able to assemble BACs for all the designed fragments (see **Appendix B.1** for efficiencies) except for fragment 37, where we failed to identify a yeast clone harbouring a complete BAC. This fragment was split into two separate assemblies, each of around 50 kbp (37a/b in **Figure 3.2d**), and these were successful.

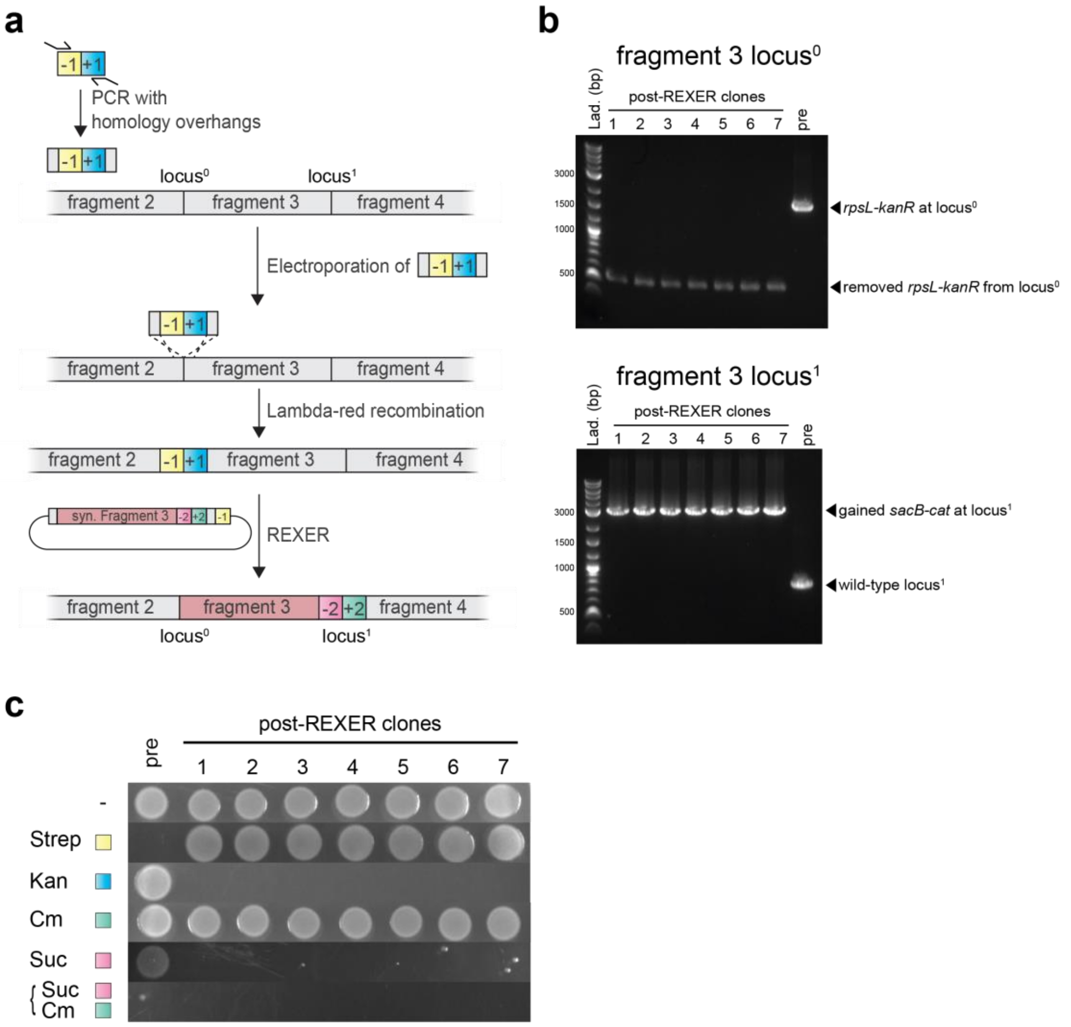


Figure 3.5 - REXER of synthetic BACs. a) Schematic of genome preparation for REXER of individual fragments, using fragment 3 as an example. A double selection cassette (-1/+1, yellow and blue boxes) is integrated at locus^0 via lambda-red recombination. The resulting genome is ready for REXER with a BAC that contains a complementary -2/+2 cassette (green and magenta boxes). REXER results in integration of -2/+2 at locus^1, which marks the 3' boundary of the fragment. **b)** Locus-specific PCR after REXER of fragment 3. The size of the products at each locus is consistent with the presence or absence of a double selection cassette. 'Pre' is the same strain before REXER. **c)** Stamping of fragment 3 REXER clones in selective LB medium. The pre-REXER cells (pre, contain the fragment 3 BAC) are sensitive to 100 µg/mL streptomycin (Strep) and resistant to 50 µg/mL kanamycin (Kan) due to the *rpsL-kanR* cassette at locus^0. REXER removes this cassette and renders cells resistant to

(continued from previous page) streptomycin, but sensitive to kanamycin (post-REXER clones 1-7). Pre-REXER cells are resistant to 7.5% sucrose (Suc) alone due to loss of the BAC, but die in the presence of both sucrose and 20 µg/mL chloramphenicol (Cm). Post-REXER cells have *sacB-cat* integrated at locus^1, and consequently die in the presence of sucrose alone. Both pre- and post-REXER cells are resistant to chloramphenicol. Wild-type DNA is indicated by grey bars, synthetic DNA is indicated by pink bars.

In parallel to REXER of individual fragments, we began GENESIS to create 500 kbp synthetic sections. We started REXER-mediated replacements at the beginning of all sections A-H except for B; because the construction of A was completed before the others, GENESIS of section B was performed directly on top of A. Consistently with Chapter 2, during creation of the synthetic sections we found that the success of electroporation of BACs harbouring synthetic fragments was highly variable. We sometimes failed to electroporate BACs into partially synthetic strains from both yeast and *E. coli* DNA extracts, but succeeded in wild-type MDS42. In these cases, we delivered the BAC to the target cells by conjugation, using the successfully transformed *E. coli* strains as donors. For this, we integrated an origin of transfer (*oriT*) into the BAC backbone by lambda-red mediated recombination in donor cells. We then transformed a version of the F' plasmid, prepared by Julius Fredens, in which the nick region of *oriT* is truncated. Since the *oriT* in this F' plasmid is non-functional, it could mediate the transfer of the *oriT*-containing BACs, but not its own. As a result, recipient cells received only the BAC of interest, and were immediately ready for REXER.

Once the technical issues regarding *sacB* selection and BAC delivery were resolved, we could proceed with creation of the synthetic sections. After each round of REXER, we sequenced the genomes of multiple clones that had the expected genotypes at locus^0 and locus^1, and the correct growth phenotypes on selective media. NGS identified clones that contained the desired synthetic sequence across the entire fragment, and these were carried forward to the next replacement step. When multiple clones had undergone complete replacement, we prioritised the ones that contained no or few mutations in the rest of the genome. GENESIS proceeded smoothly for sections A, C, D, E, F and G. A registry of the rates of full replacement at each fragment is provided in **Appendix B.2 and B.3**. In most REXER steps, we employed linear spacer arrays, as described in Chapter 2. Some REXERs however were only efficient when we used plasmid-encoded spacers. We did not encounter deleterious sequences in 35 out of the 38 fragments that formed the synthetic genome. However, REXER identified 4 lethal synthetic sequences localized to fragments 37a, 1 (section H) and 9 (section B). The process of identifying and fixing these sequences, as well as the construction of the synthetic sections that contained them, is discussed in the next section.

Note: I performed REXER of fragments 3, 9, 11, 12, 13 and 18 individually to verify their viability, and of fragments 3, 11, 12 and 13 as part of building sections. REXERs for the remaining fragments were performed either individually or as part of sections by all other authors in the original publication.

3.2.3 Construction of recoded sections containing deleterious sequences

Identifying and fixing design flaws

The genomes of the clones resulting from a REXER experiment can be either recoded or not recoded at each of the target codons in the fragment of interest. Compiling the genome sequences of multiple clones allows for constructing what we term a 'recoding landscape', which provides the frequency of recoding at each target codon across the sample of clones being analysed. When a synthetic sequence is deleterious, there is a selective pressure for excluding that sequence from the genome, and the maintenance of the wild-type sequence; this presumably occurs via recombination between the synthetic and wild-type DNA while they both co-exist within the cytoplasm. Crossovers can happen stochastically and independently of sequence viability. However, when the recoding landscape shows regions with a recoding frequency of 0 (i.e. none of the clones accepted the synthetic sequence at that position), it constitutes a good indication that the region is problematic.

We performed REXER of fragment 9 by itself, and shortly after on top of section A (fragments 4-8, generated by Julius Fredens). A recoding landscape was built using the genome sequences from both experiments, and allowed for identifying a relatively large region of approximately 26 kbp where none of the clones were recoded (**Figure 3.6a**). In order to narrow down the area that contained the problematic sequence, we opted for a synthetic complementation approach.

We returned to the strain where section A was recoded and the BAC containing fragment 9 was present, but REXER had not yet been performed. We then attempted to delete 6 stretches of 6-10 kbp from the host genome (named G2-G7, **Figure 3.6a**), corresponding to the problematic region. Deletion attempts were performed in separate strains, by integration of *pheS*-hygR* markers that contained homology arms to the flanking sequences of each stretch via lambda-red recombination. For 5 of these deletions the episomal, recoded fragment 9 could complement the deletion in the host genome. However, the remaining deletion in G4 failed to work, indicating that the deleterious synthetic design resided within this stretch.

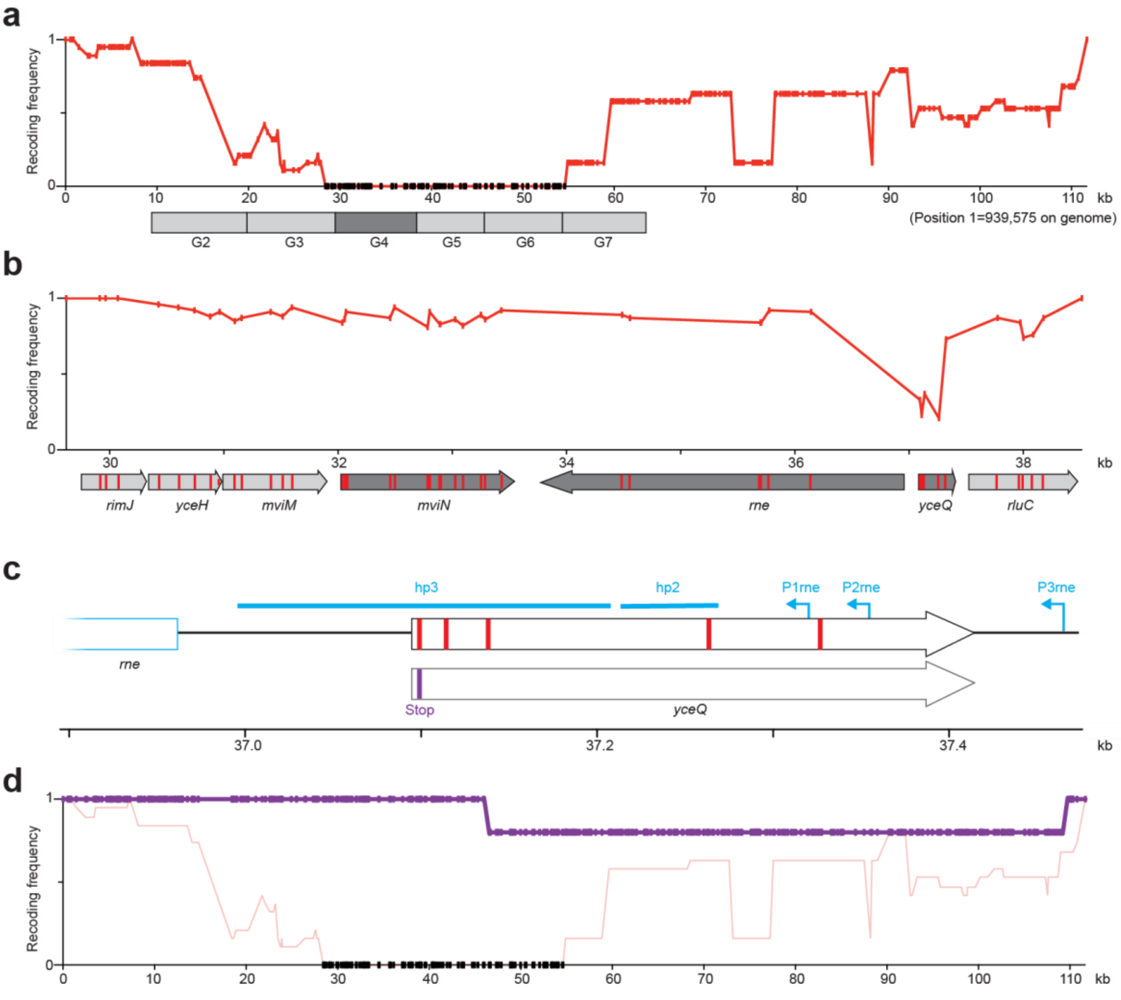


Figure 3.6 - Identifying and troubleshooting a deleterious sequence in fragment 9. a) Recoding landscape of fragment 9 REXER, compiled from 19 clones. Positions recalcitrant to recoding are indicated by black dots. The relative position of synthetic stretches G2 to G7 is shown. An episomal recoded copy of G4 (dark grey) did not support deletion in the genome. **b)** Recoding landscape of REXER with stretch G4, compiled from 10 clones. The x-axis indicates the relative position within fragment 9. Annotated genes are depicted below the landscape in light grey (non-essential) or dark grey (essential). Red lines indicate target

(continued from previous page) codons. The recoding landscape shows two minima at *yceQ*, which contains 5 target codons. **c)** *yceQ* overlaps regulatory elements in *rne* (shown in blue). The 3 leftmost codon substitutions (red lines) overlap regulatory hairpin hp3, another codon overlaps regulatory hairpin hp2. The rightmost codon sits within the -10 element of transcriptional start site P1rne. In order to address this, a fragment 9 BAC was constructed where a stop codon is introduced at the beginning of the *yceQ* gene (purple line), and the rest of the *yceQ* remains wild-type (bottom). **d)** Recoding landscape of REXER with a fixed fragment 9 (purple line), compiled from five clones. The pre-fix recoding landscape from a) is included in grey and black, for contrast.

Next, a REXER BAC was constructed that contained only G4, and it was used to perform REXER in the strain that contained a recoded section A. We hypothesised that, since the region of homology between the BAC and the genome had been reduced to less than 10 kbp, crossovers between the synthetic and wild-type DNA would pinpoint the deleterious sequence at higher resolution. Indeed, the recoding landscape resulting from this experiment highlighted a minimum at hypothetical gene *yceQ*, which contained 5 target codons (**Figure 3.6b**). While no single codon was rejected in all strains, we never observed simultaneous recoding of all 5 codons within the same strain.

While there is, to our knowledge, no evidence of translation of *yceQ*, it was originally classified as essential in the Keio collection[233]. However, a more recent transposon mutagenesis study found that disruption of *yceQ* was tolerated in one orientation, but rejected in the other[234]. *yceQ* sits within the promoter region of *rne* (**Figure 3.6c, Figure 3.7a),** which encodes the essential ribonuclease E (RNAse E). RNAse E plays a crucial role in mRNA decay and rRNA/tRNA processing, and its expression at appropriate levels is critical to cell viability[235]. Its homeostasis results from a complex regulatory network that involves multiple promoters[236] and a repressive feedback loop in which RNAse E accelerates the rate of decay of its own *rne* transcript, preventing overproduction[237]. The 5'UTR region of *rne* (which overlaps target codons in *yceQ*, **Figure 3.6c, Figure 3.7b**) is crucial for this post-transcriptional regulation and some of the structural domains in the UTR have been shown to recruit RNAse E to the transcript, promoting its degradation[238,239]. Taken together, this suggested that the recoding scheme was non-viable at this locus because it was disrupting the mRNA hairpins responsible for RNAse E homeostasis (hp2 and hp3 in **Figure 3.6c, Figure 3.7b**). With views of maintaining native *rne* regulation, we created a new BAC for fragment 9; a stop codon was introduced at the beginning of the *yceQ* ORF to prevent translation of its hypothetical product, and rest of the gene sequence was left as wild-type (**Figure 3.6c**). Following REXER with this new BAC, 4 out of 5 clones accepted the entirety of synthetic fragment 9 (**Figure 3.6d**). These results confirmed that we had successfully identified and fixed the deleterious sequence in fragment 9.

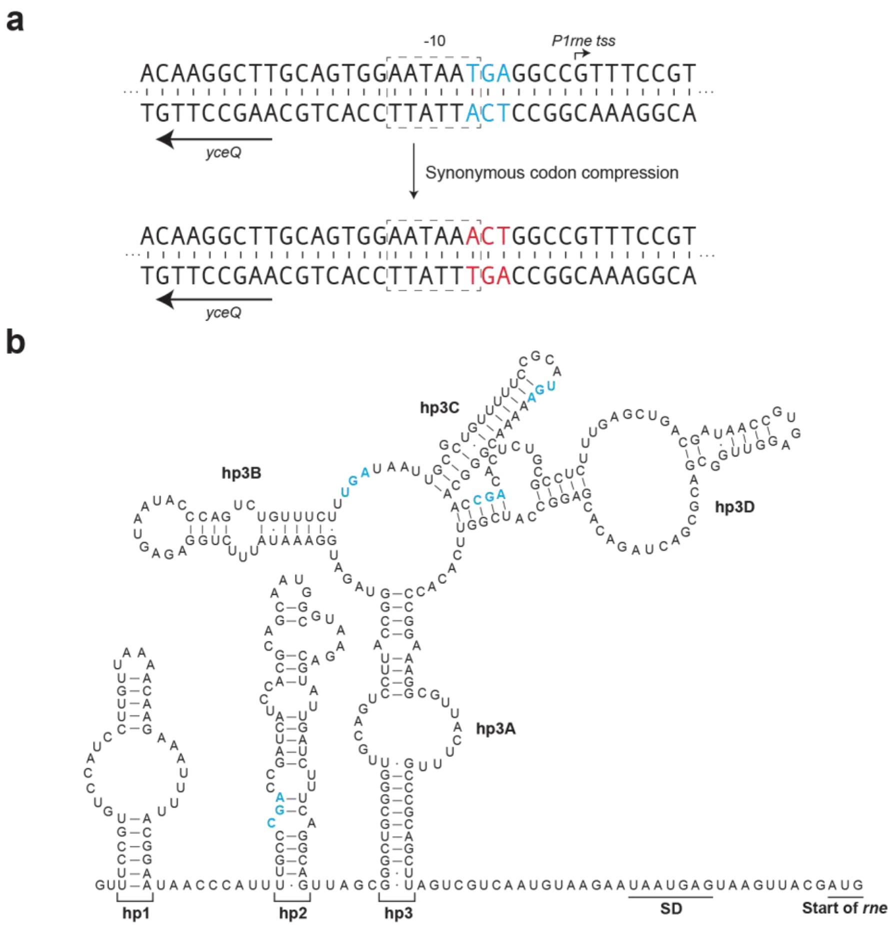

a
-10
P1rne tss
ACAAGGCTTGCAGTGGAATAATGAGGCCGTTTCCGT
TGTTCCGAACGTCACCTTATTACTCCGGCAAAGGCA
yceQ
Synonymous codon compression
ACAAGGCTTGCAGTGGAATAAACTGGCCGTTTCCGT
TGTTCCGAACGTCACCTTATTTGACCGGCAAAGGCA
yceQ
b
hp3C
hp3B
hp3D
hp3A
GUU-AUAACCCAUUU·GUUAGCG·UAGUCGUCAAUGUAAGAAUAAUGAGUAAGUUACGAUG
hp1
hp2
hp3
SD
Start of rne

Figure 3.7 – Substitutions in hypothetical gene *yceQ* overlap regulatory elements of *rne*. a) Top, wild-type sequence context of major *rne* transcriptional start site (*tss*) P1rne, which sits within the *yceQ* CDS (in the opposite orientation). A target codon TCA is indicated in blue. This codon sits within the annotated -10 element of this *tss*. Bottom, the same sequence after synonymous codon compression. The TCA codon is replaced by AGT (in red) and disrupts the -10 element. **b)** Schematic of the wild-type 5'UTR secondary structure in the *rne* transcript[238]. The figure shows regulatory hairpins 1 to 3 (hp1, hp2, hp3), the Shine-Dalgarno sequence (SD) and the translation initiation codon AUG (Start of *rne*). Target codons for replacement are indicated in blue, and locate to hp2 and hp3. These hairpins are recognised by RNAse E, which mediates their degradation as part of a critical self-regulatory loop. Substitution of these codons by their synonyms may cause disruption of these structures, preventing RNAse E binding.

Note: I performed REXER of fragment 9 to narrow down the problematic sequence, and contributed to data analysis and literature search. All other experiments for troubleshooting fragment 9 were performed by Julius Fredens. The troubleshooting of fragments 37a and 1 outlined below was performed by Julius Fredens and Louise Funke, and is described here for completeness.

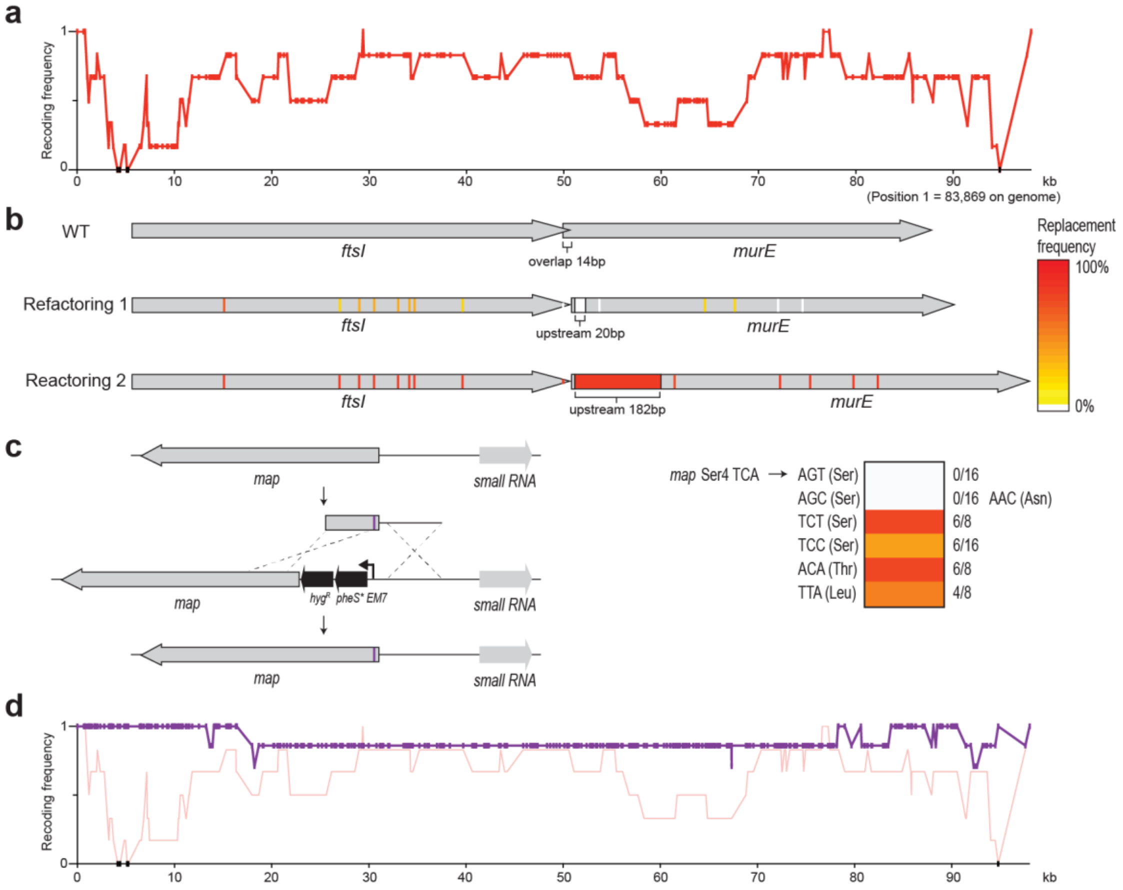


Figure 3.8 - Identifying and troubleshooting a deleterious sequence in fragment 1. a) Recoding landscape of fragment 1 REXER, compiled from 6 clones. The landscape shows two minima at i) the *ftsI-murE* overlap (left minimum) and ii) *map* Ser4 TCA (right minimum). Target codons that were never recoded are shown in black. **b)** Schematic of the local architecture at the *ftsI-murE* overlap. In the wild-type genome (WT, top), the two CDS overlap by 14 bp. In the Syn61 design (Refactoring 1), the two CDS are separated using the rules described in the text. Refactoring 2 (bottom) shows the architecture employed by Wang *et al.* (ref[185]), where the 182 bp upstream of the overlap are duplicated as opposed to only 20 bp. Target codons are highlighted within the genes as coloured lines; the colours indicate the frequency with which they were recoded in their respective experiments. **c)** Schematic of the oligo-mediated fixing of a recalcitrant codon in *map*. The chart on the right indicates the fraction of colonies that removed the *pheS*-hygR* cassette by replacement with alternative codons. Experiments with oligos bearing AGC or AGT did not generate positive colonies; one colony in the AGC experiment contained a replacement at *map* position 4, but it was AAC

(continued from previous page) (Asn). **d)** A new BAC for fragment 1 was constructed, containing Refactoring 2 at the *ftsI-murE* overlap and a TCT codon at *map* position 4. The graph shows a recoding landscape with this fixed fragment 1 BAC, compiled from 7 clones. The recoding landscape from **a)** is reproduced for comparison.

We encountered 2 other problematic fragments; 37a and 1. In fragment 1, the recoding landscape revealed 2 minima (**Figure 3.8a**). One corresponded to a single codon (TCA, Ser4) in the essential *map* gene, which encodes methionine aminopeptidase, and the other mapped to an overlap region between genes *ftsI* and *murE*, which had been refactored. In previous work, our group was able to recode this overlapping region, but the length of the duplication was 182 bp as opposed to the 20 bp used in this work (**Figure 3.8b**). Hence, we hypothesised that extending the overlap region to 182 bp should rescue the viability of the sequence at this site. In order to find replacements for the recalcitrant codon in *map* we performed two steps of lambda-red recombination (**Figure 3.8c**). In the first, we introduced a *pheS*-hygR* cassette immediately upstream of *map,* selecting for hygromycin. In the second, we removed the marker by counter-selection with 4-chloro-phenylalanine, providing ssDNA oligos that contained alternative replacements for the recalcitrant *map* codon. Sanger sequencing of the resulting clones revealed that TCT, TCC, ACA and TTA were tolerated, but AGT and AGC were not. The BAC for fragment 1 was fixed so that it incorporated i) an extended overlap between *ftsI* and *murE* and ii) TCA->TCT substitution in Ser4 of *map*. These amendments rescued the viability of the sequence, and REXER with this new BAC produced recoding across the entirety of fragment 1 in 2 out of 7 clones (**Figure 3.8d**).

In fragment 37a, a recoding landscape derived from 6 clones located the problematic region to a 6.5 kbp section (**Figure 3.9a**). Analysis of additional clones by successive rounds of Sanger sequencing, targeted at the essential genes within this section, progressively narrowed down the problem to a single TCA codon in Ser70 of hypothetical gene *yaaY* (**Figure 3.9b**), which is immediately upstream of the essential *ribF* (a bifunctional riboflavin kinase/FMN adenylyltransferase). A 2-step recombination approach similar to the one used for troubleshooting *map* revealed that TCC, TCG, TCT and AGC were suitable replacements, but AGT was not (**Figure 3.9c**). Construction of a synthetic BAC for fragment 37a with a TCA->AGC mutation in

Ser70 of *yaaY,* followed by REXER, achieved synthetic replacement across 37a in 1 out of 7 clones (**Figure 3.9d**).

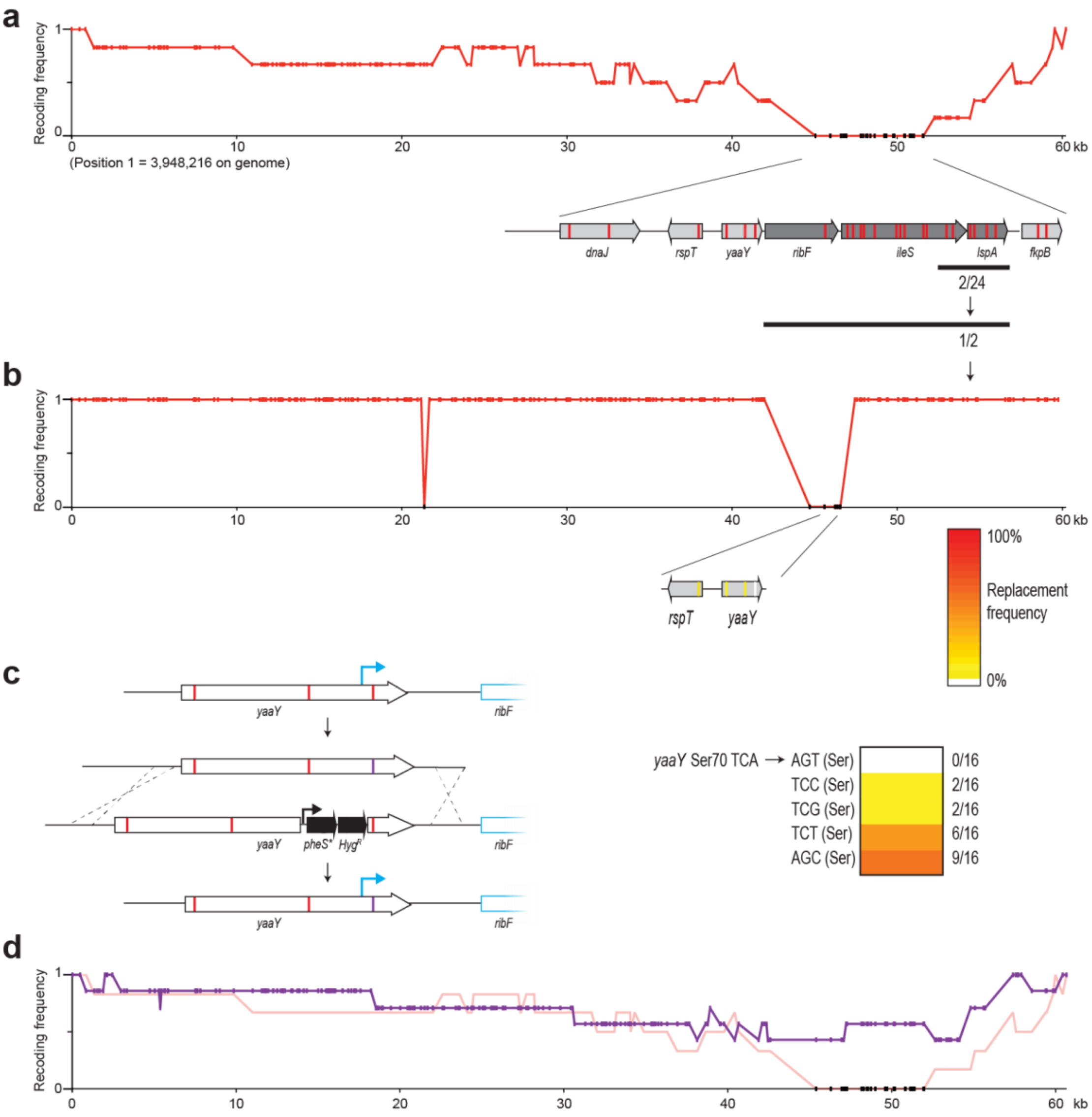


Figure 3.9 - Identifying and troubleshooting a deleterious sequence in fragment 37a. a) Recoding landscape of REXER of fragment 37a, compiled from 6 clones. The genes that comprise the region are shown in light (non-essential) or dark grey (essential). Additional clones resulting from this experiment were analysed, by PCR of the regions containing essential genes (black bars) followed by Sanger sequencing. Sequencing of 24 clones across the region indicated by the top black bar identified 2 clones where all six codons were recoded. These two clones were further analysed across the region indicated by the bottom black bar, and 1 of the 2 was completely recoded across this region. **b)** The genome of the clone identified in a) was sequenced by next-generation sequencing. The graph shows the recoding landscape of this clone, where several codons in *rspT* and *yaaY* were not recoded. This region was analysed by Sanger sequencing in 33 additional clones from the 37a REXER. The codons in *rspT* and *yaaY* are shown as coloured lines, where the frequency of recoding of each is indicated by the colour according to the scale on the right. Only Ser70 of non-

(continued from previous page) essential *yaaY* was never recoded in this analysis. **c)** Schematic of an oligo-mediated approach to testing alternative substitutions in *yaaY* Ser70, analogous to **Figure 3.8c.** Here, the *pheS*-hygR* cassette is introduced within the *yaaY* CDS. The blue arrow indicates the transcriptional start site of essential *ribF*, which is upstream of *yaaY* Ser70. **d)** Recoding landscape of a fixed fragment 37a REXER where Ser70 *yaaY* is AGC, compiled from 7 clones (blue line). The recoding landscape from a) is reproduced for comparison. 1 clone out of 7 was completely recoded.

Completing section B

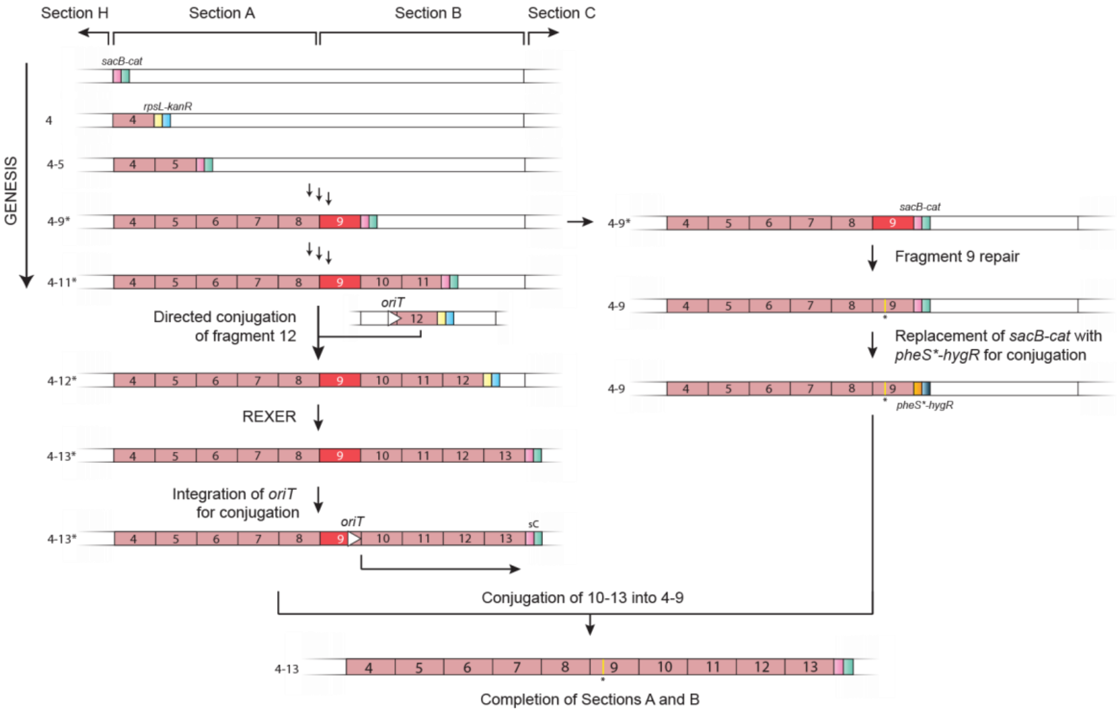

Section H
Section A
Section B
Section C
GENESIS
sacB-cat
rpsL-kanR
4
4-5
4-9*
4-11*
Directed conjugation of fragment 12
oriT
4-12*
REXER
4-13*
Integration of oriT for conjugation
oriT
sC
4-13*
4-9*
sacB-cat
Fragment 9 repair
4-9
Replacement of sacB-cat with pheS*-hygR for conjugation
4-9
pheS*-hygR
Conjugation of 10-13 into 4-9
4-13
Completion of Sections A and B

Figure 3.10 - Overall strategy for construction of Sections A-B. GENESIS of fragments 4 to 8 proceeded smoothly. REXER of fragment 9 (red) identified a deleterious sequence in *yceQ*, which was troubleshot separately as in **Figure 3.6** (right). The fix is indicated by a yellow line and an asterisk. After troubleshooting of fragment 9, the *sacB-cat* at its 3' exerted suboptimal counter-selection with sucrose, and was replaced with a *pheS*-hygR* cassette in preparation for conjugation. In parallel to fragment 9 troubleshooting, Section B continued to be constructed in a strain bearing an incomplete fragment 9 (4-9*); the details are discussed in the text. The 4-13* strain served as a conjugation donor for the transfer of synthetic fragments 10-13 into the complete 4-9 strain, yielding a strain with synthetic Sections A and B. The white arrows represent *oriT* and indicate the beginning and directionality of conjugal transfer.

A schematic of the completion of Section B on top of A is provided in **Figure 3.10.** While the troubleshooting of fragment 9 was taking place, we continued building on section B, performing REXER of fragment 10 on the strain that contained a synthetic section A (4-8) plus an incomplete fragment 9. This strain is referred to as 4-9*;

throughout the following text, the asterisk denotes the incompleteness of fragment 9. REXER of fragments 10 and 11 was not problematic. However, we could not successfully perform REXER of fragment 12 as a continuation of 4-11*, as we failed to obtain any colonies. This was unexpected, as REXER of fragment 12 in a wild-type strain had previously been successful (**Figure 3.11a, b**).

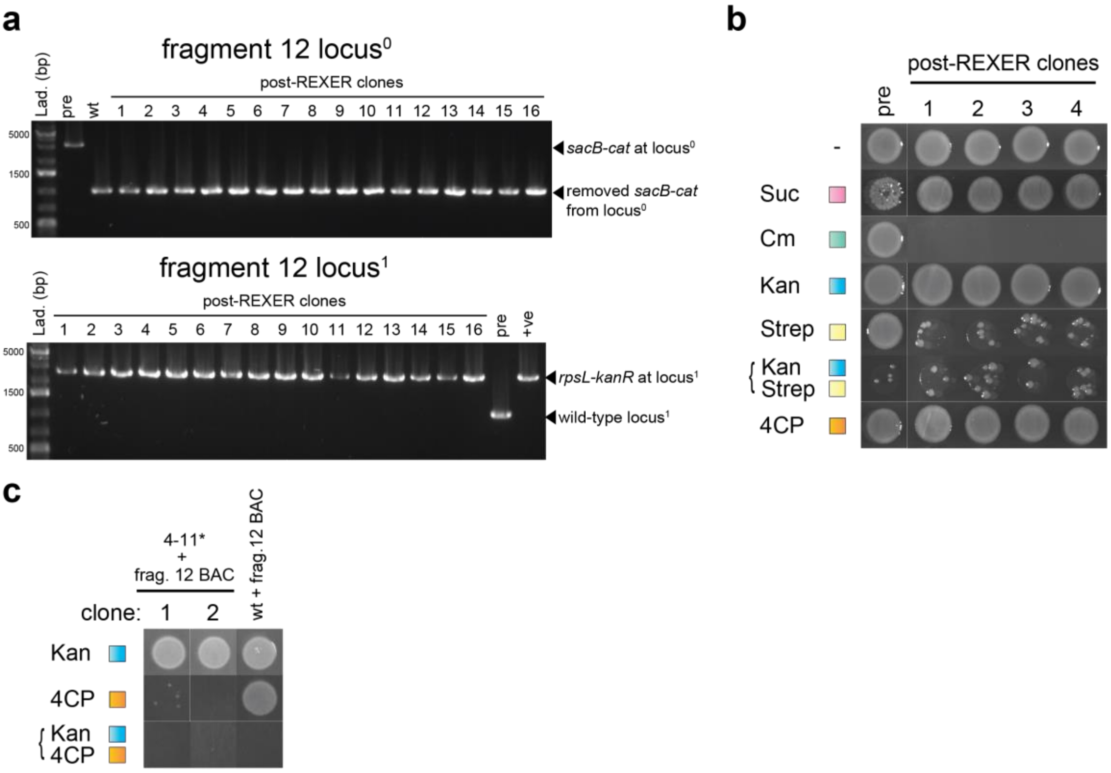


Figure 3.11 - REXER integrates fragment 12 individually, but not on top of 4-11. a) Locus-specific colony PCR of clones after individual REXER of fragment 12, in an otherwise wild-type MDS42 background. The pre-REXER strain (pre) was prepared by integration of *sacB-cat* at locus^0. Post-REXER clones (1-16) exhibit products consistent with loss of *sacB-cat* at locus^0 and gain of *rpsL-kanR* at locus^1. A strain of wild-type MDS42 with an *rpsL-kanR* integrated at locus^1 serves as a positive control (+ve). **b)** Stamping cells after individual REXER of fragment 12 in selective media. The growth phenotypes of pre- and post-conjugation (1-16) clones are consistent with successful REXER. A number of escapees are observed in 100 µg/mL streptomycin (Strep) selection. Concentrations of selective agents are 20 µg/mL chloramphenicol (Cm), 50 µg/mL kanamycin (Kan), 7.5% sucrose (Suc), and 2.5 mM 4-chloro-phenylalanine (4CP). **c)** Pre-REXER wild-type MDS42 cells containing a fragment 12 BAC ('wt + frag. 12 BAC') survive on kanamycin, as well as 4-chloro-phenylalanine by losing the BAC. Loss of the BAC is prevented by selecting with both kanamycin and 4-chloro-phenylalanine, which results in cell death. However, in pre-REXER 4-11* cells containing fragment 12, cell death is observed upon selection with 4-chloro-phenylalanine alone, suggesting that the cells are incapable of losing the BAC. This may

(continued from previous page) explain why REXER of fragment 12 in 4-11* yielded no colonies, or exclusively false positives. Concentrations of Kan and 4CP are as in b).

We analysed the phenotype of pre-REXER 4-11* or wild-type strains containing fragment 12 BAC. Following replacement of *sacB*, this BAC contained an *rpsL-kanR* double selection cassette flanking the synthetic DNA at locus[1], plus a *pheS*-hygR* cassette in the backbone. From this marker combination, in a pre-REXER strain we would expect growth in 4-chloro-phenylalanine alone, by virtue of losing the BAC, but cell death when selecting for both 4-chloro-phenylalanine and kanamycin, which impedes survival by BAC loss. This was the case for the wild-type strain. However we found that 4-11* cells containing fragment 12 BAC died upon selection with 4-chloro-phenylalanine alone (**Figure 3.11c**). The genome sequence of this strain did not reveal any mutations in the BAC or chromosomal copies of *pheS*, which may have helped explain the dominant negative phenotype under 4-chloro-phenylalanine selection. The reasons underlying lethality of 4-chloro-phenylalanine in this particular strain remain unclear.

As an alternative to REXER, in order to implement synthetic fragment 12 on top of 4-11* we adapted a method for transfer of genomic DNA between the chromosomes of distinct strains[99], illustrated in **Figure 3.12**. We also used to this method to convergently assemble multiple synthetic sections into a single synthetic strain (see section 3.4.2). Partially synthetic strains generated by REXER and GENESIS naturally contain double selection cassettes at the end of the synthetic region, and these provide convenient selectable handles to direct conjugation between them, creating larger synthetic regions in a clockwise manner.

Simple manipulations to these strains with lambda-red recombination allow them to act as donors or recipients during conjugation. Generally, we prepared donor strains by integrating an *oriT* at the beginning of the target region. An additional 3 kbp of recoded homology between donor and recipient strains were provided by either i) adding a recoded tail in the recipient (Approach 1 in **Figure 3.12**) or ii) extending the 5´ recoded region in the donor (Approach 2). Through either approach, 3 kbp of recoding overlap is shared between the two strains, at the 3' of recoding in the recipient and at the 5´ of recoding in the donor. During conjugation, *recBCD* recombination mediates the integration of incoming DNA into the host genome. The selectable markers placed at either side of this 3 kbp window allow for confining recombination events to this region. Additionally, the helper plasmid (tetracycline resistance) is cured by passaging in the donor strain but not in the recipient, which provides an additional selectable marker to distinguish donor and recipient cells.

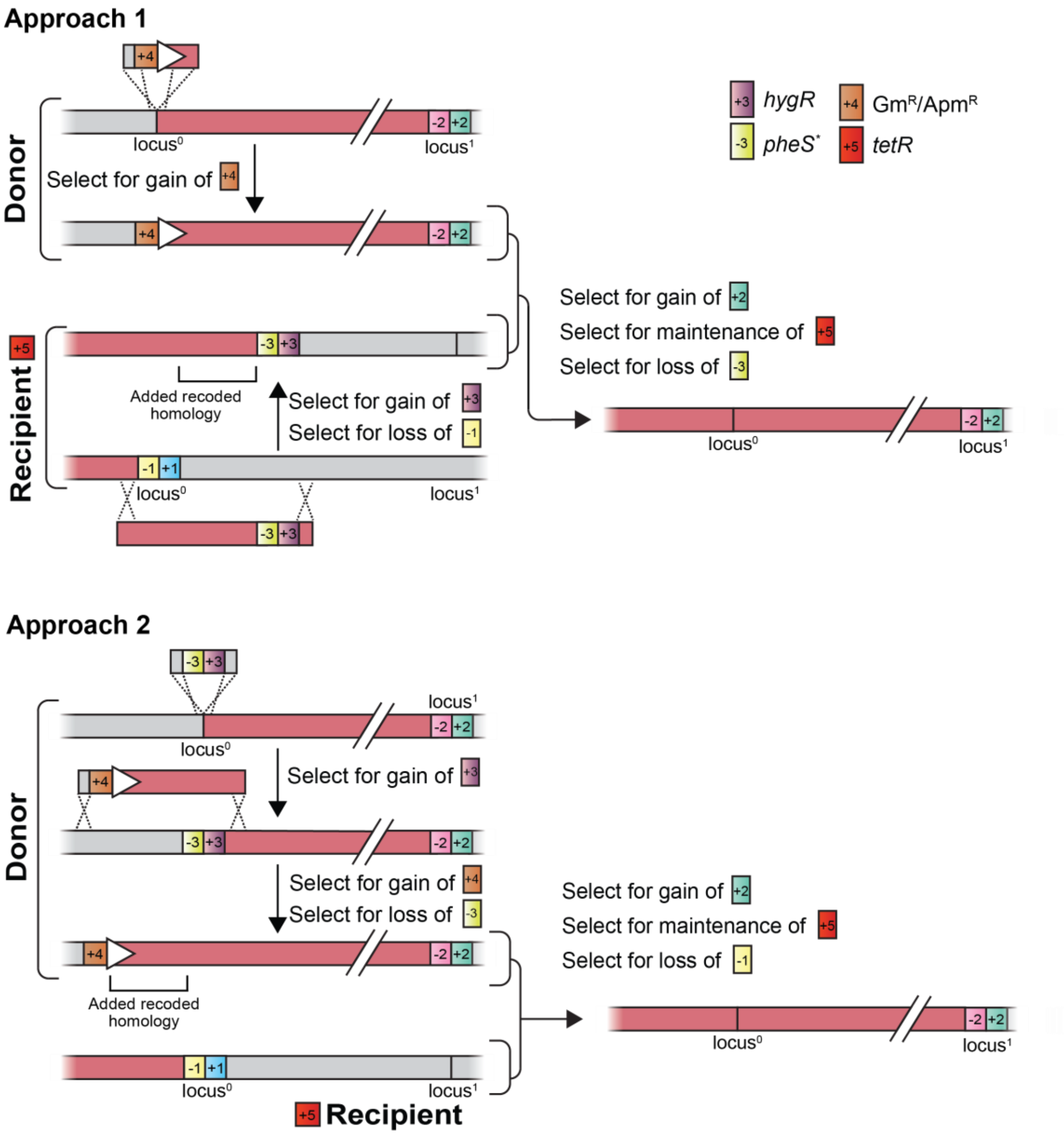


Figure 3.12 - Directed conjugation of synthetic regions. This diagram shows strategies for the transfer of a synthetic region (from locus⁰ to locus¹) from a donor to a recipient strain. The recipient genome contains a helper plasmid (+5) and a double selection cassette from REXER (here -1/+1), whereas the donor contains only a double selection cassette at locus¹ (here -2/+2). In **Approach 1**, a 3 kbp recoded homology tail, coupled to a -3/+3 double selection cassette, is added just after locus⁰ in the recipient genome via lambda-red recombination. An *oriT* (white arrow) coupled to a positive selectable marker (+4, orange) is integrated at locus⁰ in the donor, and marks the beginning of conjugal transfer. Following transformation of the F' plasmid and conjugation, selection for the helper plasmid ensures

(continued from previous page) that only recipient cells are recovered. Additionally, selection for the loss of the negative marker in the recipient and the gain of the positive marker from the donor recovers cells where the region from locus^0 to locus^1 has been replaced. In **Approach 2**, the donor genome is prepared in two steps. First a -3/+3 marker is integrated at locus^0 by lambda-red recombination. A second step introduces 3 kbp of recoded homology upstream of locus^0. Selection for the loss of -3 and the gain of +4 ensure that lambda-red recombination replaces the entire 3 kbp. Conjugation and selection then proceed as in Approach 1.

Because recombination events occur randomly, we cannot precisely define the 3' boundary for transfer. However, the selectable markers that flank the synthetic sequences in donor and recipient strains allow us to select for cells in which transfer and recombination events cover, minimally, our target region (**Figure 3.12**). Like in REXER, additional crossovers between the synthetic and wild-type sequence may occur, and post-conjugation genomes need to be verified by next-generation sequencing.

In order to perform conjugation of fragment 12 on top of 4-11*, we leveraged a strain which contained an individually integrated fragment 12 (flanked by at its 3' end by an *rpsL-kanR* cassette, **Figure 3.11**) as a conjugation donor. The 4-11* recipient contained a *sacB-cat* at the 3' end of fragment 11 (**Figure 3.13a**). We prepared the donor strain for conjugation by integrating an *oriT-apmR* cassette at the beginning of fragment 12 (+3 kbp of recoded homology) by lambda-red mediated recombination, curation of the helper plasmid and transformation of a non-transferrable F' plasmid. We then performed conjugation, and spread the cells in medium supplemented with:

i) sucrose (-2) in order to ensure recombination prior to the beginning of fragment 12.
ii) kanamycin (+1) to ensure transfer up to at least the end of fragment 12.
iii) tetracycline (helper plasmid) to ensure we recovered recipient cells.

We analysed 16 of the resulting colonies by colony PCR at the fragment boundaries and growth assay on selective media. We found that all 16 had undergone recombination at both fragment 12 boundaries (**Figure 3.13b**), and displayed growth phenotypes consistent with their expected combination of selectable markers (**Figure 3.13c**). 4 of these clones progressed to analysis by next-generation sequencing, out of which 2 were completely recoded across fragment 12. We termed one of the clones 4-12*, and carried it forward to the next step of GENESIS. 12 was the only fragment, out of 38, that was not implemented by REXER.

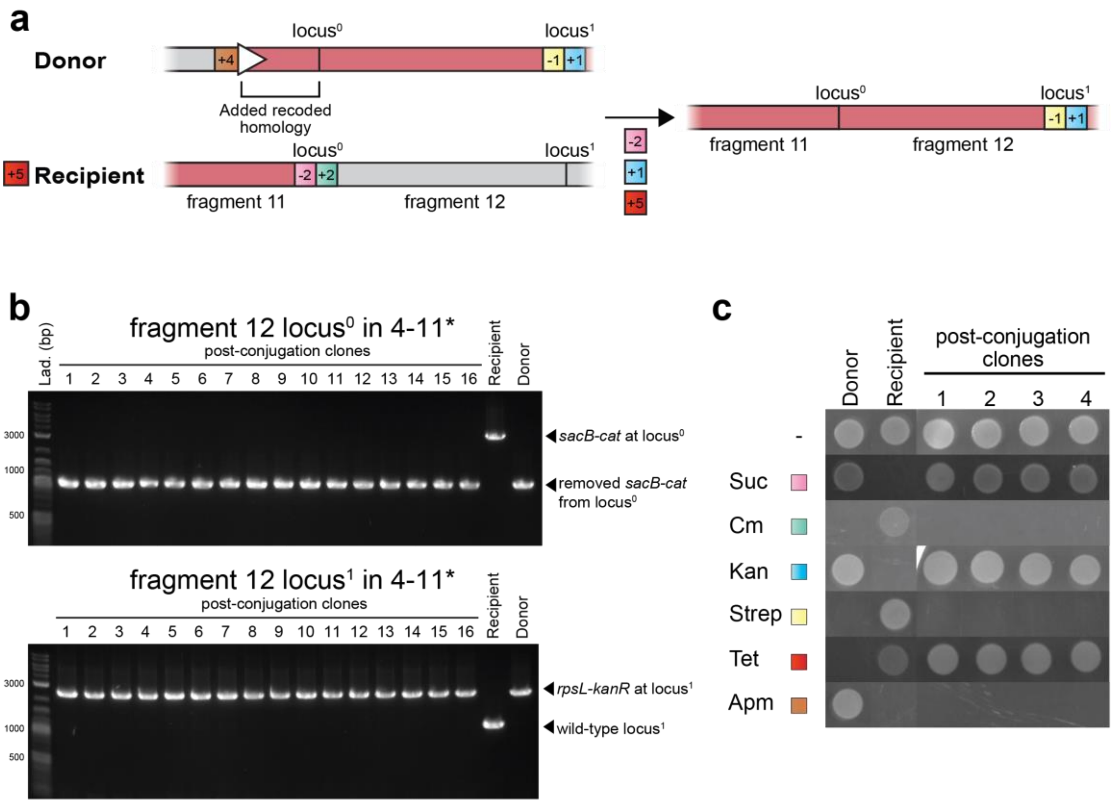


Figure 3.13 - Conjugation of fragment 12 into 4-11*. a) Schematic of donor and recipient cells, prepared as in Approach 1 (**Figure 3.12**). The *rpsL-kanR* cassette is depicted as -1/+1 (yellow and blue), and the *sacB-cat* is depicted as -2/+2 (pink and green). +5 represents tetracycline resistance encoded by the helper plasmid. **b)** Locus-specific PCR in post-conjugation clones (1-16) shows loss of the *sacB-cat* cassette at locus^0, and the gain of *rpsL-kanR* at locus^1. Donor and Recipient only controls serve to indicate the expected bands resulting from the presence and absence of the cassettes at each locus. **c)** Stampings of post-conjugation clones in selective media are consistent with the recovery of recipient cells that have lost *sacB-cat* marker and gained *rpsL-kanR* cassette.

The 4-12* strain was a substrate for the REXER-mediated replacement of fragment 13, but this step also presented a technical idiosyncrasy. While the BAC for fragment 13 was transformed and propagated intact in *E. coli* DH10b (as confirmed by NGS), upon transformation into MDS42 derivatives containing the helper plasmid we consistently observed a ~85 kbp truncation in the BAC, corresponding to the region between genes *rstB* and *ydiM*, and were unable to isolate any colonies containing an intact synthetic fragment 13. We hypothesised that this may be due to the leaky expression of the recombination machinery of the helper plasmid - unwanted recombination activity may have led to the deletion of this region, or the re-circularisation of linear BAC fragments in the electroporation preparation. The 4-12* synthetic strain was passaged to remove the helper plasmid. In these cells, transformation of the fragment 13 BAC yielded intact BACs. We could then re-transform the helper plasmid and perform REXER, which was successful in generating the synthetic 4-13* strain. Fragment 13 was also the only instance in which electroporation of a BAC yielded a consistently truncated product.

Once we had i) a strain containing a full, fixed synthetic fragment 9 in synthetic strain 4-9, and ii) a contiguous stretch of synthetic fragments 10-13 in the 4-13* strain, we combined both into a complete 4-13 synthetic strain by conjugation. Similarly to the strategy described in **Figure 3.12**, an *oriT* was placed 3 kbp upstream of the beginning of fragment 10 in the 4-13* strain, which acted as a conjugation donor. Conjugal transfer of the recoded 10-13 into the synthetic 4-9 strain was then induced, generating a clone with synthetic sections A and B.

Note: REXER of fragment 10 was performed jointly with Andres Gonzalez. Julius Fredens performed conjugation of 10-13 into 4-9 to generate A-B. I performed all other experiments to generate section B.

Completing section H

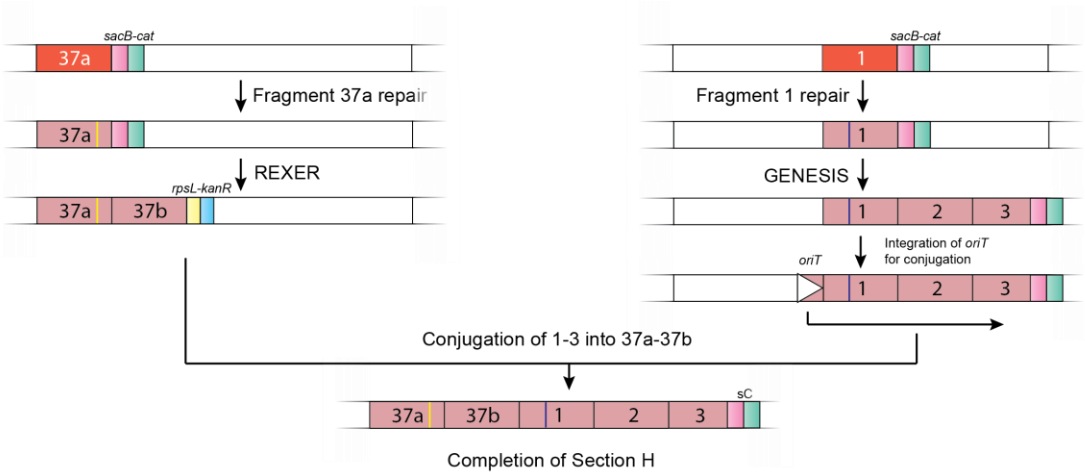


Figure 3.14 - Construction of section H. The individual REXER of fragments 1 and 37a identified deleterious sequences, and these were fixed as described in **Figure 3.8** and **Figure 3.9**. Two parallel sets of GENESIS led to the creation of 37ab and 1-3 synthetic sections. The strain bearing synthetic fragments 1-3 was prepared as a conjugation donor as in Approach

(continued from previous page) 2 of **Figure 3.12**, and conjugation into the 37ab strain yielded a strain with a completely recoded Section H.

Section H contained 2 fragments with deleterious sequences; 37a and 1. After their troubleshooting was complete (see above), we proceeded with two parallel sets of REXERs; 37b on top of fragment 37a, and 2 and 3 on top of fragment 1 (**Figure 3.14**). The resulting synthetic 37a/b and 1-3 strains were amenable for directed conjugation to generate a strain with a complete section H.

Note: Louise Funke performed GENESIS for 37a and 37b. Julius Fredens performed GENESIS for fragments 1 and 2. I performed REXER of fragment 3 and conjugation of 1-3 into 37a/b to generate a complete section H.

3.2.4 Assembling a recoded genome

At this point, we had 7 partially synthetic strains, corresponding to sections A+B, C, D, E, F, G and H. We employed the directed conjugation approach outlined in **Figure 3.12** to combine these sections and assemble them into a single synthetic genome. We performed conjugations in a clockwise manner, as illustrated in **Figure 3.15**. Since most sections were flanked at their 3' ends by REXER double selection cassettes, these provided convenient markers to define the extent of conjugal DNA transfer. Similarly to conjugation of BACs, throughout genome assembly we used a non-transferrable F' plasmid. As a result, there was no need to cure F' after each conjugation step, and the resulting strains could immediately act as recipients. This accelerated the transition from one conjugation to another.

As described above, our conjugation strategy relied on providing 3 kbp of recoded homology overlap between donor and recipient cells, which constitutes a window for *recBCD*-mediated recombination. Often, 3 kbp of recoded homology were sufficient to achieve conjugation. However, in some cases conjugation was only successful when the homology overlap was extended. For example, when conjugating section G on top of sections A-F, donor and recipient shared entire section F as a recoded homology (~400 kbp). These extended homologies were introduced into donor or recipient strains by REXER or conjugation (**Figure 3.15**).

The final conjugation necessary to achieve a fully recoded genome was section H on top of A-G, where the strain bearing section H was the donor and the A-G strain was the recipient. Because crossovers between the donor and recipient genome during conjugation are stochastic, and the donor strain was only recoded in section H, we were concerned that transfer of the donor genome beyond section H would 'erase' the recoding in section A in the recipient strain. Therefore, we set out to add a 'buffer' homology tail after section H in the donor strain (**Figure 3.16**).

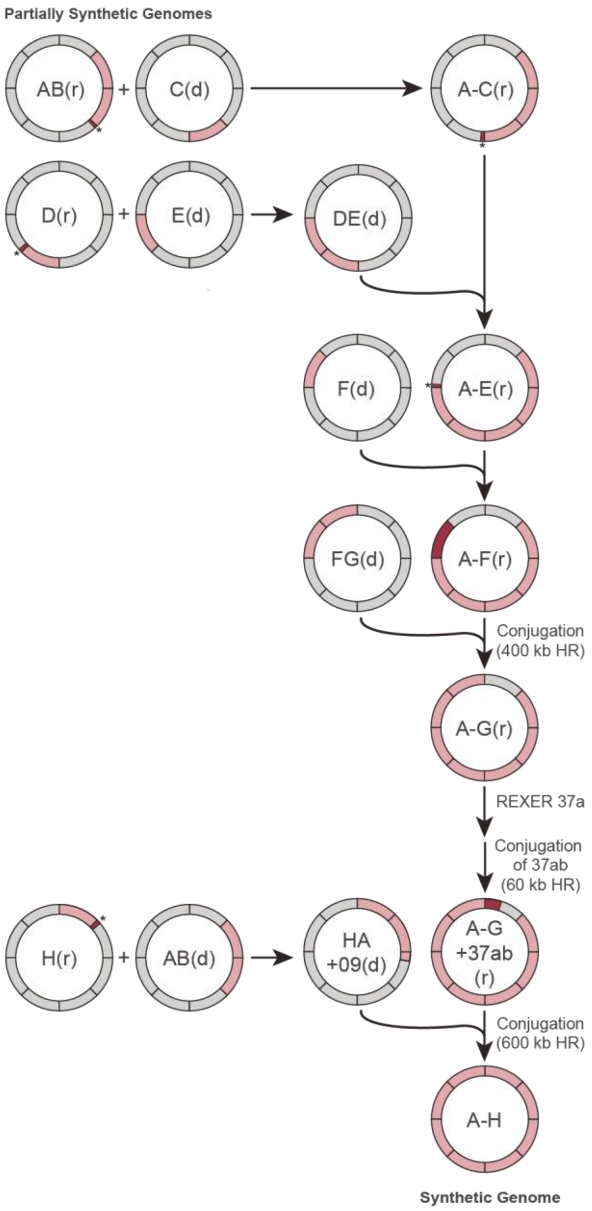


Figure 3.15 - General conjugation strategy. Directed conjugation was used to combine synthetic DNA (pink) from partially recoded genomes in a clockwise manner. The names of each strain reflect the sections recoded in their genomes; (d) denotes donors and (r) denotes recipients. Recoded homology overlaps are shown in dark pink. Short homology overlaps (~3 kbp) are denoted by an asterisk, but these could occasionally be as long as ~400 kbp, as described in the text. The details of the final conjugation steps are illustrated in **Figure 3.16**.

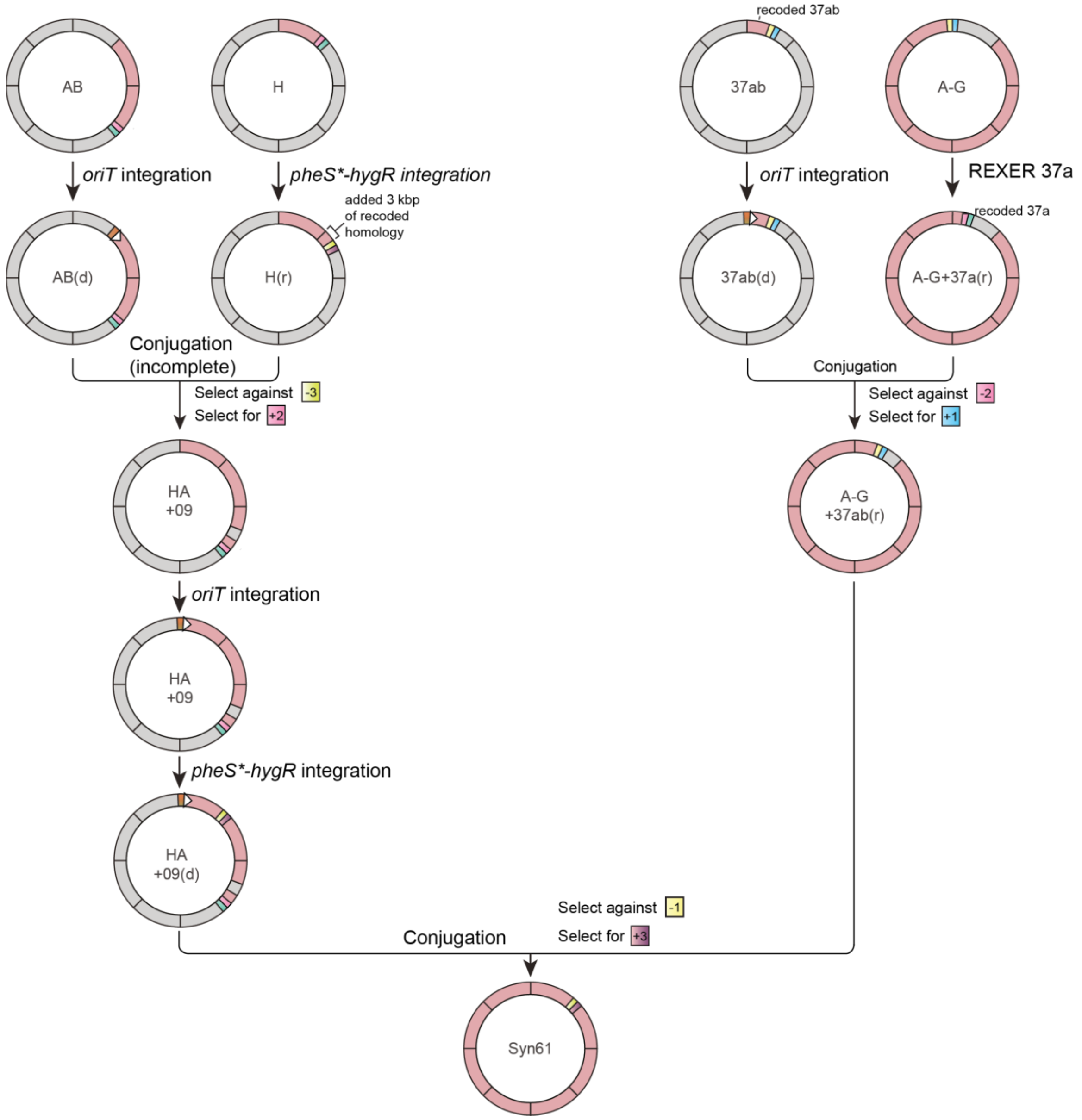


Figure 3.16 - Preparation for final conjugation. Schematic of the preparation of donor and recipient cells for combining synthetic sections A-G and H. A strain bearing synthetic sections AB was prepared as a donor by the lambda-red mediated integration of an *oriT* at the beginning of A, and conjugation was directed into a recipient strain bearing a synthetic section H with a 3' recoded tail. The resulting clones were not completely recoded across HAB but one (HA+09) was carried forward as a donor in the final step, and prepared by integration of *oriT* and a *pheS*-hygR* cassette to yield HA+09(d). In parallel, a recoded tail in the strain bearing synthetic sections A-G was added via REXER of fragment 37a followed by conjugation of fragment 37b, yielding AG+37ab(r) which was immediately ready to act as a recipient in conjugation.

We attempted to conjugate sections A and B on top of section H; this would, in principle, provide a recoded window of 1 Mb for recombination to occur. Among the resulting clones, next-generation sequencing identified one that was recoded from section H to approximately fragment 9 in section B (termed 'HA+09(d)'). This provided a buffer homology of 700 kbp, which we deemed sufficient to attempt the final conjugation. We prepared HA+09(d) as a conjugation donor by i) integrating an *oriT* at the beginning of section H and ii) a *pheS*-hygR* cassette at the end of section H. In the A-G recipient, we extended the recoded homology tail at the end of section G, by integrating synthetic fragments 37a/b via REXER and conjugation (**Figure 3.16**). Although conjugation did not replace the entire fragment 37b, the resulting strain contained 60 kbp of homology with the beginning of section H; we termed this strain 'AH+37ab(r)'.

Conjugation between HA+09(d) and AH+37ab(r) yielded a fully recoded strain, which we termed Syn61. We sequenced the genome of Syn61 (GenBank accession CP040347.1) and identified only 8 non-programmed mutations with respect to the original design. 4 of these arose during preparation of REXER BACs, whereas we assigned the other 4 to the recoding process. None of these mutations affected recoding at target codons (**Appendix B.4**)

Note: I performed all manipulations for the generation of AB(d) and H(r). Conjugation to generate HA+09 failed twice in my hands, but was performed successfully by Louise Funke. I then performed manipulations to generate HA+09(d). A-G+37ab(r) was prepared by Louise Funke and Julius Fredens. I performed conjugation to generate Syn61 jointly with Wesley Robertson, Julius Fredens and Louise Funke. All other conjugations of sections were performed by Louise Funke, Julius Fredens and Wesley Robertson.

3.2.5 Emerging properties of Syn61

We set out to characterise the properties of Syn61. While handling the increasingly synthetic strains, we noticed a progressive decrease in fitness. We measured the growth rate of Syn61 in a plate reader and determined that it grew 1.6x slower than its progenitor, MDS42, in LB medium with glucose at 37 °C. The fitness defect with respect to the MDS42 parent was accentuated at 25 °C (2.5x), but surprisingly seemed to improve at 42 °C (1.3x) (**Figure 3.17a**). In Syn61, almost 18,000 TCG and TCA codons have been replaced by AGT and AGC, which are decoded by the seryl-tRNA *serV*. Since the decoding burden of *serV* was increased by 64% with respect to wild-type MDS42, we hypothesised that the reduced growth rate may be partially rescued by providing additional copies of *serV*. However, the provision of a plasmid containing *serV* under control of a constitutive *lpp* promoter had no apparent effect in growth rate in LB at 37 °C (**Figure 3.17a**).

Under the microscope, Syn61 cells appeared slightly elongated with respect to MDS42, but the difference was minimal (**Figure 3.17b,c**). Both strains had highly similar proteomes, as judged by comparing relative protein abundances through tandem mass-spectrometry (**Figure 3.17d**).

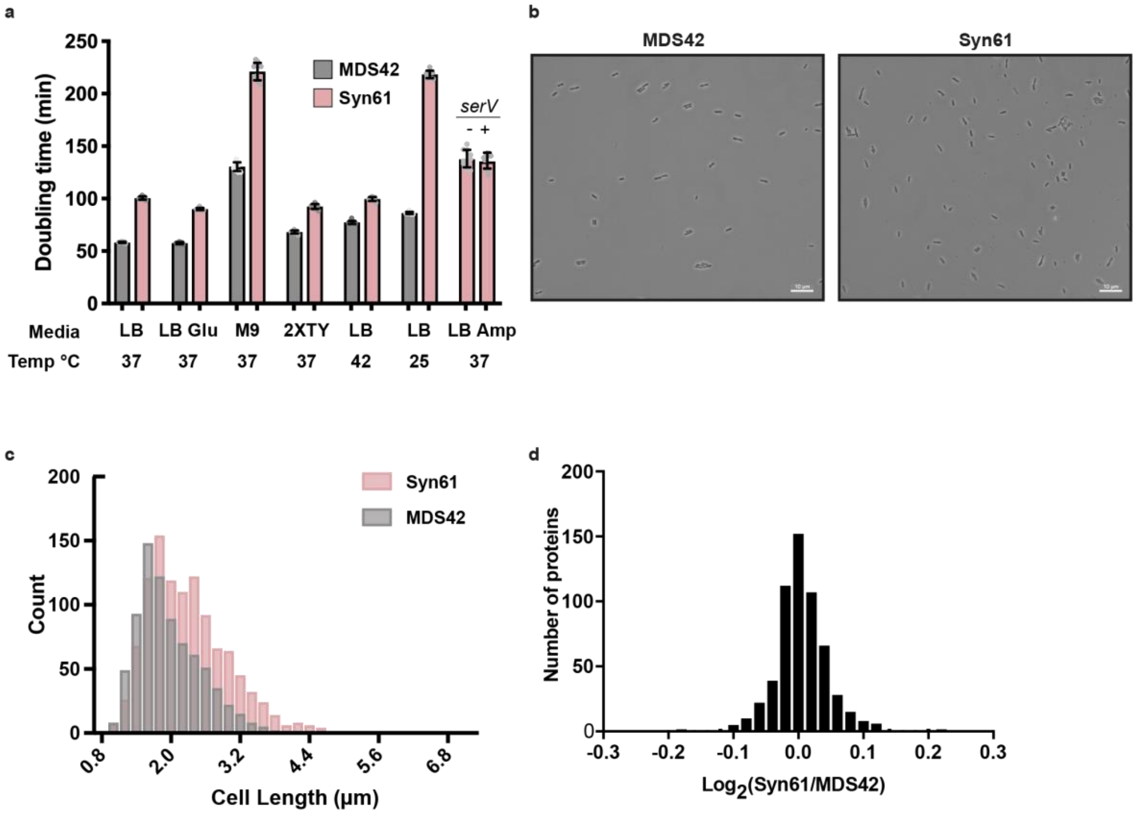


Figure 3.17 - Emerging properties of Syn61. a) Comparison of maximum doubling times for Syn61 and its parent MDS42. The maximum doubling times in LB+2% glucose are 57.6 min and 90.1 min, respectively. LB only, 58.3 min and 100.6 min; M9 minimal medium, 130.5 min and 221.1 min; LB at 42 °C, 77.4 min and 99.7 min; LB at 25 °C, 86.3 min and 218.4 min. Providing extra copies of *serV* did not improve doubling time (138.2 min, +*serV*) compared to the empty vector control (136.2 min, -*serV*). **b)** Representative phase-contrast microscopy images of MDS42 and Syn61. **c)** Histogram of cell lengths, quantified from microscopy images of strains MDS42 and Syn61. The mean cell lengths (±s.d.) for MDS42 and Syn61 were

(continued from previous page) 1.97±0.57 µm and 2.3±0.74 µm, respectively. **d)** Comparison of protein abundances in MDS42 and Syn61, quantified by tandem mass-spectrometry. Out of the 1,084 proteins quantified across samples, the maximum variation observed between strains was 1.16-fold.

We next devised an experiment aimed at verifying that the target codons for replacement, TCG and TCA, no longer encoded the proteome. Syn61 and MDS42 cells were transformed with orthogonal pyrrolysyl-tRNA synthetase (PylRS)/tRNA^{Pyl}$_{NNN}$ pairs from *Methanosarcina mazei*, where the anticodon of tRNA^{Pyl} was variable. This pair directs the incorporation of *Nε*-(((2-methylcycloprop-2-en-1-yl) methoxy) carbonyl)-L-lysine (CYPK) into the codons dictated by the anticodon of the tRNA (**Figure 3.18a**). When using PylRS/tRNA^{Pyl}$_{CGA}$, we found that increasing concentrations of CYPK caused toxicity in MDS42 but not in Syn61 (**Figure 3.18b**). This was consistent with the notion that CYPK was being incorporated in TCG codons in MDS42, causing mis-synthesis of the proteome, but not in Syn61 as TCG codons have been removed. PylRS/tRNA^{Pyl}$_{UGA}$ caused mild toxicity in both, but slightly less in Syn61 – this was expected because the UGA anticodon reads 4 codons in MDS42, but only 2 in Syn61 (**Figure 3.18c**). PylRS/tRNA^{Pyl}$_{GCU}$ incorporates CYPK in response to AGC and AGT; consistently with this, it caused greater toxicity in Syn61 (where AGC and AGT instances are increased by 64%) than in MDS42 (**Figure 3.18d**).

If our target codons for replacement are no longer present in the genome of Syn61, then we would expect their decoding elements to be redundant, and we should be able to delete them (**Figure 3.19a**). We attempted each deletion independently by lambda-red mediated recombination, inserting a double selection cassette at the relevant locus. *serU* and *serT* deletions were performed with both *rpsL-kanR* and *pheS*-hygR* cassettes. *serU* is not essential in *E. coli* because its decoding target TCG is also decoded by *serT*, and its deletion was successful (**Figure 3.19b**). *serT* is the only tRNA expected to decode TCA, and is a well-established essential gene in *E. coli*[234]. Notably, we found that we were able to delete *serT* in Syn61, indicating that it is no longer needed for synthesis of the proteome (**Figure 3.19c**). We attempted an analogous *serT* deletion in wild-type MDS42, and did not obtain any colonies. Similarly, an RF-1 knockout is lethal when amber codons are not removed[98,100], or in the absence of compensatory mutations in RF-2[94,95]. In Syn61, RF-1 deletion was successful, consistent with the absence of amber codons (**Figure 3.19d**).

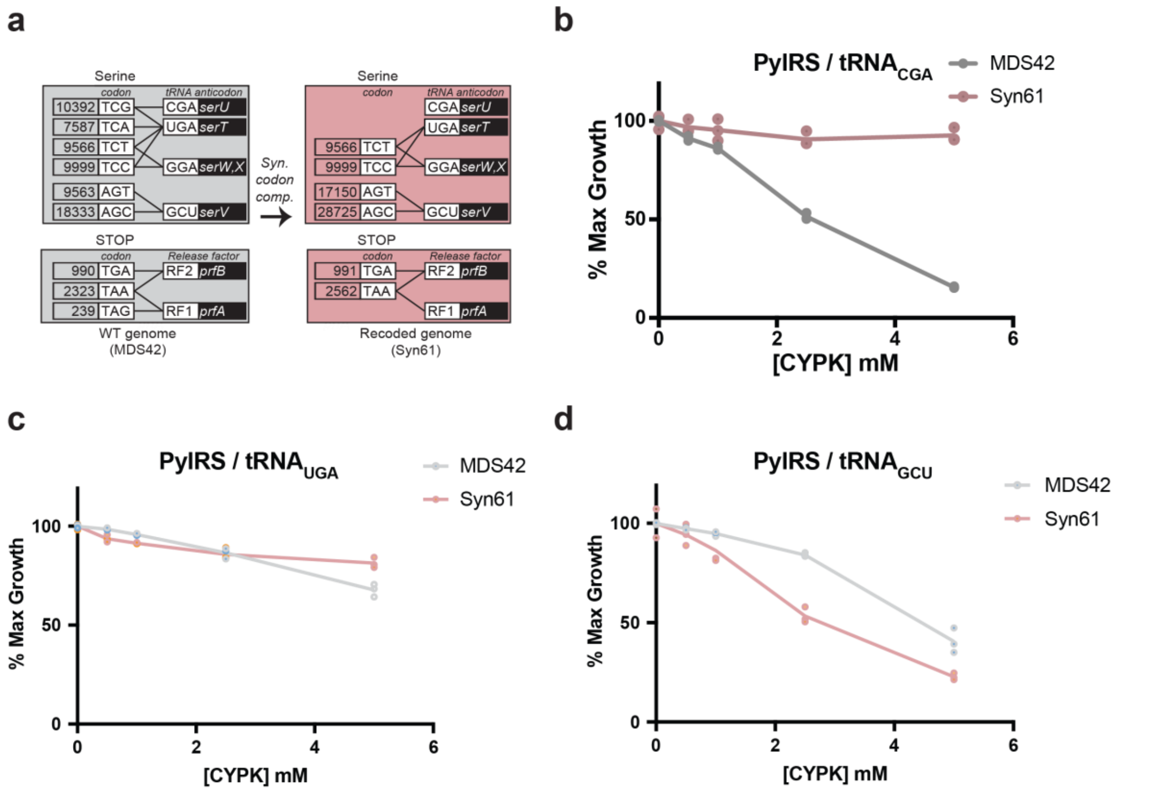


Figure 3.18 - Toxicity of CYPK incorporation into serine codons. a) Schematic of the serine decoding box, before (left, MDS42) and after (right, Syn61) synonymous codon compression. The numbers on the left of each codon indicate annotated instances in the genome. The decoding relationships between seryl-tRNA isoacceptors and codons is shown. **b)** Incorporation of Nε-(((2-methylcycloprop-2-en-1-yl) methoxy) carbonyl)-L-lysine (CYPK) into TCG with PylRS/tRNA$^{Pyl}_{CGA}$ produces dose-dependent toxicity in MDS42, but not in Syn61, as measured by maximum OD600 of the cultures. **c)** Incorporation of CYPK with PylRS/tRNA$^{Pyl}_{UGA}$ (targeting TCG, TCA, TCT and TCC) causes mild but slightly higher toxicity in MDS42 than in Syn61. **d)** Incorporation of CYPK with PylRS/tRNA$^{Pyl}_{GCU}$ (into AGC and AGT) causes greater toxicity in Syn61 than in MDS42. In each case, the relative toxicity observed correlates positively with the number of target codons for PylRS/tRNA$^{Pyl}_{UGA}$ in each strain.

Insertion of a *pheS*-hygR* or *rpsL-kanR* cassette at the loci of *serU* or *serT* decreased fitness, and conferred a clumpy phenotype when grown in culture, which seemed accentuated for *serT*. This phenotype prevented accurate measurement of growth rates by spectrophotometry in a plate reader. It was unclear whether this was a direct consequence of their absence in Syn61, or due to the alteration of the local expression dynamics caused by the presence of a double selection cassette. Indeed, drops in fitness have been previously observed when perturbing non-essential tRNA loci in *E. coli*[115].

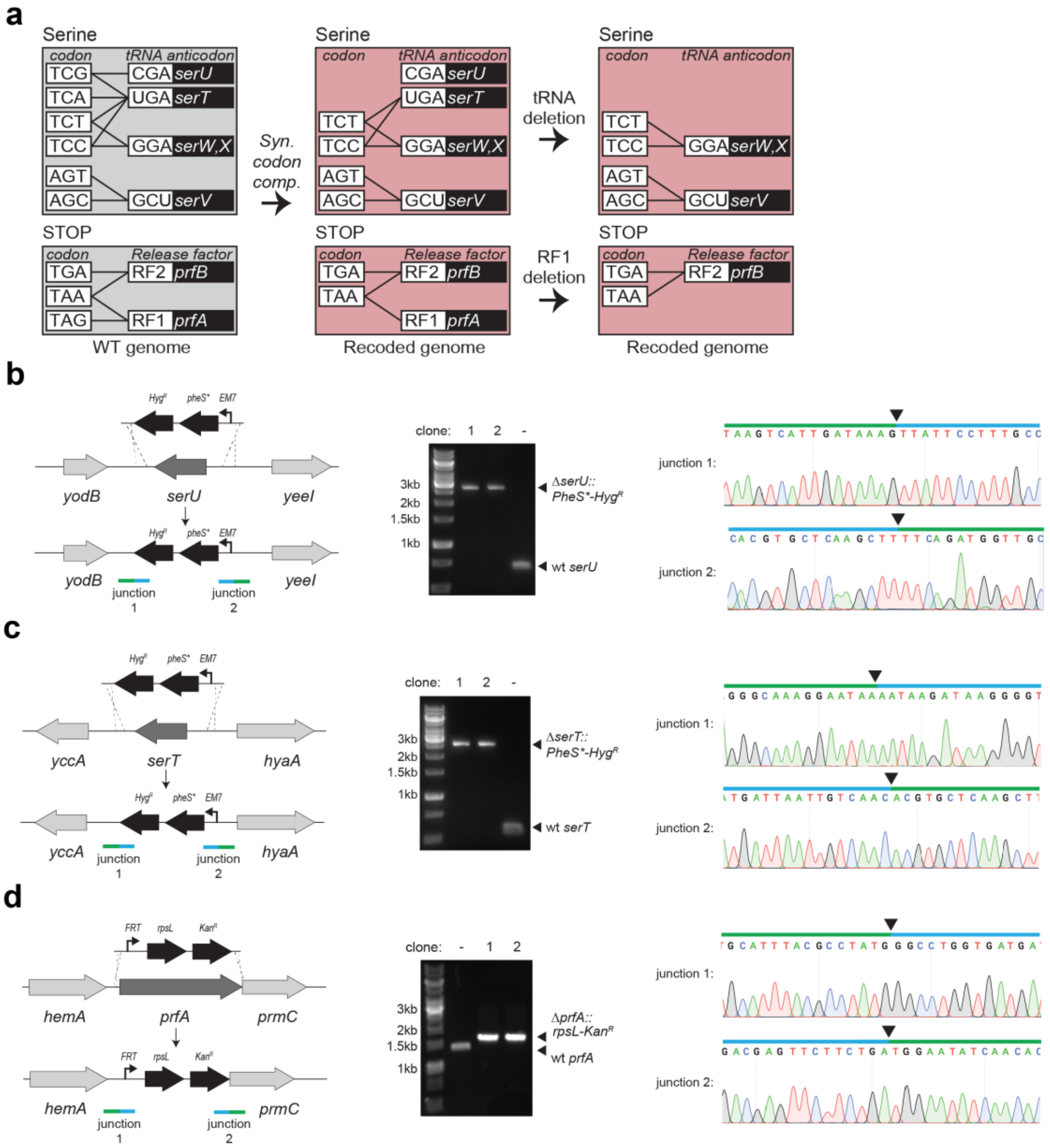


Figure 3.19 – Deleting decoding elements. a) Schematic of synonymous codon compression and tRNA deletion in Syn61. Removal of target codons TCG, TCA, and TAG is followed by the deletion of their cognate decoding elements *serU*, *serT* and *prfA*. **b)** On the left is a schematic of deletion of *serU*, by the lambda-red mediated integration of a *pheS*-hygR* cassette at its native locus. The same experiment was performed using an *rpsL-kanR* cassette (data not shown). Colony PCR at the *serU* locus in post-recombination clones (1 and 2) yields a longer product than the pre-recombination control (-), consistent with the presence of a *pheS*-hygR* marker (middle). The right panel shows a Sanger sequencing chromatogram of the newly generated junctions at the *serU* locus. The black arrows indicate the precise nucleotide that separates the native sequence from the *pheS*-hygR* cassette. **c)** Deletion of *serT*, as in b). This data corresponds to *serT* deletion in a separate Syn61 strain. **d)** Deletion of *prfA*, as in b) and c). This data corresponds to *prfA* deletion in a separate Syn61 strain.

Note: Wesley Robertson analysed wild-type Syn61 doubling times (**Figure 3.17**) and performed CYPK incorporation experiments jointly with Thomas Elliott. Julius Fredens performed mass spectrometry. Microscope images were taken by Yonka Christova with assistance from Nick Barry. Wesley Robertson performed RF-1 knockout (**Figure 3.19**). I performed deletions of tRNAs.

3.2.6 Sense codon reassignment in Syn61

Louise Funke performed adaptive laboratory evolution on Syn61, through several rounds of genome-wide mutagenesis followed by outgrowth and selection of the fastest growers. This process resulted in the generation of Syn61evo, which had a doubling time 30% shorter than that of wild-type Syn61. The strain contained 127 mutations, none of which affected annotated target codons. We performed deletion of *serU* with an *rpsL-kanR* marker as in **Figure 3.19a**, and then removed the marker via a second round of lambda-red recombination, providing a dsDNA template which restores a native *serU* environment where only the sequence corresponding to *serU* is missing (**Figure 3.20a**). Subsequently, an analogous two-step recombination approach was used to delete *serT* in the same strain, yielding Syn61evo_Δ*serU/T*; this strain provided a platform to probe sense codon reassignment.

We began testing reassignment of TCG. CGA is the cognate anticodon of TCG, and is expected to decode TCG exclusively. Reassignment of TCA would be, in principle, more challenging, as in *E. coli* its cognate anticodon UGA also decodes TCG, TCT and TCC by virtue of a cmo^{5}U modification in *serT* at the wobble position[240]. Consequently, here we focus on reassignment of TCG only.

Before investigating incorporation of unnatural amino acids, we set out to test the reassignment of TCG into other natural amino acids. We reasoned that by leveraging an endogenous tRNA and synthetase pair that we knew was highly active in Syn61, we could better assess the intrinsic availability of the TCG codon for reassignment, and remove unknowns related to the activity or efficiency of orthogonal aminoacyl-tRNA synthetase/tRNA pairs in Syn61. Alanine was a good candidate; since the endogenous *E. coli* alanyl-tRNA synthetase does not recognise the anticodon stem loop of alanyl-tRNA as an identity element[241], we could generate an alanyl-tRNA variant with a CGA anticodon (tRNA$^{Ala}_{CGA}$, denoted herein as *alaT*$_{CGA}$) while maintaining recognition as alanine by its cognate synthetase.

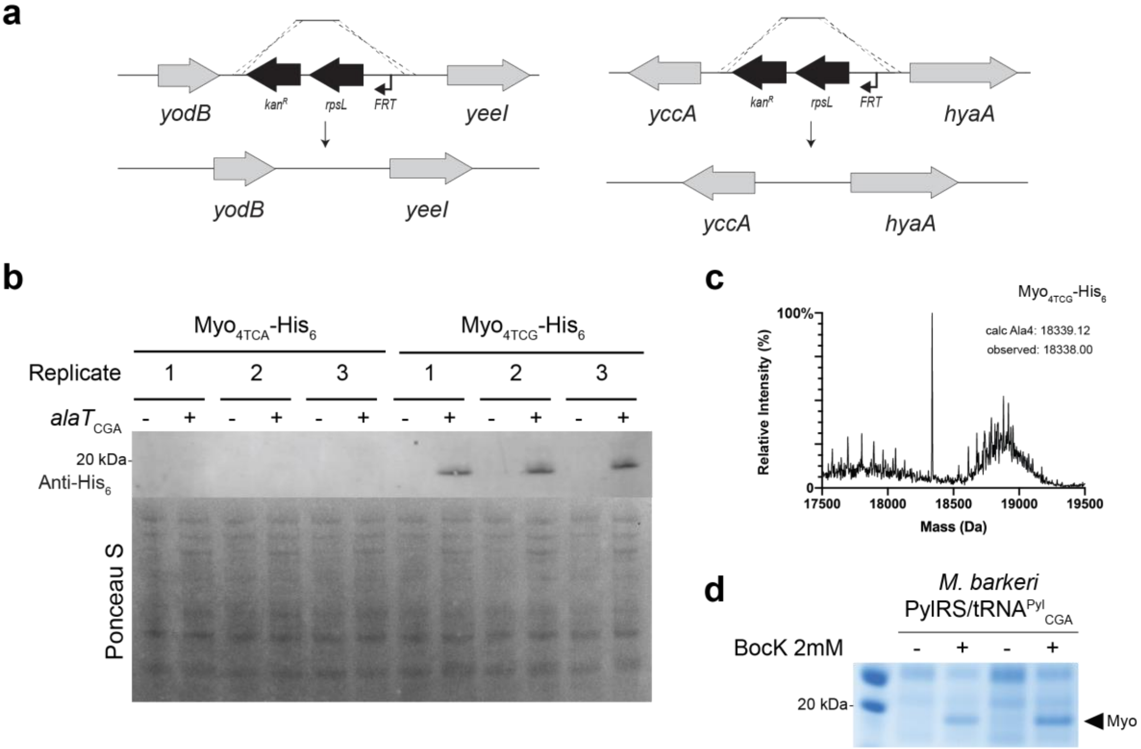


Figure 3.20 - Reassignment of TCG to natural and unnatural amino acids. a) Schematic of tRNA deletion to generate Syn61evo_Δ*serU/T*. *rpsL-kanR* cassettes at *serU* and *serT* loci were removed by lambda-red recombination, providing a construct with homology to either side of the selection cassette and counter-selecting with streptomycin. **b)** Anti-His$_6$ western blot of Syn61evo_Δ*serU/T* lysates after expression of Myo$_4$-His$_6$. Cells were co-transformed with Myo$_{4TCG/TCA}$-His$_6$ variants, and either a pSC101 plasmid containing *alaT*$_{CGA}$ (+) or an empty vector control (-). For each variant, replicates indicate independent expressions in 5 mL cultures from the same pool of transformed cells. Expression of myoglobin is observed in response to the expected cognate TCG codon, but not in response to the non-cognate TCA. Loading was normalized by harvesting and lysing equal amounts of cells based on the density (OD$_{600}$) of the expression cultures. A Ponceau S stain prior to the antibody stain is shown for loading comparison. **c)** Albeit highly noisy, the ESI-MS spectrum of purified Myo$_{4TCG}$-His$_6$ shows a peak corresponding to the calculated molecular weight of Myo-His$_6$ bearing alanine at position 4. **d)** Coomassie of Ni^{2+}-purified Myo$_{4TCG}$-His$_6$ expressed in Syn61evo_Δ*serU/T* harbouring a *M. barkeri* (PylRS)/tRNA$^{Pyl}_{CGA}$, in the presence (+) or absence (-) of 2 mM BocK.

We decided to investigate codon reassignment in myoglobin, a well-established model protein for genetic code expansion[67,121,242]. We transformed Syn61evo_Δ*serU/T* with i) a plasmid containing an *alaT*$_{CGA}$ tRNA under control of a native *serT* promoter region (or an empty vector control), and ii) a plasmid containing an arabinose-inducible His_6-tagged myoglobin reporter bearing TCG (Myo_{4TCG}-$His_{6)}$ or TCA (Myo_{4TCA}-His_6) at position 4. Upon induction of expression with arabinose, we observed *alaT*$_{CGA}$-dependent expression of myoglobin containing TCG, but not TCA (**Figure 3.20b**), as judged by Western blot of the cell lysates. We affinity-purified Myo_{4TCG}-His_6 with Ni^{2+}-NTA beads at very low yield, and performed ESI-MS. Due to the low yield, the spectrum was noisy, but we observed a peak consistent with the presence of alanine at position 4 (**Figure 3.20c**).

We next tested whether we could re-assign TCG into a non-canonical amino acid. We transformed the Myo_{4TCG}-His_6 reporter into Syn61evo_Δ*serU/T* together with a *M. barkeri* PylRS/$tRNA^{Pyl}_{CGA}$ orthogonal pair. We expressed Myo_{4TCG}-His_6 in 1L cultures in the presence and absence of cognate PylRS substrate N^{ε}-(*tert*-butoxycarbonyl)-L-lysine (BocK). Analysis of the purified samples by SDS-PAGE showed BocK-dependent expression of myoglobin (**Figure 3.20d**). This suggested that, in Syn61evo_Δ*serU/T*, we could reassign TCG to non-canonical amino acids. However, the protein yield was too low to obtain a conclusive ESI-MS spectrum.

We observed that in our myoglobin purifications, endogenous proteins co-eluted with myoglobin at high levels; the low expression yields presumably resulted in low occupancy of myoglobin in the nickel beads, which may increase unspecific binding. However, we did not perform extensive optimization of expression and purification conditions. Instead, we tried another model protein, T4 Lysozyme (T4L)[243–245], to try to obtain better yields, and to further test whether TCG reassignment into unnatural amino acids was clean and generalizable.

We co-transformed the *M. barkeri* PylRS/$tRNA^{Pyl}_{CGA}$ orthogonal pair with T4L constructs containing a TCG codon at positions 38, 72, 83, 153 or 157, cloned in the same plasmid backbone as Myo_{4TCG}-His_6. Expression levels varied significantly with the position of TCG, but in all cases, we observed BocK-dependent expression of T4L (**Figure 3.21a**). While we mostly did not observe T4L expression in the absence of BocK, in $T4L_{157TCG}$-His_6 there seemed to be a low level of TCG read-through. This underscores that the local context of the TCG codon within the mRNA message can influence translation dynamics and the extent of reassignment.

We next tested whether we could achieve incorporation of BocK at two TCG codons within the same polypeptide. We observed BocK-dependent expression of T4L$_{72,83TCG}$-His$_6$ and T4L$_{83,157TCG}$-His$_6$, but not T4L$_{38,72TCG}$-His$_6$ (**Figure 3.21b**). The efficiency of expression of T4L bearing dual TCG codons seemed to correlate well with their individual efficiencies (**Figure 3.21a**). T4L$_{72,83TCG}$-His$_6$ was purified and subject to ESI-MS, which was consistent with the presence of BocK residues at positions 72 and 83 (**Figure 3.21c**).

Wesley Robertson and Louise Funke performed deletion of RF-1 in Syn61evo_Δ*serU/T*, yielding the strain Syn61evo_Δ*serU/T*Δ*prfA*. We used Syn61evo_Δ*serU/T*Δ*prfA* to assess the efficiency of incorporation of three non-canonical amino acids in response to TCG within the same polypeptide. In a preliminary experiment, BocK-dependent expression of T4L was observed in all tested codon combinations (**Figure 3.21d**). Overall, these results show that we can effectively reassign TCG in Syn61 following deletion of the cognate tRNAs; the extent and 'cleanliness' of the reassignment seems to be subject to multiple variables which will be investigated in future work.

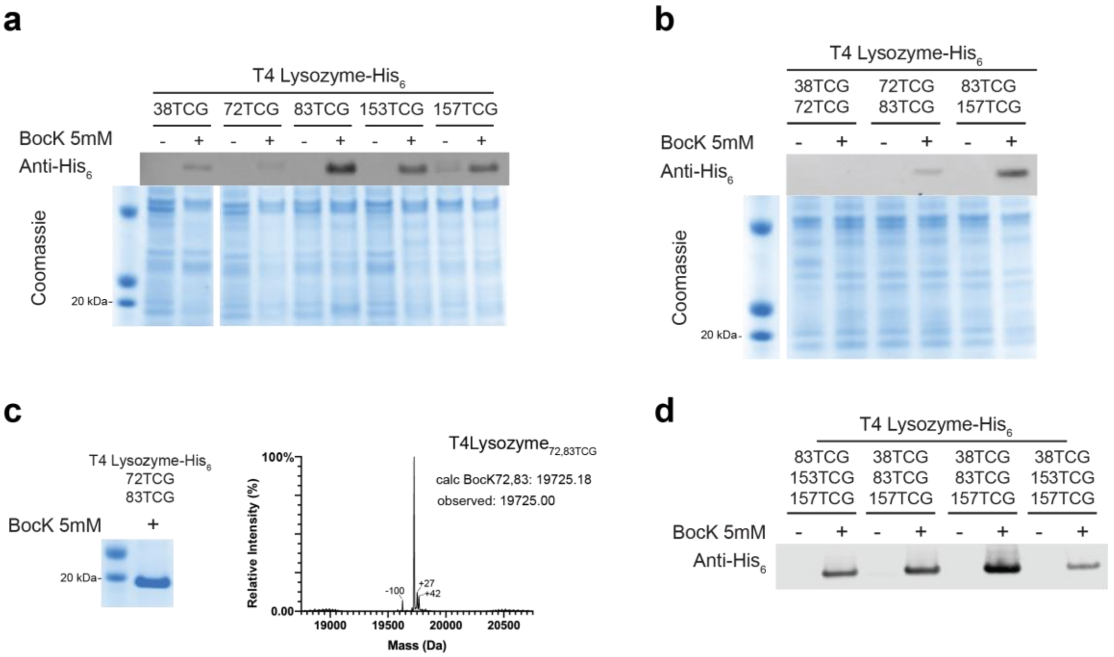


Figure 3.21 - Multisite TCG-dependent BocK incorporation in T4 Lysozyme. a) Anti-His$_6$ western blot of Syn61evo_Δ*serU/T* lysates expressing multiple T4Lysozyme-His$_6$ bearing TCG codons at the indicated positions. Cells contained the *M. barkeri* (PylRS)/tRNA$^{Pyl}_{CGA}$ and expressed T4Lysozyme in the presence (+) or absence (-) of 5 mM BocK. BocK-dependent expression is observable in all cases, albeit with varying efficiency. A Coomassie stain of the

same lysates is shown below as a loading comparison, re-constructed from 2 regions of the same gel. **b)** Analogous to a), expressions of T4Lysozyme-His$_6$ bearing two TCG codons within the same message, at the indicated positions. **c)** On the left, a Coomassie stain of purified T4Lysozyme$_{72,83TCG}$-His$_6$, expressed in Syn61evo_Δ*serU/T* with *M. barkeri* (PylRS)/tRNA$^{Pyl}_{CGA}$ supplemented with 5 mM BocK. On the right is the ESI-MS spectrum of the same purification, showing a major peak corresponding to the calculated molecular weight of T4Lysozyme-His$_6$ bearing BocK residues at positions 72 and 83. Three additional peaks are observed at -100 Da, +27 Da and +42 Da. The -100 peak may correspond to lysine, or to the loss of one *tert*-butoxycarbonyl group from a BocK residue due to fragmentation during ionization or sample preparation. The nature of the +27 and +42 peaks is unclear. **d)** Anti-His$_6$ western blot of Syn61evo_Δ*serU/T* lysates expressing T4Lysozyme-His$_6$ variants that contain 3 TCG codons at the indicated positions. Loading was normalized by harvesting and lysing equal numbers of cells based on spectrophotometric readings (OD$_{600}$).

Note: tRNA deletions in Syn61evo were performed jointly with Salvador Buse and Louise Funke. I performed all reassignment experiments in this section.

3.3 – Discussion

We designed, synthesised and assembled a recoded *E. coli* genome where all annotated instances of serine codons TCG and TCA, and stop codon TAG, were systematically replaced by their synonyms. We performed partial stepwise genome replacements in parallel strains using GENESIS, and then used directed conjugation to assemble the multiple synthetic sections into a single, fully recoded genome. During this process, we identified 4 deleterious sequence designs; we used REXER to progressively narrow down the problematic regions, which allowed us to investigate the underlying causes. We could then apply solutions on a case-by-case basis.

It is known that synonymous substitutions may be deleterious, and identifying suitable substitutions in non-trivial. In prior work, we defined candidate schemes for whole genome recoding based on the maintenance of relevant codon parameters such as tRNA adaptation index (tAI), codon adaptation index (cAI), and a combination of both (tE), and tested them on a ~17 kbp region of the *E. coli* genome[185]. Some recoding schemes, including the one chosen for carrying out this work, were viable. Others failed, despite being *a priori* equally good candidates based on the metrics. It is remarkable that a set of recoding rules validated on a relatively small section of the genome could be used to recode the entire genome with very few corrections. This underscores the highly empirical nature of recoding, and shows that testing recoding rules in subsections of the genome is a viable approach for identifying recoding patterns that can be applied genome-wide.

As discussed in Chapter 1, mRNA secondary structures are emerging as a major determinant for translation dynamics. Out of the 4 deleterious sequences that we found, we could associate one of them to the disruption of regulatory motifs in the 5' UTR of the essential gene *rne*. A similar effect may underlie the lethality of a codon substitution in *yaaY*, which overlaps the 5' UTR of essential gene *ribF*. *ribF* is transcribed polycistronically with *ileS,* which encodes the isoleucyl-tRNA synthetase and is also essential, from at least one promoter[246]. The reasons underlying sequence lethality at *map* and the *ftsI-murE* overlap are less clear, and will require further investigation.

The ability to realize genome designs at high fidelity will allow for deconvoluting which deviations from the designed synthetic sequence are direct consequences of faults in the genome design, as opposed to spontaneous errors and mutations. In previous recoding efforts, a significant number of off-target mutations and reversions of target codons have confounded the assessment of the effects of recoding[185,188]. In this work, the final Syn61 genome contained 8 non-programmed mutations with respect to the design; 4 of them originated during assembly, and the remaining 4 during construction of the recoded sections. This was a surprisingly low number of errors, considering the scale of the synthesis; at 3,978,937 bp, the Syn61 genome synthesis is almost 4 times larger than the previous record holder[171](1.08 Mb).

The recoding rules employed in this work were overall well tolerated; it seems logical that, when using recoding schemes that pose a larger burden on cellular functions, the number of non-programmed mutations may increase due to selective pressures for compensating the deleterious effects of recoding. The high-fidelity workflow presented here may prove particularly valuable in such cases, as it may allow for accurately identifying bottlenecks and deleterious sequences with minimal biological noise. Overall, we have demonstrated that our pipeline for genome assembly, combining REXER, GENESIS and directed conjugation, is suitable for the precise implementation of genome designs *in vivo*, and may serve as a foundation in future prokaryotic genome syntheses.

Considering the large number and high density of codon changes introduced, it is remarkable that the synthetic strain doubled only 1.6x times slower than its progenitor. No significant morphological changes were observed either, and the proteomes of both strains were similar, indicating that a reduced set of codons (61 as opposed to 64) is suitable for encoding a viable proteome. This was further confirmed by experiments in which we directed the co-translational incorporation of a non-canonical amino acid into codons TCG, TCA and AGC, in both Syn61 and the parent MDS42. The tRNA/synthetase pair that incorporates CYPK is in competition with the endogenous tRNAs for each of the tested codons. Hence, we expected dosage-dependent toxicity through poisoning of the proteome, presumably due to misfolding or catalytic inactivation of endogenous proteins. The toxicity of incorporation into each codon correlated with the number of codons that we expected to be present in the respective genomes.

Immediately after completion of Syn61, we could delete all TCG, TCA and TAG cognate decoding elements individually. In an evolved derivative of Syn61, termed Syn61evo, we deleted *serU*, *serT* and RF-1 in the same strain to generate Syn61evo_Δ*serU/T*Δ*prfA.* When provided with an alanyl-tRNA or an *orthogonal M. barkeri* pyrrolysine pair bearing CGA anticodons, alanine or BocK were incorporated in response to TCG in an amino acid-dependent fashion, in myoglobin and T4 Lysozyme. A preliminary experiment suggested that unnatural amino acids could be encoded up to three times within a single polypeptide, although this must be further confirmed by mass-spectrometry of the resulting proteins. Together, these results suggest that upon removal of the tRNAs, at least one of the 3 new spaces in the genetic code of Syn61 is effectively blank, and available for reassignment into other amino acids. The expression efficiency and the level of background TCG read-through seemed to depend on context effects related to the position of TCG within the transcript.

One of the model proteins we used, T4 Lysozyme, has previously been useful for demonstrating non-canonical amino acid incorporation[243–245]. In our hands however, expression of T4 Lysozyme caused some level of toxicity, presumably due to digestion of the cell wall, which resulted in decreased OD_{600} of expression cultures. Accurate quantification of protein yields in future reassignment efforts in Syn61 may benefit from the use of inactivated versions of T4 Lysozyme and alternative model proteins. Such experiments will serve to further probe the extent and efficiency of TCG reassignment *in vivo*.

Although not explored here, previous work suggests that the TAG amber codon may also be amenable to clean reassignment[100]. Future work will also investigate reassignment into the TCA codon removed in Syn61; this may require the engineering of tRNAs whose anticodon recognizes TCA exclusively, and not TCG, TCT or TCC as is the case for *serT*. Provided that sufficient mutually orthogonal aminoacyl-tRNA synthetase/tRNA pairs are available, this may enable the creation of a 23 amino acid genetic code in Syn61.

Chapter 4 – Programmed genome fission and chromosomal fusion in *Escherichia coli*

Where appropriate, figures in this chapter are edited and reproduced with permission from their original publication:

Wang, K., **de la Torre, D.**, Robertson, W. E. & Chin, J. W. Programmed chromosome fission and fusion enable large-scale genome rearrangement and assembly. *Science* **365,** 922-926 (2019).

Note: This project was carried out jointly with Kaihang Wang and Wesley Robertson. Our specific contributions are clarified throughout the text.

4.1 – Introduction

So far, this dissertation has discussed and exemplified our ability to write synthetic designer genomes, which constitutes a novel bottom-up approach for addressing biological questions, and may prove useful in the creation of synthetic organisms with desirable properties.

The recoding design described in Chapters 2 and 3 of this dissertation, as well as others[186,188], have been based on the known sequence and genome architecture of natural microorganisms. Despite a high density of codon substitutions, these designs have maintained the pre-existing genomic arrangement. Future genome designs may re-organize the genome into alternative configurations, perhaps by clustering genes that are functionally related. This may make genomes more modular and intuitive for investigators, and facilitate the exchange and engineering of gene clusters in a programmable manner. Genome re-organization and clustering of related genes (on a small scale) have been explored in previous synthesis efforts[187,189]. These were directly implemented in the synthetic genome design and assembled in the absence of a pre-existing template in *S. cerevisiae,* for subsequent transfer into their target host. In general, the ability to perform such manipulations *a posteriori* directly in the genome of a host organism may facilitate the identification of viable re-organizations, and directly couple different genomic configurations to functional outcomes.

Normally, efforts to synthesize and assemble genomes break down the synthesis in a number of intermediates, which then have to be combined into a single genome in a given host. When the synthetic genome does not introduce structural rearrangements with respect to the natural counterpart, directed conjugation has proven to be useful in combining sub-sections of genomes from two progenitors (Chapter 3, ref [100]). It does, however, have caveats, as recombination between donor and recipient genomes requires long homology regions of at least 3 kbp. Moreover, the precise sites of recombination cannot be controlled; these can be modulated by fine-tuning conjugation times and strategically placing selectable markers in the donor and recipient genome, however this only provides resolution on the kilobase scale. Finally, conjugation offers little flexibility with respect to the location or configuration of donor DNA in the recipient.

Overall, the genome synthesis field would benefit from a set of technologies that allow the precise creation of large-scale structural modifications to genomes, including insertions, deletions, inversions, translocations ('cut and paste' somewhere else in the genome) and the combination of large DNA pieces from diverse progenitors into a single strain at high resolution. In *E. coli*, which has accumulated much genome engineering attention, some of these are still lacking.

A common intermediate step to performing all of these operations could be the splitting of defined sub-sections of a chromosome into multiple sub-chromosomes. Chromosome splitting occurs naturally in evolution[247–250], but few studies have tried to recapitulate or engineer this type of process in the laboratory. In *S. cervisiae*, the 16 linear chromosomes have been combined into two[251] or one[252] mega-chromosomes by progressive removal of telomeres and creation of inter-chromosomal linkages through PCR products. Conversely, the integration of telomere sequences or *Tetrahymena* rDNA ends is naturally resolved in yeast by splitting the chromosome at the site of insertion[253]; derivatives of this process have been used to probe the structural constraints and malleability of the yeast genome[254,255]. In *Bacillus subtilis*, whose genome is known to be highly malleable[181], chromosome splitting has also been observed[256]. Two halves of a CDS for kanamycin resistance were installed 300 kbp apart. Homologous recombination at these sites rescued the frame, and chromosomal splits could be selected with kanamycin[256]. All these examples of chromosomal manipulation leveraged high endogenous levels of homologous recombination in their respective hosts. Additionally, non-programmed spontaneous instances of chromosome splitting in the laboratory have been observed in *Haloferax volcanii*[257].

In *E. coli,* one study reported the excision of up to 720 kbp of the genome into sub-chromosomes[258] using a split selection marker approach, analogous to that used in *Bacillus subtilis* above. However, only a small fraction of the *E. coli* genome was targeted in these experiments. Moreover, the low levels of endogenous recombination in *E. coli* presumably make this process very inefficient, and its generalizability is unclear. Another interesting report achieved the splitting of the *E. coli* circular genome into two linear chromosomes[259]. A pair of *tos* sequences from bacteriophage N15 were integrated in the *E. coli* genome. In N15, protelomerase *telN* recognizes and cleaves *tos* sequences, sealing each end by formation of a hairpin[260]. Expression of *telN* in *E. coli* cells harbouring two *tos* sequences resulted in splitting of the genome at these sites, and the generation of two linear sub-chromosomes. However, only 1 characterized arrangement was viable out of multiple attempts. In both of these examples, it is unclear how the new chromosomes may be precursors for further manipulations.

We envisioned that the ability to split the *E. coli* genome into pairs of chromosomes, in a format amenable for their fusion back into a single chromosome in a variety of configurations, may provide a broader set of tools for probing genomic plasticity, and facilitate genome synthesis, re-organization and assembly. Consequently, here we set out to develop a suite of technologies for manipulating the genome of *E. coli*, including chromosome splitting, inversions, translocations and chromosomal transplant.

4.2 – Results

4.2.1 A strategy for genome fission in *E. coli*

Chapters 2 and 3 have described REXER, a method for the replacement of large segments of genomic DNA in *E. coli*. REXER couples CRISPR/Cas9 and lambda-red mediated recombination to achieve the exchange of DNA between the host genome and an episomal BAC. Building on this approach, we envisioned that if we prepared an artificial BAC with appropriate homologies for recombination with the genome, we may be able to use CRISPR/Cas9 and the lambda-red machinery to catalyse the permanent transfer of host chromosome regions into the BAC. This would allow us to effectively split the wild-type *E. coli* chromosome into a pair of synthetic chromosomes, and we may be able to precisely control their sequence by defining appropriate Cas9 cut sites and homology regions.

We devised a system to split the *E. coli* chromosome into two, which we term 'Genome Fission' (**Figure 4.1**). We set out to test it by performing fission of the *E. coli* MDS42$^{\mathit{rpsL}\mathrm{K43R}}$ genome into two chromosomes, of 3.43 and 0.56 Mb. We began by specifying the precise nucleotides that defined our 0.56 Mb region of interest, with the only constraints that they were in intergenic regions and sat adjacent a PAM sequence in order to enable Cas9 cleavage. The ~50 bp upstream and downstream the 5' boundary of the 0.56 Mb region are termed HR1 and HR2. Similarly, the 3' boundary is flanked by HR3 and HR4.

We then constructed a 'fission BAC' that contained:

1. A *luxABCDE* operon, which we term 'linker 1'
2. A BAC vector comprising the BAC replication and segregation components, coupled to a *sacB-cat* double selection cassette. We term this 'linker 2'
3. Two copies of wild-type *rpsL*, sandwiched between linkers 1 and 2. Each copy is flanked on both sides by an artificial and unique Cas9 cut site

These components are separated by homology regions HR1-HR4, as shown in **Figure 4.1**. The BAC encodes 4 Cas9 cut sites, at the junctions of linkers 1 and 2 and the two *rpsL* copies. The PAM of the cut sites is encoded within the linker sequences, whereas each protospacer is defined by HR1, HR2, HR3 or HR4.

In order to perform fission, we first transformed a helper plasmid (encoding Cas9 and the lambda-red recombination machinery) into the target cells, and followed by transforming the fission BAC. Subsequently, we induced expression of Cas9 and lambda-red recombination machinery, and provided a set of spacers to direct Cas9 cleavage of the genome (twice) and the fission BAC (4 times). As a result, the genome is split into what we term 'fragment 1' (here, 3.43 Mb) and 'fragment 2' (here, 0.56 Mb). In the BAC, linkers 1 and 2 are released, and decoupled from *rpsL*. Cas9 cuts are designed such that the resulting fragments and linkers contain exposed homologous ends for re-circularisation by lambda-red mediated recombination. The products of this circularisation are:

- Chromosome 1 (Chr. 1), formed by fragment 1 and linker 1
- Chromosome 2 (Chr. 2), formed by fragment 2 and linker 2

In order to identify cells with two chromosomes, we spread cells in medium supplemented with chloramphenicol and streptomycin. This selects for i) maintenance of linker 1, and ii) loss of *rpsL*. Because these two components are originally coupled in a single replicon, by selecting for linker 1 and against *rpsL* we enrich for cells which have decoupled them via Cas9 cleavage of the fission BAC.

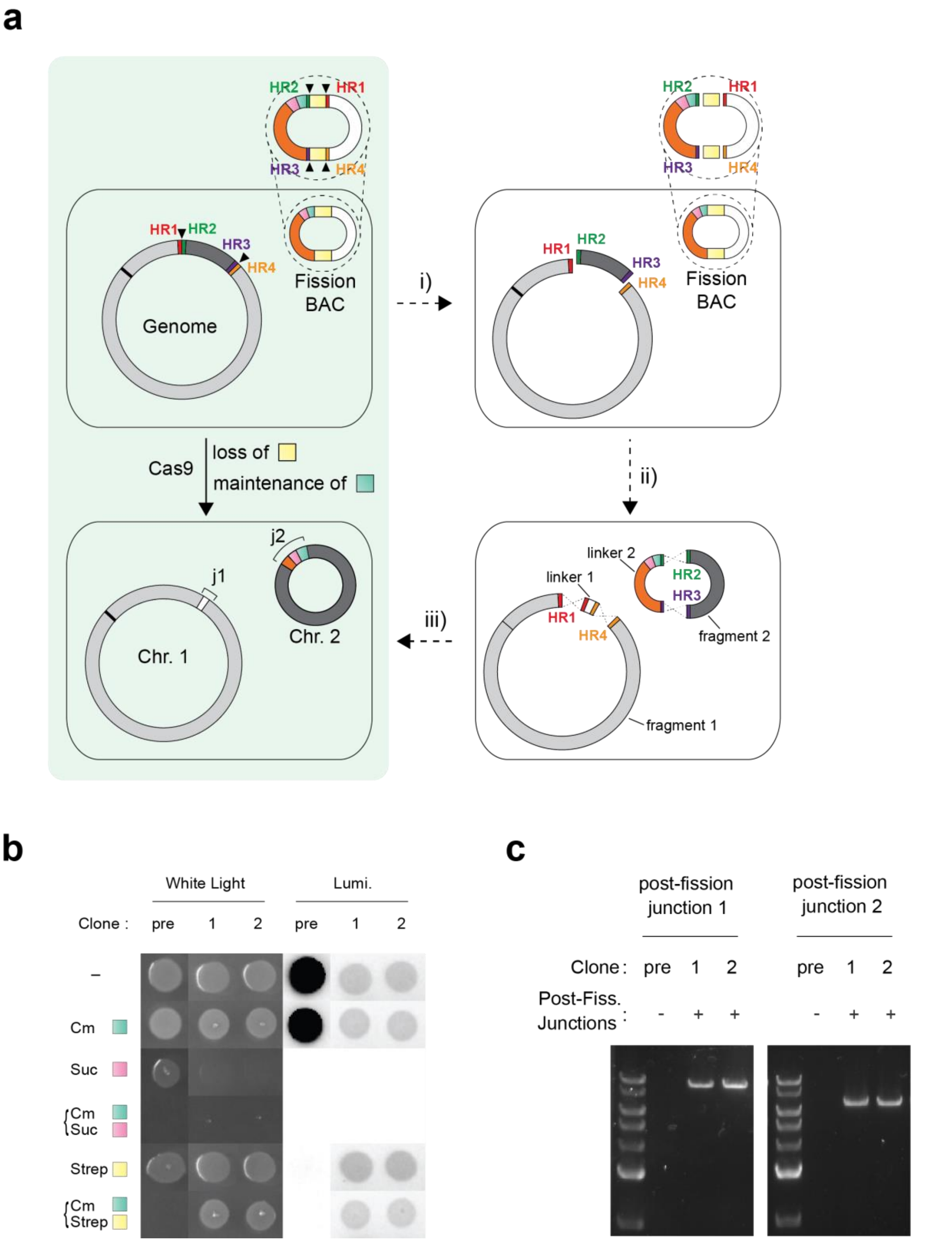


Figure 4.1 - Genome fission. a) Schematic of the genome fission process. The pre-fission cell (top left) contains a single-chromosome, circular genome. The portion of the chromosome targeted for splitting into Chr. 2 is in dark grey. Cas9 cut sites are indicated by black arrows. Homology regions HR1-4 are indicated. The fission BAC consists of a *sacB-cat* double selection cassette (pink and green, respectively), two copies of *rpsL* (yellow), a *luxABCDE* operon (white) and the BAC segregation and replication machinery (orange). i) Cas9 cleavage splits the genome into fragments 1 and 2 (light and dark grey, respectively), and the BAC into

(continued from previous page) linkers 1 and 2. ii) The resulting fragments and linkers contain exposed homology regions for recombination. iii) Lambda-red mediated recombination generates chromosome 1 (Chr. 1, 3.43 Mb) and chromosome 2 (Chr. 2, 0.56 Mb). The two new chromosomes contain novel junctions j1 and j2. *oriC* is indicated by a black line. **b)** Stamping of genome fission products in selective media. Pre-fission cells (pre) display growth phenotypes consistent with the presence of a fission BAC, whereas post-fission clones (1 and 2) show phenotypes consistent with the splitting of the genome into two chromosomes. Luminescence measurements are denoted by 'Lumi.'. Cells are stamped into LB agar supplemented with 20 µg/mL chloramphenicol (Cm), 7.5% sucrose (Suc), 100 µg/mL streptomycin (Strep) or a combination of these. **c)** Junction-specific PCR yields products corresponding to the presence of new junctions j1 and j2 in post-fission clones (1 and 2), but not in pre-fission cells (pre).

We preliminarily identified successful post-fission clones by stamping the colonies in selective media, and assessing their sensitivity to sucrose and streptomycin, as well as their luminescence (**Figure 4.1b**). In the pre-fission strain, the two copies of *rpsL* and the *sacB-cat* (in linker 1) are episomal, and not coupled to the genome. This means that pre-fission cells can survive in the presence of streptomycin or sucrose alone, via loss of the fission BAC, but not in the presence of both chloramphenicol and sucrose/streptomycin, as loss of the fission BAC is not possible. After fission, however, *sacB-cat* is linked to an essential 0.56 Mb piece of the genome (Chr. 2), which prevents sucrose survival via loss of the BACs. Additionally, in post-fission clones *rpsL* is lost, which allows the cells to grow in the presence of chloramphenicol and streptomycin. Finally, in the pre-fission BAC the *luxABCDE* operon in linker 1 is under control of the strong *rpsL* promoter upstream. In post-fission cells, where linker 1 is part of Chr. 1, it is expressed under control of only its own weaker promoter, which results in a drop in luminescence in post-fission cells with respect to pre-fission cells.

The generation of Chr. 1 and Chr. 2 during fission results in the formation of two new junctions, which we term j1 and j2 (**Figure 4.1c**). In order to further confirm that post-fission clones contained a split genomic arrangement, we performed PCR across j1 and j2. We observed products of the correct size in post-fission clones, but not in the pre-fission control. These results suggested that we had split the *E. coli* genome into Chr. 1 (~3.43 Mb) and Chr. 2 (~0.56 Mb).

Note: The experiment in this section was performed by Kaihang Wang.

4.2.2 Robustness and flexibility of genome fission

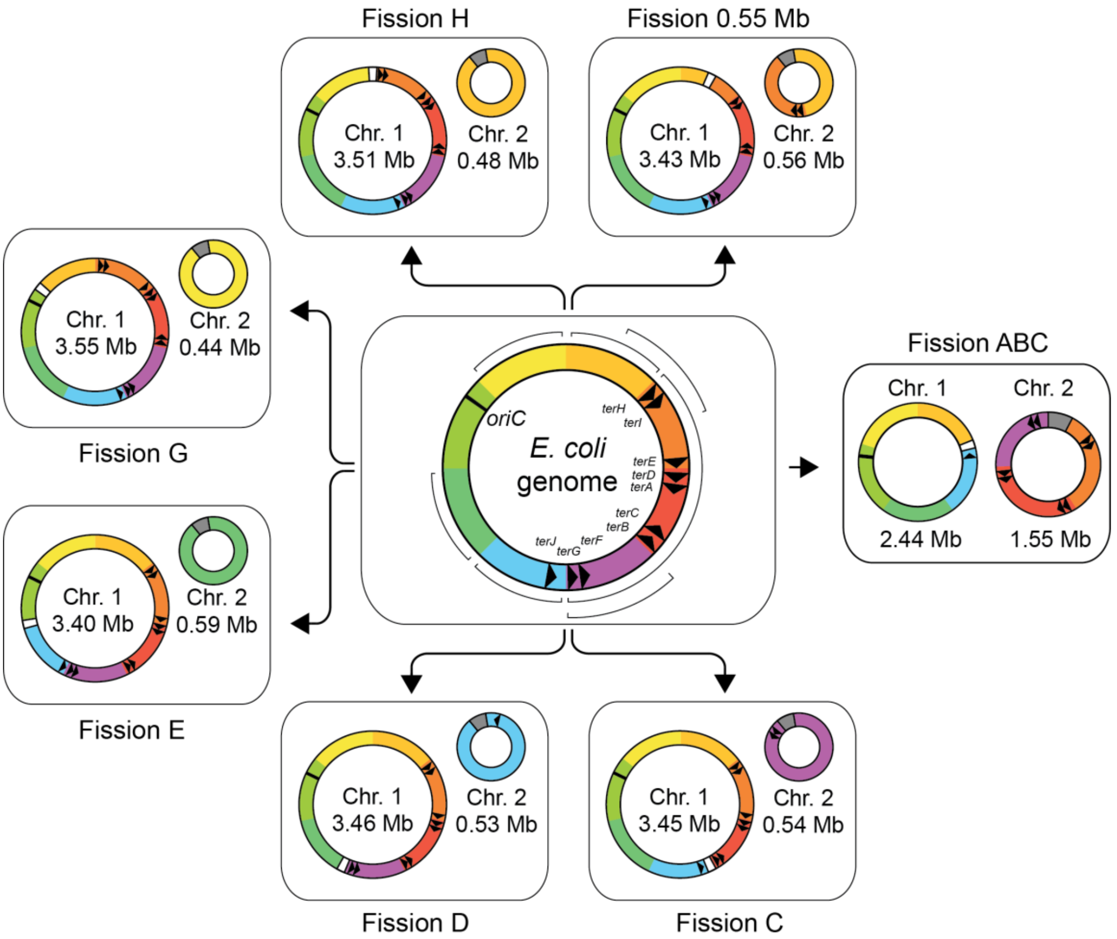


Figure 4.2 - Genome fission throughout the genome. Schematic of the different successful fission experiments performed in this Chapter. The *E. coli* genome in the centre is colour-coded according to the sections A-H defined in Chapter 3. The section contained by Chr. 2 in each case is indicated. Fission 0.55 Mb corresponds to the fission described in **Figure 4.1**. Fission C was performed in a strain where section C is recoded, as described in the text (Section 4.2.3). Linkers 1 and 2 are represented in white and grey, respectively. *oriC* is represented by a black line. *ter* sites (A-J) are thought to act as unidirectional checkpoints for the replication fork during cell division, eventually trapping replication forks between *terA* and *terC*. Their location and directionality before and after each fission experiment is indicated by black arrows.

We set out to test whether our method for splitting the genome was generalizable to other sites. We envisioned that fission may be useful in the assembly of synthetic *E. coli* genomes and complementary to the retrosynthetic strategy outlined in Chapter 3, which may form the basis for future *E. coli* genome syntheses. For this reason, we decided to test the fission of ~0.5 Mb pieces corresponding to the Sections defined in our recoded genome design (**Figure 3.2**). Additionally, in order to test the size limitations of genome fission, we attempted fission of a ~1.5 Mb segment, corresponding to sections A, B and C.

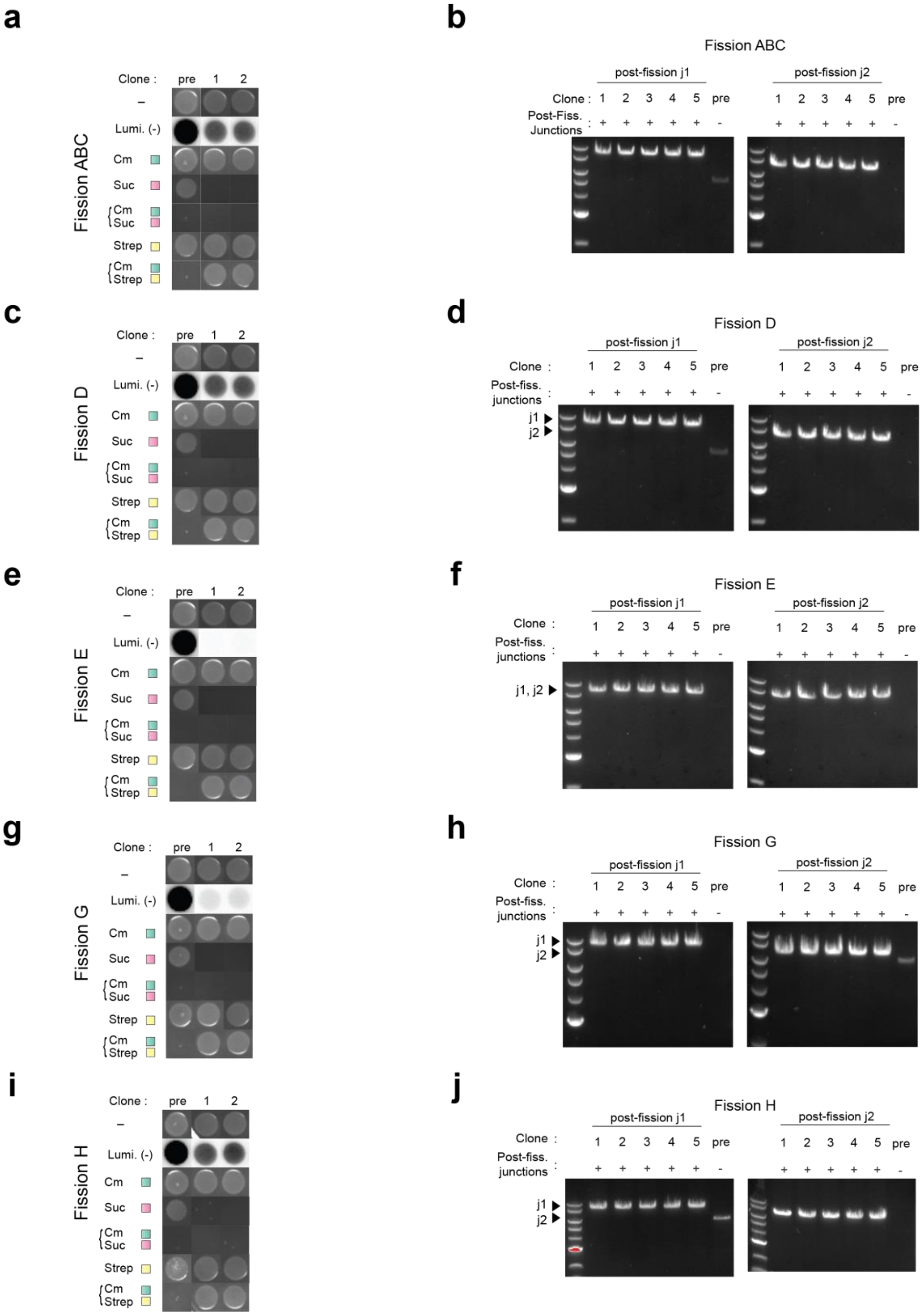
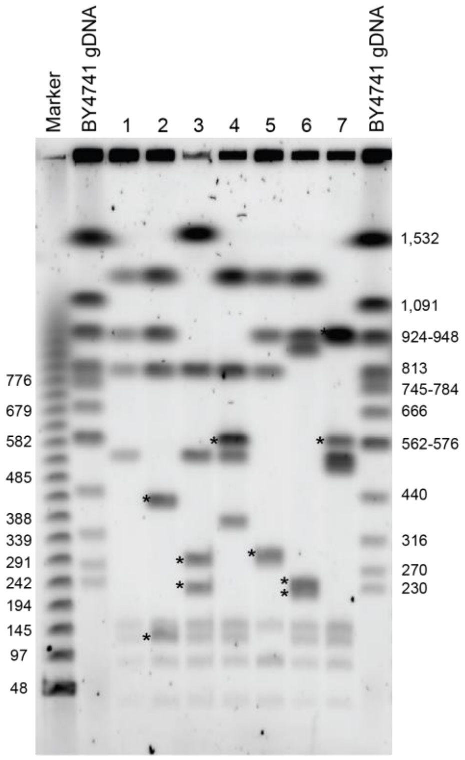


Figure 4.3 - Selective growth and junction-specific PCR for multiple fissions. Stampings in selective LB agar medium and PCR at j1 and j2 for pre- and post-fission clones, for the experiments shown in **Figure 4.2**. In each case, the observed growth phenotypes are

(continued from previous page) consistent with the genome being split into two chromosomes by genome fission (as in **Figure 4.1b**); cells are sensitive to sucrose, and the transfer of linker 1 from the fission BAC to chromosome 1 results in a drop in luminescence intensity (Lumi). Pre-fission cells (pre) display phenotypes consistent with the presence of a fission BAC. The selective agents are 20 μg/mL chloramphenicol (Cm), 7.5% sucrose (Suc), 100 μg/mL streptomycin (Strep) or a combination of these. '-' indicates plain LB agar. PCR across newly generated junctions j1 and j2 yields products of the expected size in post-fission clones (1-5), but not in pre-fission cells (pre).

All these attempts at fission yielded clones whose growth phenotypes and genotypes at j1 and j2 were consistent with genome fission at the defined boundaries (**Figure 4.2**). For all these experiments, phenotyping and genotyping served to confirm i) cleavage of the BAC and loss of *rpsL*, ii) the coupling of linker 2 to essential DNA, and iii) generation of the expected junctions (**Figure 4.3**). However, we could not rule out the possibility that post-fissions genomes had adopted incorrect configurations compatible with all of the above, via additional compensatory rearrangements. We set out to obtain additional evidence of our desired two-chromosome configuration in all post-fission strains, by two separate methods.



MDS42 wild-type	Fission 0.55 Mb	Fission D	Fission E	Fission G	Fission H	Fission ABC
1	2	3	4	5	6	7
1,246,785	1,246,785	1,681,672	1,246,785	1,246,785	1,246,785	956,536
952,029	952,009	815,317	815,317	952,009	952,009	952,009
815,317	814,274	550,099	587,360	815,317	903,828	593,467
550,099	430,043	298,054	550,099	312,112	245,792	550,099
144,278	144,278	232,941	378,508	293,369	229,655	525,958
131,128	134,946	144,278	144,278	144,278	144,278	144,278
92,304	131,128	131,128	131,128	92,304	131,128	131,128
39,983	92,304	92,304	92,304	89,605	92,304	92,304
1,505	39,983	39,983	39,983	39,983	39,983	39,983
1,419	1,505	1,505	1,505	1,505	1,505	1,505
1,419	1,419	1,419	1,419	1,419	1,419	1,419
	1,419	1,419	1,419	1,419	1,419	1,419

Figure 4.4 - Pulse-field gel electrophoresis of diverse fission experiments. Pulse-field electrophoresis gel showing AvrII restriction patterns of the chromosomal arrangements generated by the different genome fission experiments (**Figure 4.2**). The table specifies the expected AvrII restriction fragments for each experiment. Values in grey correspond to

(continued from previous page) restriction fragments which derive from the cleavage of Chr. 2, and the corresponding bands are indicated on the gel with an asterisk. AvrII digestion generated the expected restriction patterns in all experiments. 'Marker' corresponds to New England Biolabs' Lambda PFG ladder. Agarose plugs of *S. cerevisiae* BY4741 were prepared in-house and used as a ladder ('BY4741 gDNA'). The size of the bands in the ladders are indicated in kbp.

We embedded post-fission cells in agarose plugs, performed in-gel restriction digestion with AvrII, and analysed the genomic restriction pattern by pulse-field gel electrophoresis. Based on a virtual map of the two chromosomes that formed the split genome, we calculated the expected restriction products pre- and post-fission strains. AvrII cuts the wild-type MDS42 genome 7 times, but since linker 2 contains an additional AvrII site, the total number of cuts in post-fission strains is 8. The restriction patterns we observed were consistent with the predicted fragment sizes (**Figure 4.4**).

A higher resolution, unbiased way to assess the new chromosomal arrangements would be a *de novo* genome assembly, where next-generation sequencing data is used to re-construct the genome sequence. This essentially occurs through the overlapping of sequencing reads to generate larger contiguous sequences called 'contigs'. Because the result of our experiment is expected to be two circular chromosomes, an ideal *de novo* assembly would yield two independent circular contigs, corresponding to Chr. 1 and Chr. 2. Short reads (e.g. Illumina) have low error rates, but they cannot produce complete assemblies since structural information is lost when they encounter repeats that are longer than the read length. On the other hand, long reads (e.g. Oxford Nanopore, IonTorrent, PacBio) can be up to several hundred kbp long, and hence provide a good structural picture of the genome. However, they suffer from low accuracy.

We decided to perform a hybrid assembly, in which a scaffold of the genome architecture is generated using long reads, and the short accurate reads improve the quality of the assembly at the nucleotide scale. This type of assembly retains the desirable properties of both types of data, and has been previously shown to generate complete high-quality assemblies of prokaryotic genomes[261]. We sequenced the genome of a post-fission clone using both short-read (Illumina MiSeq 2x300 bp) and long-read (Oxford Nanopore MinION) technologies. The N50 of the Oxford Nanopore reads ranged from 8.3 kbp to 60 kbp. For each fission experiment, we used both datasets as inputs in a hybrid *de novo* assembly with the Unicycler package[262]. Briefly, Unicycler first builds contigs with the short reads using SPAdes[263], and the uses the long reads to bridge the gaps, repeats and unresolved junctions between the short-read contigs. Finally, the short reads are aligned to this draft (Pilon[264]) to refine small-scale mutations, insertions and deletions that may have been introduced by the noisy long reads. We visualized the resulting assembly graphs in Bandage[265], and in all cases observed two circular contigs corresponding to Chr. 1 and Chr. 2 (**Figure 4.5**). Together, these results confirmed that we could precisely split the *E. coli* genome into multiple pairs of chromosomes.

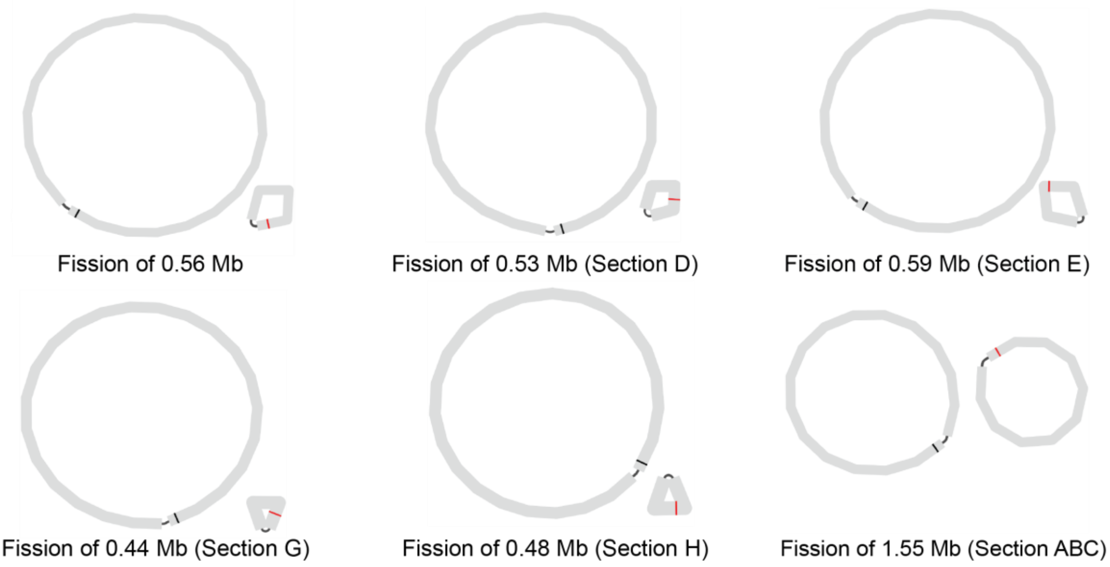

Fission of 0.56 Mb
Fission of 0.53 Mb (Section D)
Fission of 0.59 Mb (Section E)
Fission of 0.44 Mb (Section G)
Fission of 0.48 Mb (Section H)
Fission of 1.55 Mb (Section ABC)

Figure 4.5 - *De novo* assembly of fission experiments. Assembly graphs for the *de novo* hybrid assembly of the different post-fission chromosomal arrangements (**Figure 4.2**), generated by Unicycler[262] and visualized in Bandage[265]. Each assembly yielded one contig per replicon, corresponding to the expected post-fission Chr. 1 and Chr. 2. The regions corresponding to *oriC* and the BAC replication and segregation components (linker 2) are shown by black and red lines, respectively. Contigs corresponding to the helper plasmid or the spacer plasmid are not shown.

Characterization of post-fission clones

Unicycler calculates the relative copy number of the different contigs in an assembly by comparing their median read-depth. It assigns a depth of 1x to the largest contig, which serves as a reference to calculate the relative depth of the remaining contigs. All of our assemblies yielded one contig per replicon, and Chr. 1 was always the largest contig. This allowed us to estimate the copy number of BAC-supported Chr. 2 to be between 1 and 2; this is consistent with previous reports of BAC copy number[217] (**Appendix C.1**). We did not observe obvious biases in copy number related to the size or sequence of DNA carried by the BAC across fission experiments.

We further characterised our post-fission strains by measuring their growth rates in rich medium. We were surprised to find that despite the dramatic changes in genomic configuration that we induced, the effects of splitting the genome on fitness seemed relatively mild (**Figure 4.6**).

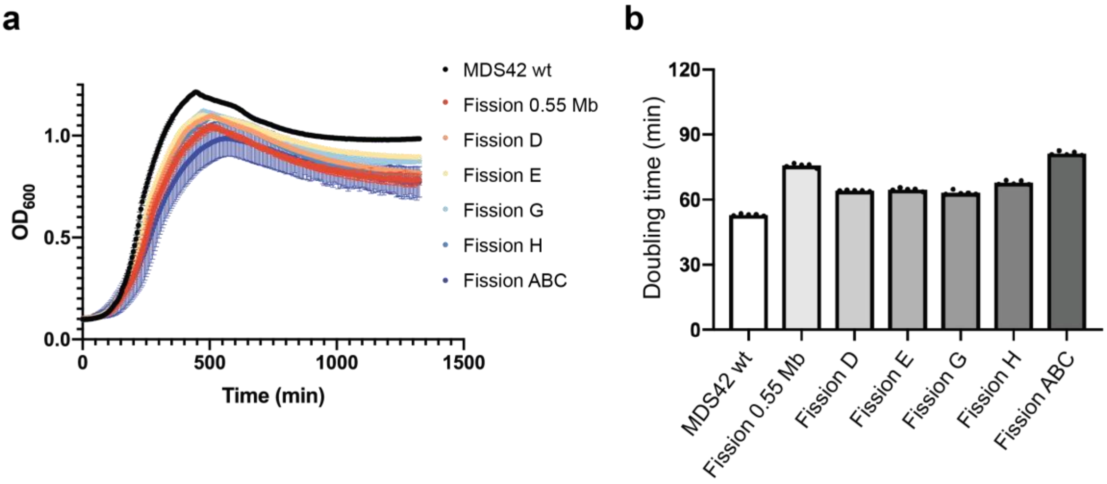
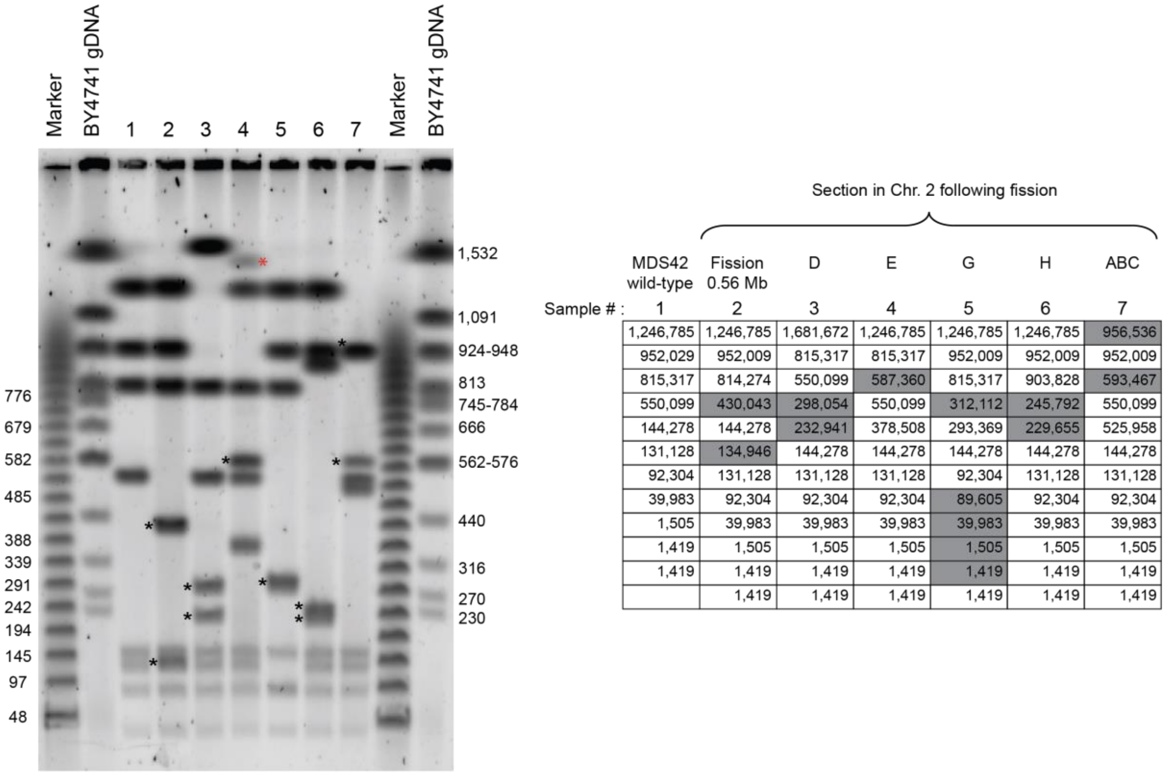


Figure 4.6 - Growth assays of post-fission strains. a) Growth curves of different post-fission strains (**Figure 4.2**). A pre-fission wild-type MDS42 control is shown in black. Individual data points represent the mean of 5 independently grown biological replicates, ±s.d. **b)** Maximum doubling times of different post-fission strains. Bars represent the mean of 5 independently grown biological replicates; individual data points are shown. The measured max. doubling time for MDS42 is 52.97 min ± 0.39; Fission 0.56 Mb, 75.76 min ± 0.8; Fission D, 64.18 min ± 0.21; Fission E, 64.67 min ± 0.61; Fission G, 63.1 min ± 0.85; Fission H, 67.83 min ± 0.83; Fission ABC, 81.28 min ± 0.81.

In order to assess the stability of the new chromosomal configurations, we grew the cells continuously in rich medium for 5 days, diluting 1/1500x every 12 hours. We hypothesised that, if these configurations were unstable, cells may revert to more favourable genomic arrangements and become over-represented in the population during continuous growth. After passaging, we prepared agarose plugs directly from the cultures and performed AvrII digestion followed by PFGE electrophoresis, as above. We found that for most fission experiments, the restriction patterns of the post-passaging populations (**Figure 4.7**) were identical to those of the individual starting clones (**Figure 4.4**). For fission of section D, we observed a faint band of ~1.4 Mb that was not predicted from its hypothetical chromosomal configuration, nor observed at the single clone level. It is possible that this band reflects the adoption of an alternative genome architecture in the passaging population, but PFGE does not provide the resolution to speculate about what these may be. This band may also result from a contamination - determining its origin will require further investigation. Overall, the alternative arrangements appear highly stable, and permit cell growth at practical rates.



Section in Chr. 2 following fission

	MDS42 wild-type	Fission 0.56 Mb	D	E	G	H	ABC
Sample # :	1	2	3	4	5	6	7
	1,246,785	1,246,785	1,681,672	1,246,785	1,246,785	1,246,785	956,536
	952,029	952,009	815,317	815,317	952,009	952,009	952,009
	815,317	814,274	550,099	587,360	815,317	903,828	593,467
	550,099	430,043	298,054	550,099	312,112	245,792	550,099
	144,278	144,278	232,941	378,508	293,369	229,655	525,958
	131,128	134,946	144,278	144,278	144,278	144,278	144,278
	92,304	131,128	131,128	131,128	92,304	131,128	131,128
	39,983	92,304	92,304	92,304	89,605	92,304	92,304
	1,505	39,983	39,983	39,983	39,983	39,983	39,983
	1,419	1,505	1,505	1,505	1,505	1,505	1,505
	1,419	1,419	1,419	1,419	1,419	1,419	1,419
		1,419	1,419	1,419	1,419	1,419	1,419

Figure 4.7 - PFGE Genomic stability. Pulse-field electrophoresis gel, showing AvrII restriction patterns of post-passaging genome fusion products. Each post-fusion product was passaged continuously for a week. Agarose plugs were then prepared from the resulting cultures, followed by AvrII digestion. The gel is labelled as in **Figure 4.4**. All post-passaging populations yielded the expected AvrII restriction patterns. Fission E (4 on the gel) shows an additional weak band at ~1.4 Mb, indicated by a red asterisk; the nature of this band is unclear.

Failed fission of 2 Mb

We attempted an additional fission experiment, where we aimed to generate 2 chromosomes, each ~2 Mb in size. We designed the boundaries such that Chr. 2 would contain sections ABC (**Figure 4.8**). We analysed several post-fission clones and found that the growth phenotypes and genotypes at j1 and j2 we consistent with genome fission. However, we performed hybrid *de novo* assemblies of two of these clones and found that their genome consisted of only 1 chromosome instead of 2. While the programmed recombination junctions were as expected, further recombination between Chr. 1 and Chr. 2 through rRNA operons appeared to have mediated reversion into a single chromosome. The two clones analysed were the product of two independent recombination events, as their genomic arrangements were different: the genome of clone 1 seemed to result from a rearrangement between rRNA operons H and B, whereas in clone 2 it seemed to have been caused by D and G. These recombinations may have occurred simultaneously to Cas9 cleavage and induction of lambda-red during the fission protocol, or during recovery to correct an unstable genomic architecture. This was the only instance in which fission failed to generate a split genome architecture.

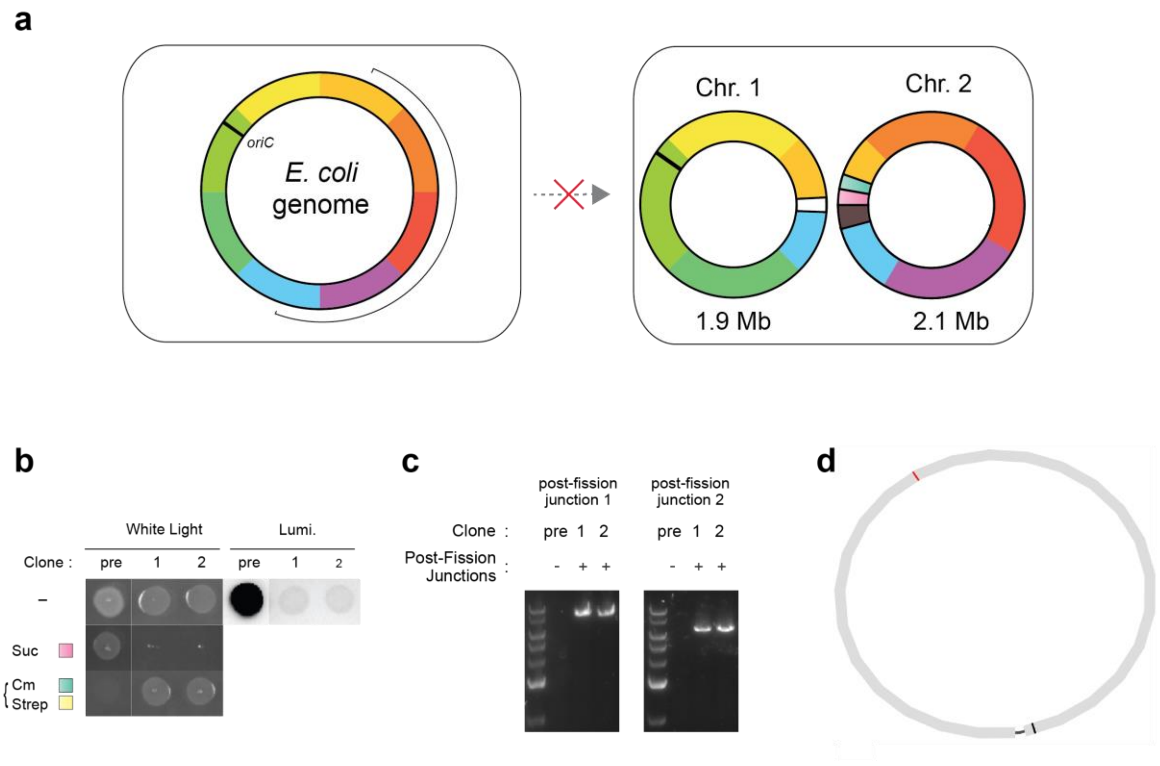


Figure 4.8 - Characterisation of a failed fission. a) Schematic of a failed fission attempt, intended to split the *E. coli* genome at the indicated sites into two chromosomes of ~2 Mb. In the design, Chr. 1 contains sections E, F, G and part of sections D and H; Chr. 2 contains sections A, B, C and part of sections H and D. The BAC replication and segregation components are shown in grey, linker 1 is shown in white. *oriC* is shown as a black line. **b)** Selective growth assays of post-fission clones 1 and 2 are consistent with the genome being split into two chromosomes. **c)** Junction-specific PCR yields the expected products for the new junctions j1

(continued from previous page) and j2 in post-fission clones, but not in pre-fission cells (pre). **d)** Representative *de novo* assembly graph of a post-fission clone, consisting of a single circular contig. *oriC* is shown as a black line, the BAC replication machinery (linker 2) is shown in red. In two independent instances, the genome assemblies showed that in post-fission cells linker 1 and linker 2 are flanked by the correct sequences, but additional recombination events between ribosomal operons precluded the separation of the genome into two chromosomes, and led to the maintenance a single-chromosome architecture.

Note: Fission experiments and characterization by PCR and selective growth assay were performed jointly by Kaihang Wang, Wesley Robertson and me. I performed all other experiments and characterization.

4.2.3 Genome fusion

Having demonstrated a method for the programmable splitting of the *E. coli* genome into multiple sets of two chromosomes, we next asked whether we could perform the reverse operation, and carry out the fusion of multiple chromosomes into a single chromosome. We envisioned that this, combined with the ability to precisely transfer chromosomes between strains, may enable the creation of chimeric genomes from multiple progenitors, and in user-defined configurations.

We devised a general workflow for fusing chromosomes. During fission, Chr. 2 is formed by the joining of linker 2 with fragment 2. We designated the two 50 bp termini of fragment 2 as HR5 and HR6. In the fission BAC, linker 2 encoded two cut sites; the PAMs were at the linker 2 termini, and the protospacers were encoded by HR2 and HR3. After fission, PAM sequences remain in linker 2, now adjacent to the termini of fragment 2. As a result, post-fission Chr. 2 naturally contains two new Cas9 cut sites, where the protospacers are encoded by HR5 and HR6. We can use these to release fragment 2 and expose HR5 and HR6 for recombination. We hypothesised that, by providing suitable HR5 and HR6 homologies in Chr. 1 plus appropriate Cas9 cut sites, we should be able to direct recombination between Chr. 1 and Chr. 2 to generate a single chromosome.

Fusion regeneration

To test the workflow, we first needed to perform a fission experiment; the resulting strain would serve as a common progenitor for multiple types of fission experiments, which will be described later on. We anticipated that, eventually, we may want to transfer the split chromosomes between strains. Consequently, we decided to perform fission on one of the intermediate strains from the synthetic genome assembly described in Chapter 2, where the ~500 kbp Section C had been recoded; we term this strain MDS42_SynC. The substitution of TCG, TCA and TAG codons by their synonyms in Section C would provide convenient high-density watermarks for differentiating this sequence from the genome of any future hosts.

We performed fission of Section C in MDS42_SynC, yielding a Chr. 1 of 3.45 Mb and a watermarked Chr. 2 of 0.54 Mb. After fission, we prepared for fusion by replacing the *sacB-cat* double selection marker in Chr. 2 with an *oriT-pheS*-kanR* cassette. This new marker retains the functionality of double selection, and allows for the conjugal transfer of Chr. 2 down the line. We performed this replacement with lambda-red recombination, selecting for the gain of *kanR* and the loss of *sacB* with kanamycin and sucrose, respectively (**Figure 4.9a**).

We used this strain to test the fusion workflow. First, we attempted to combine Chr. 1 and Chr. 2 into a single chromosome, re-generating the original genome of MDS42_SynC. After fission, Chr. 1 contains a Linker 2 sequence (*luxABCDE*), which bridges the gap of what used to be Section C. We used lambda-red mediated recombination to replace Linker 2 with a 'fusion sequence', which contains a *pheS*-hygR* cassette flanked by HR5 and HR6, as well as two Cas9 cut sites (**Figure 4.9b**).

We provided cells with suitable spacers and induced expression of Cas9 and lambda-red recombination machinery. Cas9 generates 4 cuts; two in Chr. 1 and two in Chr. 2. In both chromosomes, this exposes homologies HR5 and HR6, which are substrates for recombination by lambda-red. After outgrowth, selection with 4-chloro-phenylalanine for the loss of *pheS** recovers clones which have undergone recombination, and rearrangement of their genomes into a single chromosome.

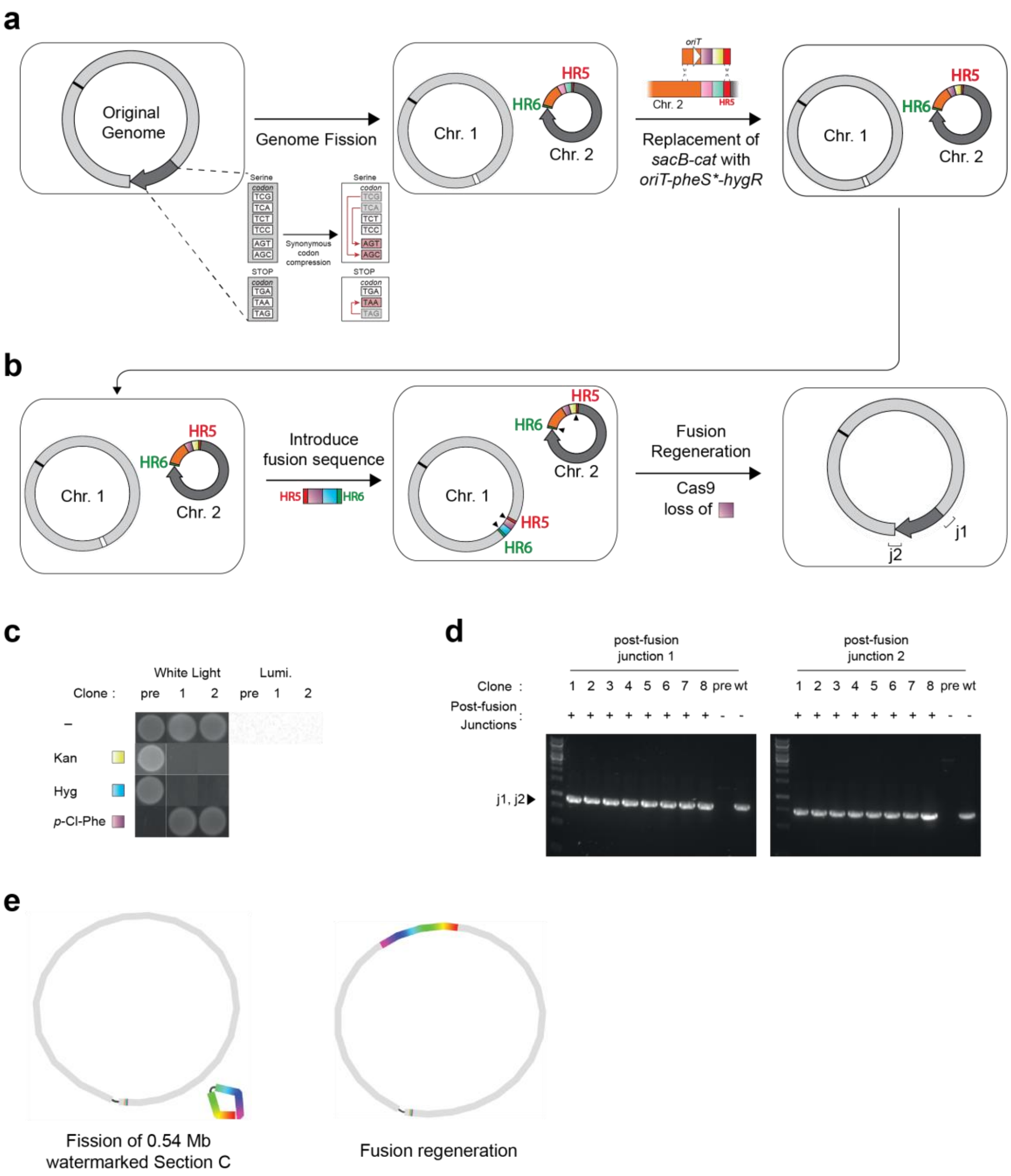


Figure 4.9 - Fusion regeneration. a) Schematic of genome fission and chromosomal fusion, performed in a strain where section C (dark grey) is recoded according to the synonymous codon compression scheme described in Chapter 3. The process starts with genome fission as in **Figure 4.1**; the termini of fragment 2 in Chr. 2 are marked as HR5 and HR6. *oriC* is indicated by a black line, linker 1 is in white, and the BAC replication and segregation components are in orange. After fission, the *sacB-cat* cassette (pink and green, respectively) is replaced by a *pheS*-kanR* cassette by lambda-red recombination. The cassette is flanked at its 5' by an *oriT* sequence (white arrow). **b)** A fusion sequence, consisting of a *pheS*-hygR* cassette (magenta and blue, respectively) flanked by HR5 and HR6 homology arms, is introduced in Chr. 1 by lambda-red recombination, replacing linker 1. Cas9 cleavage and lambda-red recombination are then induced. Selection against *pheS** with 4-chloro-phenylalanine isolates clones which have undergone cleavage and recombination. **c)** Selective

(continued from previous page) growth assay of post-fusion regeneration clones (1 and 2) is consistent with genome fusion. Pre-fusion cells (pre), which contain Chr. 1 and Chr. 2 and a fusion sequence replacing linker 1, are sensitive to 2.5 mM 4-chloro-phenylalanine (4CP) but resistant to 200 µg/mL hygromycin (Hyg) and 50 µg/mL kanamycin (Kan). The converse is observed in post-fusion clones. **d)** Junction-specific PCR of post-fission regeneration clones (1-8) regenerates the wild-type j1 and j2 products of the pre-fission strain (wt), and which were transiently lost in the post-fission, pre-fusion strain (pre). **e)** *De novo* assembly graphs of post-fission (left) and post-fusion regeneration cells (right). The sequence corresponding to recoded section C is indicated by rainbow colouring. The separate, short rainbow line indicates *oriC*. Fusion regeneration establishes the original genome architecture.

During the fusion process, the cell loses selection cassettes *pheS*-hygR* and *oriT-pheS*-kanR*. As a result, successful fusion clones should no longer be resistant to hygromycin nor kanamycin. We characterised post-fusion clones and verified that they exhibited the expected growth phenotypes in the relevant selective media (**Figure 4.9c**). MDS42_SynC originally contained two junctions, j1 and j2, between Section C and the remaining genome. These junctions were destroyed in the fission process. We performed PCR at j1 and j2 and observed that these were restored after fusion, consistent with the regeneration of the original genome (**Figure 4.9d**). Finally, we performed a hybrid *de novo* assembly and confirmed that the architecture of the MDS42_SynC genome had been restored as expected (**Figure 4.9e**). Notably, this process was scarless; the product of fusion regeneration was exactly the same as MDS42_SynC pre-fission an fusion.

Fusion translocation

We showed that we could re-direct fusion between the two chromosomes by providing a fusion sequence, which contains suitable homologies to mediate recombination between Chr. 1 and Chr. 2 (HR5 and HR6). Because we artificially introduce this sequence after fission by lambda-red mediated recombination, we hypothesised that by introducing the fusion sequence elsewhere in the genome we may be able to program the site of fusion between Chr. 1 and Chr. 2. This would result in translocation of the sequence encoded in Chr. 2, leading to genome rearrangement. Similarly, providing a fusion sequence in the reverse orientation may allow us to perform inversions.

We set out to test fusion-mediated genome rearrangements, starting with translocation of Section C. We performed this in a post-fission MDS42_SynC strain where Chr. 2 contains an *oriT-pheS*-kanR* cassette, as above. We integrated the same fusion sequence as above 500 kbp, 700 kbp away from Linker 2 in Chr. 1 (**Figure 4.10a**). Because the Cas9 cuts are conserved with respect to the fusion re-generation experiment, we could perform the fusion protocol exactly as above (**Figure 4.9**), providing an identical set of spacers for Cas9 cleavage.

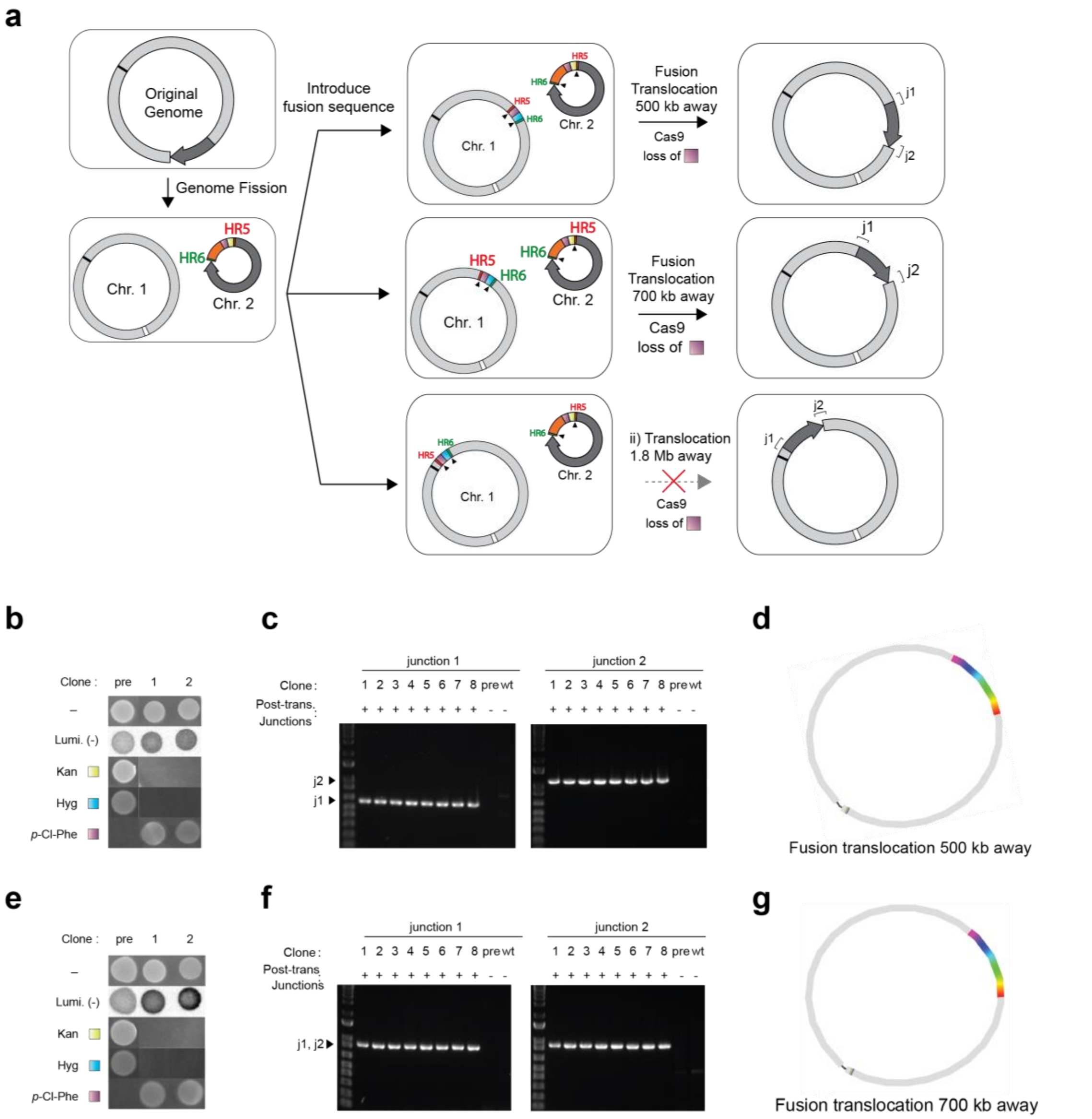

a
Original Genome
Genome Fission
Introduce fusion sequence
Chr. 1
Chr. 2
HR5
HR6
Fusion Translocation 500 kb away
Cas9 loss of
Fusion Translocation 700 kb away
Cas9 loss of
ii) Translocation 1.8 Mb away
Cas9 loss of
j1
j2
b
Clone: pre 1 2
–
Lumi. (-)
Kan
Hyg
p-Cl-Phe
c
junction 1
junction 2
Clone: 1 2 3 4 5 6 7 8 pre wt
Post-trans. Junctions + + + + + + + + - -
j2 ▸
j1 ▸
d
Fusion translocation 500 kb away
e
f
j1, j2▸
g
Fusion translocation 700 kb away

Figure 4.10 - Chromosomal fusion mediates the translocation of genomic segments. a) Schematic of fusion-mediated translocation. Fusion occurs from the same precursor as in **Figure 4.9**, with a recoded section C in Chr. 2 and the *sacB-cat* replaced with *oriT-pheS*-kanR*. A fusion sequence is introduced at multiple sites in Chr. 1 by lambda-red recombination. Subsequent Cas9 cleavage and lambda-red recombination drive the joining of Chr. 1 and Chr.

(continued from previous page) 2 at diverse loci within Chr. 1, placing section C away from its original position. Linker 1 (*luxABCDE*) remains at the original fission junction. Two translocations at 500 kbp and 700 kbp from linker 1 were successful. One attempted translocation 1.8 Mb resulted in instability. **b)** Selective growth assay of post-fusion translocation 500 kbp away from the original location. The observed growth phenotypes of post-fusion clones (1 and 2) are consistent with chromosomal fusion, losing sensitivity to 2.5 mM 4CP and resistance to 50 µg/mL Kan or 200 µg/mL Hyg. Because linker 1 (*luxABCDE*) is not replaced by the fusion sequence in fusion translocation, post-fission (pre) and post-fusion clones retain luminescence (Lumi.). **c)** Junction-specific PCR of post-fusion translocation products (1-8) 500 kbp away yields the expected products at j1 and j2, whereas these are not present in the pre-fusion (pre) or the wild-type pre-fission (wt) strains. **d)** *De novo* assembly graph of post-fusion translocation products 500 kbp away from linker 1. Recoded section C is indicated by rainbow colouring; the short rainbow line corresponds to *oriC*. The graph shows a single circular contig corresponding to the expected genome architecture. **e)** Selective growth assay post-fusion translocation 700 kbp away from the original location, as in b). **f)** Junction-specific PCR of post-fusion translocation products 700 kbp away, as in c). **g)** *De novo* assembly graph of post-fusion translocation products 700 kbp away from linker 1, as in **d)**.

We confirmed that post-fusion translocation clones displayed the expected growth phenotypes in selective media (**Figure 4.10b, e**). The criteria for assessment is the same as for fusion regeneration, with the exception that Linker 2 (*luxABCDE*) is not removed in preparation for fusion, as the fusion sequence is integrated elsewhere in the genome. Consistently, post-fusion rearrangement cells retained bioluminescence from Linker 2. We performed PCR across the new junctions and found that cells contained the expected products, which were not present in the pre-fusion strain or the wild-type MDS42_SynC strain (**Figure 4.10c, f**). *De novo* assemblies of post-fusion genomes confirmed the translocation of Section C to the defined target sites (**Figure 4.10d, g**).

We performed an additional attempt, where we aimed to translocate Section C 1.8 Mb away from its original locus, at the junction between sections F and G. The phenotype of the post-fusion clones was as expected, and *de novo* assembly of one of these clones showed a single chromosome configuration that resembled the expected architecture. However, closer inspection of this genome assembly revealed the presence of a ~40 kbp deletion spanning genes *yfbB* to *frvR*, and which erased one of the expected new junctions (j2). From this, it seems likely that our protocol generated the expected fusion arrangement using the homologies we provided, but that the resulting configuration was unstable and resulted in deletion of this region.

Note: I performed fusion regeneration and translocations 500 and 700 kbp away. Translocation 1.8 Mb away was performed jointly with Wesley Robertson. I performed all characterization of the resulting strains.

Fusion inversion

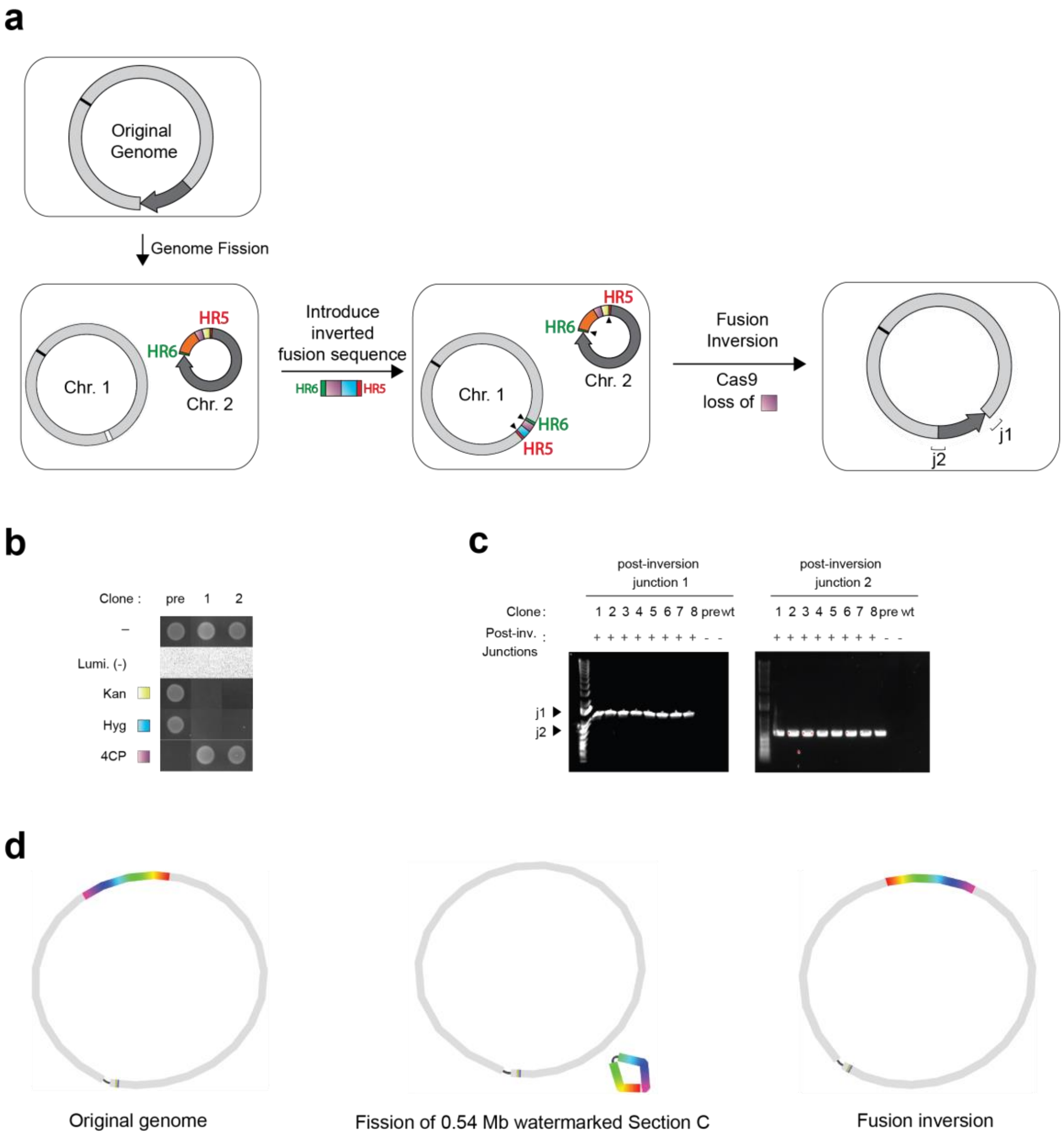

a
Original Genome
Genome Fission
HR5
HR6
Chr. 1
Chr. 2
Introduce inverted fusion sequence
HR6
HR5
HR5
HR6
Chr. 1
Chr. 2
HR6
HR5
Fusion Inversion
Cas9 loss of
j1
j2
b
Clone : pre 1 2
–
Lumi. (-)
Kan
Hyg
4CP
c
post-inversion junction 1
post-inversion junction 2
Clone : 1 2 3 4 5 6 7 8 pre wt
Post-inv. Junctions : + + + + + + + + - -
j1
j2
d
Original genome
Fission of 0.54 Mb watermarked Section C
Fusion inversion

Figure 4.11 - Chromosomal fusion mediates the inversion of genomic segments. a) Schematic of fusion inversion. The starting strain is as in **Figure 4.9** and **Figure 4.10**. A fusion sequence is introduced by lambda-red recombination in an inverted orientation with respect to **Figure 4.9**. Cas9 cleavage and lambda-red recombination generate a strain where recoded section C is at the same locus as the original genome, but in the reverse orientation. **b)** Selective growth assay is consistent with chromosomal fusion in post-fusion inversion clones (1 and 2), and a two-chromosome arrangement in pre-fusion cells (pre). **c)** Junction-specific PCR of post-fusion inversion clones (1-8) yields the expected products for new junctions j1

(continued from previous page) and j2, which are not present in the wild-type (wt) or the pre-fusion (pre) cells. **d)** *De novo* assembly graphs of post-fusion inversion clones. Graphs of the original MDS42_SynC genome and post-fission of recoded section C are provided for comparison. Recoded section C is indicated by rainbow colouring, magenta being the 5' and red being the 3'. The directionality of section C in the original genome versus the fusion inversion graph shows inversion of the region. The short rainbow line corresponds to *oriC* (5' magenta, 3' red).

We next tested inversion of Section C. We introduced a fusion sequence at the same locus as in the fusion regeneration experiment (replacing Linker 2), but in the opposite orientation (**Figure 4.11a**). We carried out the fusion protocol and characterised the cells by evaluation of the growth phenotypes in selective media (**Figure 4.11b**) and junction-specific PCR (**Figure 4.11c**), as above. Both were consistent with inversion of the region, and this was further confirmed by *de novo* genome assembly.

We analysed the fitness of clones bearing all the different arrangements we generated by fission and fusion in MDS42_SynC. The doubling time of the post-fission strain, where watermarked Section C is in Chr. 2, increased from 60 to 107 min (**Figure 4.13a**). Upon regeneration of the original genome by fusion, the growth rate was rescued completely. This was not the case when we performed fusion translocations and inversions. Interestingly, the effect of translocating Section C 500 kbp away from its original locus, but in the same orientation, was less detrimental to fitness than inverting the region. The fitness of the Section C translocation 700 kbp away was comparable to that of the split-chromosome arrangement.

Note: Fusion inversion was performed jointly with Wesley Robertson. I performed all characterizations.

4.4 – Generating chimeric genomes from diverse progenitors

Note: Fission and fusion experiments in this section were performed by Kaihang Wang, and are discussed here for completeness. I performed analysis of fitness and genome architecture.

We had demonstrated technologies for performing precise fission and fusion of the *E. coli* genome. In general, pipelines for the stepwise assembly of genomes, including the one described throughout Chapter 3 of this dissertation, often rely on the progressive assembly of genome fragments in intermediate strains. These fragments have to be somehow combined into a single, final genome, and we thought fission and fusion may be useful in executing this type of assembly. The work in Chapter 3 provided a series of partially synthetic strains. These could serve as convenient starting points; we could test whether fission and fusion could mediate the combination of multiple synthetic sections within a single strain.

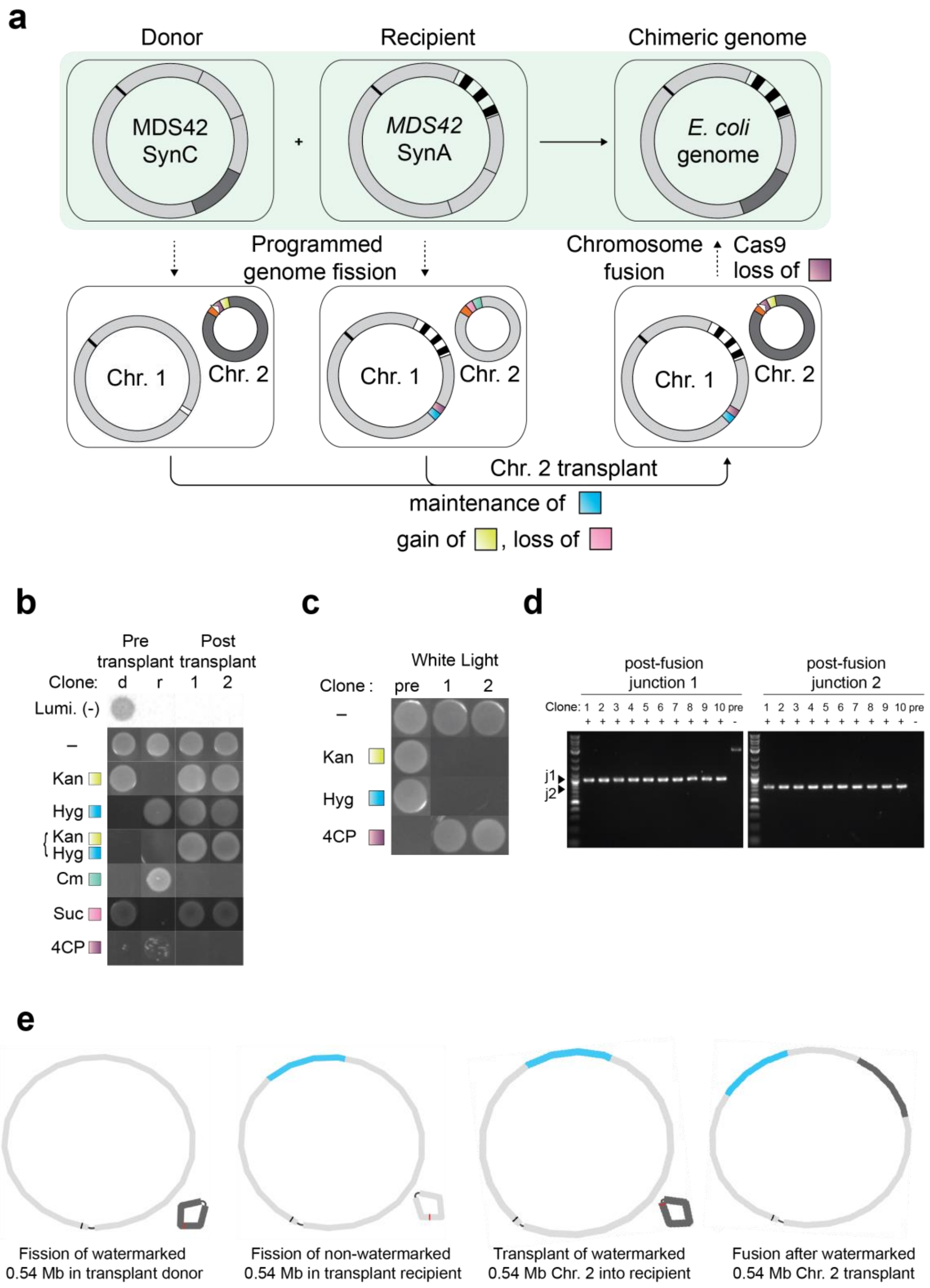
We started with two strains; one contains a recoded Section C (MDS42_SynC), and another contains a recoded Section A (MDS42_SynA). Our goal was to generate a single strain where both Section A and Section C were recoded. We defined MDS42_SynC as the donor strain, and MDS42_SynA as the recipient.

In the donor, a fission of recoded Section C had already been performed, as described in the above section. The *sacB-cat* in Chr. 2 had been replaced by an *oriT-pheS*-kanR* cassette. We performed an analogous fission of Section C in the recipient, where it is wild-type, but did not alter the resulting *sacB-cat* in Chr. 2 (**Figure 4.12a**). We confirmed fission in the recipient strain by stamping in selective media and genotyping the new junctions (data not shown), as well as *de novo* assembly (**Figure 4.12e**). In the recipient however, we replaced linker 2 with a fusion sequence, as in the fusion regeneration experiment.

We then transformed a non-transferrable F' plasmid into the post-fission donor, and performed directed conjugation between donor and recipient strains. We spread the cells in medium containing kanamycin, hygromycin, and sucrose, and the differential marker identities between the two strains allowed us to select for the desired chromosome combinations. Hygromycin ensured that we only recovered recipient cells. Kanamycin enforced uptake of the donor recoded Chr. 2, whereas sucrose selected for the loss of the recipient non-recoded Chr. 2.

The selective growth phenotypes were consistent with transplant of Chr. 2 (**Figure 4.12b**). We performed *de novo* genome assembly on one post-transplant clone, and confirmed that chromosomal transplant had occurred (**Figure 4.12e**). The recoding of Sections A and C allowed us to determine that, in the post-transplant cell, Chr. 1

derived from the recipient, and Chr. 2 derived from the donor. We did not observe any crossovers between the chromosomes.



Fission of watermarked 0.54 Mb in transplant donor

Fission of non-watermarked 0.54 Mb in transplant recipient

Transplant of watermarked 0.54 Mb Chr. 2 into recipient

Fusion after watermarked 0.54 Mb Chr. 2 transplant

Figure 4.12 - Genome fission, chromosome transplant and chromosome fusion mediate the generation of chimeric genomes. a) Fission of section C was performed in parallel in

(continued from previous page) donor MDS42_SynC (where section C is recoded, dark grey), and recipient MDS42_SynA (where section A is recoded, black/white stripes). In the donor, *sacB-cat* in Chr. 2 is replaced by *oriT-pheS*-kanR* as in **Figure 4.9**. In the recipient, linker 1 is replaced by a *pheS*-hygR* cassette in Chr. 2. After conjugation, post-transplant cells are immediately amenable for chromosome fusion, generating a single circular chromosome where recoded sections C and A are combined. **b)** Selective growth assay of chromosomal transplant. Pre-transplant donor (d) and recipient (r) cells have growth phenotypes consistent with their selectable markers in Chr. 1 and Chr. 2. Post-transplant clones (1 and 2) gain resistance to 50 μg/mL kanamycin (Kan) from donor Chr. 2 and maintain resistance to 200 μg/mL hygromycin (Hyg) from recipient Chr. 1. They are also resistant to 7.5% sucrose (Suc) and 20 μg/mL chloramphenicol (Cm) due to loss of recipient Chr. 2. Sensitivity to 2.5 mM 4-chloro-phenylalanine (4CP) is maintained throughout. **c)** Selective growth assay of chromosomal fusion after chromosome transplant. Post-fusion clones exhibit growth phenotypes consistent with fusion, as in **Figure 4.9**. 'Pre' indicates pre-fission clones. **d)** Junction-specific PCR of post-fusion clones (1-10) shows the expected products at new junctions j1 and j2, whereas these are not present in the pre-fusion (pre) control. **e)** *De novo* assembly graphs of chromosomal transplant strains. Recoded section C is shown in grey, recoded section A is shown in blue.

The configuration of this strain was directly amenable to performing fusion. We induced expression of lambda-red components and Cas9, and provided suitable spacers exactly as in the fusion regeneration described in the previous section. The result of this process was a genome consisting of a single chromosome, where both Section A and Section C were recoded (**Figure 4.12c, d, e**). This demonstrated that our set of technologies could be used to combine sections of genomes with high efficiency and precision.

We measured the maximum doubling times of the intermediate strains for chromosomal transplant. Comparison of doubling times for MDS42_SynC and MDS42_SynA before and after fission of section C shows that the fitness defect is much smaller in MDS42_SynA. The starting fitness of MDS42_SynA is also higher compared to MDS42_SynC (**Figure 4.13b**). This suggests that recoding of section C results in decreased fitness, and that this effect is exacerbated in a two-chromosome configuration. Fusion of recoded sections A and C into a single chromosome restores fitness substantially.

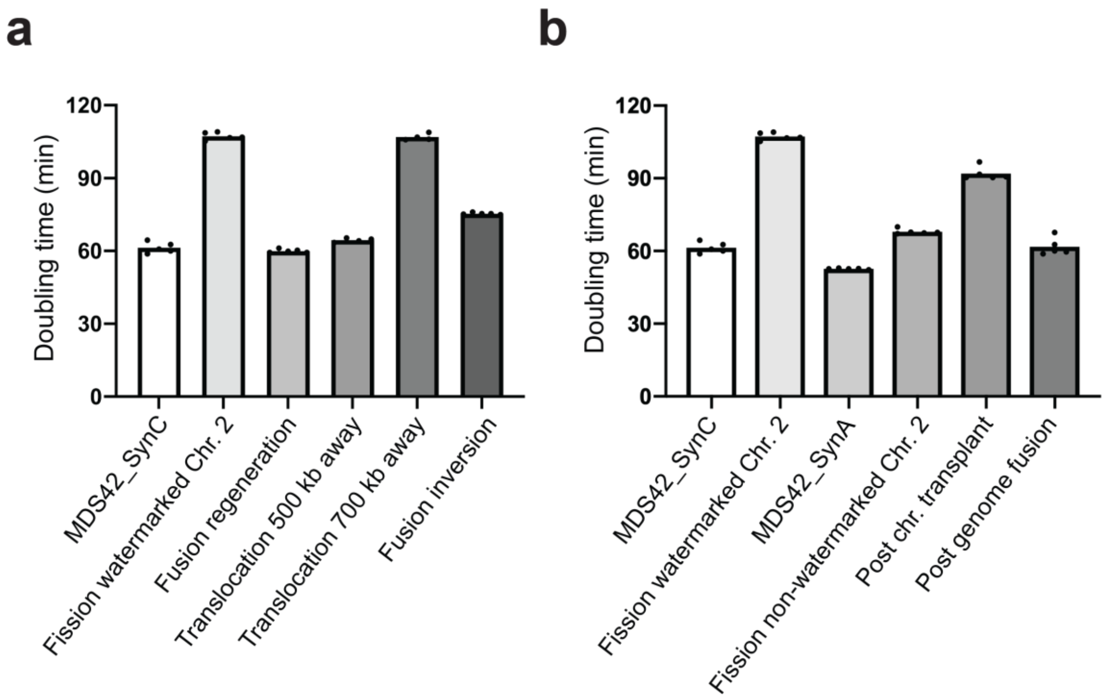


Figure 4.13 - Fitness of different fusion products. All growth rates were measured in 100 μg/mL streptomycin**. a)** Maximum doubling times of chromosomal fusion strains. Fission of recoded section C ('watermarked Chr. 2') results in a significant increase in doubling time (107.30 min ± 1.39) with respect to the progenitor MDS42_SynC (61.37 min ± 1.97); fusion re-generation completely restores fitness (59.94 min ± 0.67). Other fusions result in different doubling times; 64.45 min ± 0.70 for translocation of section C 500 kbp away, 106.93 ± 1.19 for translocation 700 kbp away, and 75.4 min ± 0.36 for inversion of section C. **b)** Maximum doubling times for strains involved in chromosomal transplant. The doubling times for the starting trains are 61.37 min ± 1.97 for MDS42_SynC and 64.45 min ± 0.70 for MDS42_SynA. The fitness effect of fission of section C in MDS42_SynA (67.98 min ± 1.01, 'Fission non-watermarked Chr. 2') is minor compared to fission in MDS42_SynC (106.93 ± 1.19, 'Fission watermarked Chr. 2'). Fitness decreases when non-recoded Chr. 2 from MDS42_SynA is replaced by recoded Chr. 2 from MDS42_SynC (91.95 ± 2.43, 'Post chr. Transplant'), and is partly restored after genome fusion (61.77 ± 3.20, 'Post genome fusion').

4.5 – Discussion

We demonstrated a set of genome engineering tools to split, re-arrange and assemble the genome of *E. coli*. We could program the division of the single *E. coli* chromosome into multiple pairs of chromosomes with nucleotide resolution. Performing genome fission did not require any prior modifications to the genome, and relied only on a helper plasmid encoding Cas9 and lambda-red recombination, a fission BAC, and a set of spacer sequences to direct Cas9 cuts.

Pairs of chromosomes resulting from fission experiments were amenable to recombination back into a single chromosome architecture. The ability to define the precise site and orientation of the recombination event allowed us to perform translocations and inversions with high precision. Moreover, by appropriately placing selectable markers in the chromosomes of separate strains, we could shuffle chromosomes between cells. We used this method to combine two recoded ~500 kbp sections from distinct strains into a single cell, and consolidated both into a single chromosome through chromosome fusion.

In Chapter 3, we described an approach for the convergent assembly of multiple partially synthetic genomes into a single strain. Similarly to previous work[100], we prepared the genome by integrating *oriT* sequences and selection markers, to induce direct genome-to-genome transfer of DNA. While this approach clearly is useful in the assembly of genomes, it has shortcoming - perhaps the most important of these is the inability to precisely define the boundaries of transfer. This is exemplified in the last conjugation step to assemble Syn61, where we needed to perform extensive genome preparations to provide a recoded homology tail where recombination could occur (3.2.4).

Here, by splitting the genome, we topologically isolate a region of interest into a separate replicon within the cell. In this form, conjugal transfer can occur similarly to natural plasmids, which do not rely on integration on the host genome and are simply re-circularized in the cytoplasm of the recipient cell. Because we can specify the nucleotides that constitute the splitting boundaries, this effectively allows us to define a region to be transferred between cells with much higher precision than conjugation. Moreover, as homologous recombination is now decoupled from the transfer process, in principle we retain the flexibility to re-integrate the transferred DNA anywhere in the genome, as long as the final arrangement is viable.

Yeast has been the platform of choice for performing genome assemblies; the high level of endogenous recombination in this organism facilitates genome manipulations. However, as described in Chapters 2 and 3 of this thesis, the transfer of genomic DNA from yeast to other bacteria (except for perhaps *Mycoplasma*) suffers from size limitations inherent to the competence of the target hosts. The general framework presented here, combining Cas9 and lambda-red to program recombination at defined sites *in vivo*, endows *E. coli* with some of the tools that are already naturally present in yeast, and may help establish it as an organism of choice in future genome assembly projects. In fact, in this framework recombination is inducible, which may help avoid unwanted rearrangements during propagation. Moreover, conjugation does not suffer from the same size limitations as electroporation or PEG-mediated transformation, meaning that the resulting DNA constructs could be directly transferred to bacterial hosts of interest, bypassing any yeast intermediates.

In *E. coli*, DNA replication begins at *oriC*, and proceeds bidirectionally until the two replication forks meet somewhere near the terminus region, which is diametrically opposite to *oriC*[266]. It has been shown that significant differences in the length of the replichores can be detrimental to cell fitness[267]. Here, each fission of a ~500 kbp section results in shortening of one replichore approximately 25%. This may account (at least in part) for the drop in fitness of post-fission strains, although the deleterious effects are minor, and the strains grow sufficiently well to act as useful intermediates in genome assembly.

The *E. coli* genome is thought to be divided into four macrodomains; Ori (around *oriC*), Left, Right and Ter (near the terminus region)[268,269]. These domains are spatially separated during the cell cycle. The terminus domain is scattered with *ter* sites (**Figure 4.2**), which bind the protein *tus*[266]. *Ter/tus* complexes allow progression of replication forks in only one orientation. These have been proposed to promote convergence of replication forks in between *terA* and *terC* where termination can occur; this region is called 'replication fork trap'. However, the biological significance of this site is unclear, as *tus* is not essential, and other bacteria such as *Vibrio cholera* lack a *tus/ter* system or equivalent[270]. According to this model, when performing translocations we may be introducing barriers to the progress of replication forks near the origin; this may effectively render one replichore much shorter than the other, or prevent convergence of replication forks. This may be one of the reasons underlying instability of the translocation of section C 1.8 Mb away (near the origin) from its original location. Under these conditions, inactivation of *tus* may alleviate stalling of replication forks and promote stability. The precise mechanisms underlying replication termination are still subject to debate, but are thought to involve an over-replicated intermediate that is further processed by helicase-nuclease SbcDC, exonuclease ExoI and recombination components recBC[271].

Here we performed fission of sections ABC, which effectively transfers the entire Ter domain to a separate replicon; this includes all *ter* sites except for one, and the *dif* site, which is normally placed in the middle of the replication fork trap. *dif* is thought to mediate the resolution of chromosome dimers and segregation, by the action of Xer recombinases[272]. Strains bearing *dif* mutations or re-locations have been shown to display segregation phenotypes[273]. While we did not thoroughly investigate cell division in the resulting strains, we showed that the strain bearing a fission of ABC grew stably under our conditions. A larger fission which also encompassed the Ter domain was unsuccessful, as cells reverted back into a single genome configuration through recombination at ribosomal operons. This may have been due to instability of the resulting chromosomes, perhaps due to termination or segregation defects, but the mechanistic causes are unclear. Alternatively, it may be due to the inability of BAC origins to support more than 2 Mb of DNA. Importantly, much of the literature concerned with bacterial chromosome structure relied exclusively on inversions and deletions to probe chromosome structure[267–269,273]. The tools provided here could allow scientists to test the phenotypic consequences of a wider range of perturbations, and allow them to re-localize entire chromosomal domains.

Here, we have only explored fission into two chromosomes. However our system should, in principle, be compatible with the generation of multiple chromosomes within the same cell. A limiting factor may be the availability of double selection cassettes to perform iterative fissions. Removing the double selection cassette in Chr. 2 after genome splitting would enable its re-use in a subsequent round of fission. Each new replicon will require its own origin of replication. Throughout this dissertation, we have generated multiple *E. coli* strains that stably propagate a BAC-based replicon (bearing recoded fragments in Chapter 3, Chr. 2 in Chapter 4) and a fertility F' plasmid; the origins of these two replicons derive from fertility factors[217]. Hence, it appears that at least two F'-derived chromosomes can co-exist within the same cell, as long as there is selection for both. Future work on generating multiple sub-chromosomes may also rely on the import of chromosomal replication origins from heterologous organisms such as *Vibrio cholerae*, which have been shown to be functional in *E. coli*[259,274,275].

The set of technologies described here may provide a foundation for the re-organization of genomes into modular compartments. This may eventually facilitate the exchange of entire pathways or multi-gene functional modules with engineered and evolved counterparts, and between different organisms.

In the future, these tools may be synergistic with the development of orthogonal DNA replication systems[136]. The ability to target defined sections of the genome and confine them to a separate replicon, replicated by an engineered orthogonal chromosome-specific polymerase, could allow for modulating the rates of replication error of the different chromosomes within the same cell. This may eventually be useful as a platform for *in vivo* directed evolution of subsets of genes or pathways of interest.

In conclusion, this dissertation has provided a set of technologies for the creation of designer genomes. These technologies have been used to create a recoded *E. coli* genome with only 61 codons in annotated genes, and the liberated codons have been shown to be amenable to reassignment into unnatural amino acids. This has been complemented with the development of several genome engineering operations that further extend our capacity to manipulate the DNA of living organisms. The workflows and lessons derived from this work may guide future genome synthesis efforts.

Chapter 5 - Experimental procedures

Where appropriate, sections of this chapter are edited or reproduced with permission from the following publications:

Fredens, J.*, Wang, K.*, **de la Torre, D**.*, Funke, L. F. H.*, Robertson, W. E.*, Christova, Y., Chia, T., Schmied, W. H., Dunkelmann, D. L., Béranek, V., Uttamapinant, C., Gonzalez Llamazares, A., Elliott, T. S. & Chin, J. W. Total synthesis of *Escherichia coli* with a recoded genome. *Nature* **569,** 514–518 (2019).

*equal contribution

Wang, K., **de la Torre, D**., Robertson, W. E. & Chin, J. W. Programmed chromosome fission and fusion enable large-scale genome rearrangement and assembly. *Science* **365,** 922-926 (2019).

5.1 – General procedures

Note: The list of primers used for cloning intermediate plasmids, genotyping of BAC construction, junction-specific PCRs in REXER for recoded sections and DGF-327 replacements, fission, fusion, etc. is too vast to be reasonably included here. Large datasets and primer spreadsheets are available in digital format from the original publications, as indicated throughout the text. Accession numbers of relevant plasmids and genome sequences are provided. All other sequences are available from the author upon request.

The positive-negative selection cassettes used throughout this dissertation are -1/+1 (*rpsL-kanR*), -2/+2 (*sacB-cat*) and -3/+3 (*pheS*-hygR*)[276]. In each cassette, both proteins are expressed polycistronically under control of the EM7 promoter (*sacB-cat* and *pheS*-hygR*) or the *rpsL* promoter (*rpsL-kanR*).

Unless explicitly indicated in each section, all lambda-red recombination, REXER, fission and fusion were carried out using a helper plasmid which contains the lambda-red recombination machinery, Cas9 and a copy of tracrRNA (Genbank accession MN927219).

5.1.1 Preparation of competent cells

For each experiment, overnight cultures of the cells of interest (from either glycerol stocks or individual colonies) were started in 10 mL LB supplemented with the relevant antibiotics. If the helper plasmid was present in the cells and it needed to be maintained for subsequent recombination, 5 µg/mL tetracycline and 2% glucose were supplemented. Additionally, for cells containing *rpsL-kanR* cassettes, 50 µg/mL kanamycin was provided. For cells containing *sacB-cat* cassettes, 20 µg/mL chloramphenicol was provided.

Overnight cultures were diluted 1:50 (OD_{600}=0.05) in 500 mL of LB supplemented with the same antibiotics as above, at 37 °C while shaking at 200 rpm. Once OD_{600} reached 0.4-0.6, cells were harvested by centrifugation for 10 minutes at 4000 x g, 4 °C. Cells were washed twice in 40 mL ice-cold Milli-Q H_2O, centrifuging as before. An additional wash was performed with 40 mL of ice-cold 16% glycerol. After centrifugation, the cell pellets were resuspended in 1 mL ice-cold 16% glycerol. At this point cells were either used directly for electroporation or snap-frozen in liquid nitrogen and stored at -80 °C.

5.1.2 Lambda-red recombination

Lambda-red recombinations were performed in cells that contained the helper plasmid, which encodes Cas9 and the lambda-red recombination machinery. 10 mL overnight cultures of the relevant cells were prepared as in 5.1.1, omitting glucose. Overnight cultures were diluted 1:50 (OD_{600}=0.05) in 500 mL of LB supplemented with 5 µg/mL tetracycline plus the relevant antibiotics, without glucose. Cells were incubated at 37 °C while shaking at 200 rpm for 1-2 hours, until OD_{600} reached 0.2-0.25. At this point 2.5 g of L-arabinose powder were added to the culture, for a final concentration of 0.5%. Cells were incubated for ~1 more hour in the same conditions. Cells were then harvested, and washed twice with Milli-Q and once with 16% glycerol as in 5.1.1. Bacterial pellets were resuspended in 1 mL 16% glycerol, and either used directly or stored at -80 °C for later use.

Generally, for lambda-red recombination experiments 100 µL of electrocompetent, arabinose-induced cells were electroporated with 2-8 µg of the relevant PCR product containing suitable homology regions, purified with QIAgen QIAquick Spin Columns as per manufacturer's instructions. The nature of the particular PCR products used in each experiment is described in the relevant sections.

5.1.3 Preparation of *S. cerevisiae* spheroplasts

S. cerevisiae BY4741 spheroplasts were prepared as described by Kouprina *et al.*[277]. Spheroplasts were co-transformed with multiple linear dsDNA pieces, depending on the type of assembly (specified in each section). PEG-mediate transformation was carried out as in ref[277].

5.1.4 Construction of spacer plasmids for REXER, genome fission and fusion

Spacer plasmids encode protospacers in an architecture that mimics natural CRISPR arrays[185]. Generally, the plasmids contain a pMB1 origin of replication plus an ampicillin resistance marker. The three variations used throughout this work are pKW1_MB1Amp_TracrK_Spacer (pKW1) (GenBank accession MK809152.1) pKW3_MB1Amp_TracrK_Spacer (pKW3; GenBank accession MN226641.1) and pKW5_MB1Kan_TracrK_Spacer (pKW5; GenBank accession MN226642.1). For each experiment, these are referred to as pKW1 Spacers, pKW3 Spacers or pKW5 Spacers. pKW1 contains only the three components described above. pKW3 contains an additional tracrRNA sequence upstream the CRISPR array, and was derived from pKW1 by Gibson assembly. pKW5 contains *kanR* instead of *tetR*, which renders it resistant to kanamycin instead of tetracycline, and was derived from pKW3 by Gibson assembly.

Each experiment needs a unique set of spacers; plasmids for REXER 2 contain two protospacers, REXER4 and genome fusion experiments required 4 protospacers, and genome fission experiments require 6 protospacers. For each experiment, the arrays were constructed by successive rounds of primer extension and PCR, illustrated in the figure below. The resulting inserts were then assembled with AccI and EcoRI digested pKW1/3/5 backbones by Gibson assembly. Spacer arrays were verified by Sanger sequencing to be free of mutations.

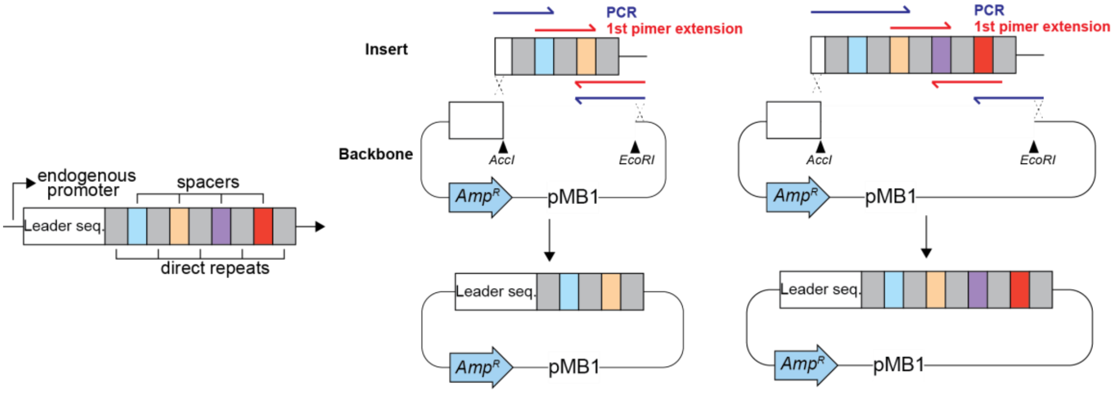


Figure 5.1 - Construction of spacer plasmids for REXER, fission and fusion

Appendix A.4 details the spacer sequences used for REXER in Chapter 2.

The spacer sequences used for REXER in Chapter 3 are detailed in Supplementary Data 9 of ref[185].

The spacer sequences used for genome fission and fusion in Chapter 4 are detailed in Table S1 of ref[278].

5.1.5 REXER

For each REXER, target cells containing the helper plasmid and a double selection cassette at the beginning of the fragment of interest (locus^0) were made competent as in 5.1.1, selecting with 5 µg/mL tetracycline throughout. 100 µL of competent cells were mixed with 5-20 µL of purified BAC DNA and subject to electroporation. Cells were recovered in 1 mL SOC for 1:30 hours, and spread in LB agar plates supplemented with 5 µg/mL tetracycline plus 50 µg/mL kanamycin (*rpsL-kanR*) or 20 µg/mL chloramphenicol (*sacB-cat*). Plates were incubated overnight at 37 ˚C.

A colony was picked from the BAC transformation plate and grown overnight in 10 mL LB supplemented with 5 µg/mL tetracycline plus 50 µg/mL kanamycin or 20 µg/mL chloramphenicol. Cells were made electrocompetent with 0.5% L-arabinose induction for 1 hour as in 5.1.2. If using spacers in a plasmid form, 2-8 µg of purified pKW1/pKW3 spacer plasmid were electroporated. If using a linear product, primers Spacer_F (5' GTTTTATTTGATGCCTCTAGCACGCGTACCATG 3') and Spacer_R (5' GATACTGAGCACATCAGCAGGACGCAC 3') were used to amplify the spacer cassette by PCR using the corresponding pKW1 or pKW3 plasmid as template. The PCR was treated with DpnI for 1 hour at 37 ˚C and then purified; 2-8 µg of purified product were electroporated into 100 µL of electrocompetent cells.

Cells were recovered in 4 mL SOB for 1 hour shaking at 200 rpm at 37 ˚C, then transferred to 100 mL of LB supplemented with 5 µg/mL tetracycline and 100 µg/mL ampicillin. If using linear spacers, ampicillin was omitted. Cells were incubated for another 4 hours, shaking (200 rpm) at 37 ˚C. Cells were harvested by centrifugation for 10 minutes at 4000 x g, then resuspended in 1 mL of Milli-Q water and spread in serial dilutions in LB agar plates supplemented with 5 µg/mL tetracycline plus:

- 20 µg/mL chloramphenicol and 100 µg/mL streptomycin, when the REXER BAC contains a *sacB-cat* cassette plus an *rpsL* in the backbone
- 50 µg/mL kanamycin and 7.5% sucrose, when the REXER BAC contains an *rpsL-kanR* cassette plus *sacB* in the backbone
- 50 µg/mL kanamycin and 2.5 mM 4-chloro-L-phenylalanine, when the REXER BAC contains an *rpsL-kanR* cassette plus *pheS*-hygR* in the backbone

Plates were incubated overnight at 37 °C. The resulting colonies were streaked in fresh LB agar plates supplemented with the same antibiotics as above. In order to assess the growth phenotypes, colonies were resuspended in 50 µL of Milli-Q water, and 5 µL volumes were stamped in selective agar plates as described in the text. To verify DNA exchange, locus-specific colony PCR was performed using 1µL of the same cell resuspension as template, with primers flanking locus^0 and locus^1. Colony PCRs to assess the presence or absence of deletions in Chapter 2 were performed analogously, using the primers described in the text.

5.1.6 Preparation of a non-transferrable F' plasmid

Note: This methodology corresponds to experiments performed by Julius Fredens, and is described here for completeness.

The nick site of the *oriT* region within an F' plasmid was deleted, similarly to a previous approach[279]. The tetracycline-resistant F' plasmid derivative pRK24 (Addgene #51950) was modified through several rounds of lambda-red recombination (as in 5.1.2), using a variant of the helper plasmid that contained *kanR* instead of *tetR*. First, the β-lactamase gene, conferring ampicillin resistance in pRK24, was replaced with the artificial T5-*luxABCDE* operon; correct colonies were identified via bioluminescence. Next, *tetR* was replaced with T3-*aac3*, which produces aminoglycoside 3-N-acetyltransferase IV, selecting with 50 µg/mL apramycin. Finally, a 24 bp deletion of the nick-site in *oriT* was introduced by integrating EM7-*bsd*, which expresses blasticidin-S deaminase, by lambda-red recombination, selecting with 50 µg/mL blasticidin in low-salt TYE/LB. The resulting F'-plasmid, termed pJF146 (Genbank accession MK809154.1), was purified using QIAprep Spin Miniprep Kit (QIAgen) and transformed by electroporation into relevant donor strains for subsequent conjugation.

5.1.7 Preparation of whole-genome and BAC Illumina libraries for next-generation sequencing

E. coli genomic DNA was purified from overnight cultures using the DNEasy Blood and Tissue Kit (QIAgen) as per manufacturer's instructions. BAC DNA was purified form *E. coli* overnight cultures using the QIAprep Spin Miniprep Kit (QIAgen) as per manufacturer’s instructions. We found that this kit was suitable for purification of BACs in excess of 130 kb. When handling BAC preparations, we avoided vigorous shaking of the samples throughout purification in order to reduce DNA shearing. Paired-end Illumina sequencing libraries were prepared using the Nextera XT Kit as per manufacturer’s instructions. Sequencing data was obtained in the Illumina MiSeq, running 2 x 300 or 2 x 75 cycles with MiSeq Reagent Kit v3.

5.1.8 Analysis of Illumina sequence data

The standard workflow for Illumina sequence analysis throughout this work is compiled in the iSeq package, available at https://github.com/TiongSun/iSeq. In short, custom reference sequences (e.g. recoded genome sequences) were generated with SnapGene (Insightful Science; available at www.snapgene.com) in fasta format. Sequencing reads were aligned to the reference using bowtie2 with soft clipping activated[280]. Aligned reads were sorted and indexed with samtools[281]. A customized Python script combines functionalities of samtools and igvtools to yield a variant calling summary. This script was used to assess mutations, indels and structural variations, in combination with visual analysis in the Integrative Genomics Viewer[282]. Processing of Illumina data processing for *de novo* genome assembly is described in 5.4.7.

We produced a custom Python script to generate recoding landscapes across a target genomic region. Briefly, the script takes a BAM alignment file, a reference in fasta and a Genbank annotation file as inputs. It identifies the target codons for recoding, and compiles the reads that align to these target codons in the alignment file. It then outputs the frequency of recoding at each target codon, and plots these frequencies across the length of the genomic region of interest. Script available at https://github.com/TiongSun/recoding_landscapes.

5.1.9 Growth rate measurement and analysis

The cells of interest were streaked in LB agar + 100 µg/mL streptomycin plates plus any other relevant antibiotics (e.g. selection for genomically integrated positive markers with 50 µg/mL kanamycin or 20 µg/mL chloramphenicol). 5 colonies from each were picked and grown overnight in 5 mL of LB + 100 µg/mL streptomycin. Overnight cultures were diluted 1:50 or 1:100 in LB + 100 µg/mL streptomycin in a

96 well plate. OD600 measurements were taken every 5 minutes in a Tecan Infinite Microplate reader, shaking at 37 ˚C.

To determine maximum doubling times, the growth curves were log2-transformed. At a linear phase of the curve during exponential growth, the first derivative was determined (d(log2(x))/dt) and ten consecutive time-points with the maximal log2-derivatives were used to calculate the doubling time for each replicate. A total of 5 independently grown biological replicates were measured for each strain. The mean doubling time and standard deviation from the mean were calculated for all n=5 replicates, except when indicated in the text.

5.2 – Procedures specific for Chapter 2

5.2.1 Double selection cassette integration at landing site loci

E. coli DGF-327 was kindly provided by Naotake Ogasawara's group at the Nara Institute of Science and Technology (Japan). As a result of the deletion process, the original DGF-327 strain contained a *sacB-cat* cassette in its genome. Prior to any other genetic manipulations, this cassette was replaced by a *luxABCDE* operon by lambda-red recombination, in order not to affect any future selections. Subsequently, we introduced a K43R mutation in *rpsL* by recombineering[283] with a ssDNA oligo that contained the mutation plus 50 bp of homology either side. Cells were plated in 100 μg/mL streptomycin, and the sequence of the *rpsL* locus was confirmed by Sanger sequencing. The resulting strain served as a starting point for all the manipulations described in the text. Double selection cassettes were integrated at landing site loci by lambda-red as in 5.1.2, with the exception that the commercial pRED/ET plasmid (GeneBridges) was used instead of the helper plasmid. The sequences of landing site loci are detailed in Supplementary Data 2 of ref[284].

5.2.2 Design and synthesis of sgRNAs

Cas9 cut sites were identified at the upstream and downstream homology regions of each landing site, preceding an NGG PAM sequence. In order to minimize off-targets, cut sites were chosen aided by the Cas9 Online Designer tool[285]. dsDNA templates for *in vitro* transcription were obtained by primer extension of a ssDNA scaffold

('Universal gDNA backbone') with a variable ssDNA fragment ('spacer-specific oligo') containing the target sequence, as in ref[286]. The sequence of these is detailed in **Appendix A.3**. sgRNAs for genomic excision were obtained by *in vitro* synthesis with the HiScribe™ T7 High Yield RNA Synthesis Kit (New England Biolabs, USA), as per manufacturer's instructions, using 1 µg of template DNA, and incubating for 16 hours. RNA products were treated with Roche DNAse I for 1 hour as per manufacturer's instructions, and purified by phenol-chloroform extraction followed by EtOH precipitation. RNA pellets were resuspended in RNAse-free water, quantified by NanoDrop, and stored at -20 ºC.

5.2.3 Construction of REXER BACs containing DGF-327 genomic fragments

Low-melting point agarose plugs of DGF-327 strains containing double selection markers at landing site loci were prepared using the CHEF Genomic DNA Plug Kit (BioRad, USA), as per manufacturer's instructions. Genomic DNA plugs were equilibrated in 1 mL 1X Cas9 Reaction Buffer (New England Biolabs) for 30 mins, providing fresh buffer half-way through. For each genomic segment, 4 plugs (400 µL total volume) were digested in a total volume of 1200 µL in 1X Cas9 Reaction Buffer, in the presence of 20 nM purified *S. pyogenes* Cas9 (New England Biolabs) and ~300 nM of each sgRNA (sequences detailed in **Appendix A.3**). Before digestion, Cas9 was pre-mixed with the two sgRNAs in 1X Cas9 Reaction Buffer in a total volume of 50 µL, and incubated for 10 mins at room-temperature. The Cas9-sgRNA mix was then added to the digestion reaction for the total 1200 µL. Reactions were incubated for 2 hours at 37 °C, and the buffer was then aspirated. Approximately half a plug (~50 µL) was cut with a clean scalpel and melted at 65 °C for 10 mins together with 10 µL of 6X Gel Loading Dye (New England Biolabs). The melted mixture was carefully loaded into a 1% agarose gel (1X TBE) with wide-bore tips. Electrophoresis was conducted in a BioRad CHEF-DR-III pulse-field electrophoresis system, in 0.5x TBE, for 14-16 hours at 14 °C, with 1-80s switch times, 120° angle and 6 V/cm. Gels were stained with 1X SybrSafe and visualized in a suitable imager.

The remaining ~350 µL of plugs were equilibrated for 30 mins in 500 µL of 1X β-agarase buffer (New England Biolabs) at 4 ˚C, providing fresh buffer half-way through. The buffer was then aspirated and the plugs were melted at 65 °C for 10 minutes, followed by equilibration to 42 °C. 2 units of β-agarase were added to the molten plugs, followed by incubation for 1 hour at 42 °C. The resulting solution was carefully transferred to a drop dialysis membrane (0.025 µm MF-Millipore Membrane Filter, Merck Millipore, USA) with wide-bore tips; membranes floated on a petri dish filled with ~25 mL of Milli-Q water. Dialysis was carried out overnight at room temperature. Samples were carefully recovered from the membranes using wide-bore tips, and concentrated using a Savant SpeeVac Plus concentrator (Thermo Fischer Scientific, USA) to a final volume of ~250 µL.

5.2.4 Pulse-field gel electrophoresis verification of DGF-327 genomic library

Low-melting point agarose plugs of MDS42 cells containing BACs with DGF-327 genomic inserts were prepared as described in 5.2.3. sgRNAs targeting the BAC-insert junctions were prepared as in 5.2.2; the specific spacer sequences are detailed in **Appendix A.4**. Cas9 cleavage was performed as in 5.2.3, scaling down reaction volumes from 1200 µL to 300 µL (digestion of one plug per fragment). Each plug was cut in 2, and half a plug was loaded onto the gel for analysis by pulse-field gel electrophoresis, as in 5.2.3. The plugs in this section were prepared exclusively for quality control purposes and were not subject to any further treatment or purification.

5.3 – Procedures specific for Chapter 3

5.3.1 Recoded genome design

We based our synthetic genome design on the sequence of the *E. coli* MDS42 genome (accession number AP012306.1, released 07-Oct-2016), which has 3547 annotated CDS. We manually curated the starting genome annotation to remove *htgA*, *ybbV*, and *yzfA* as described in the text. Conversely, the pseudogenes *ydeU*, *ygaY*, *pbl*, *yghX*, *yghY*, *agaW*, *yhiK*, *yhjQ*, *rph*, *ysdC*, *glvG*, and *cybC* were promoted to CDS. To enable negative selection with *rpsL*, we mutated the genomic copy of *rpsL* to *rpsL*K43R. Deep sequencing of our in-house MDS42 strain revealed a 51 bp insertion between *mrcB* and *hemL* which had not been reported in AP012306.1. We manually introduced and annotated this insertion in our starting genome sequence.

We produced a custom Python script that i) identifies and recodes all target codons, and ii) identifies and resolves overlapping gene sequences that contain target codons (available at https://github.com/TiongSun/genome_recoding). From our curated MDS42 starting sequence, we used the script to generate a new synthetic genome sequence in which all TCG, TCA and TAG codons were replaced with AGC, AGT and TAA respectively. The script reported 91 CDS with overlaps containing target codons. The specific overlap instances are detailed in Supplementary Data 4 of ref[284].

prfB (release factor RF-2) was not annotated as a CDS in our starting MDS42 genome due to its regulatory internal stop codon, and we therefore recoded all the target codons in the gene manually, thereby maintaining the internal stop codon. The resulting genome design contained 3556 CDS with 1,156,625 codons of which 18,218 were recoded. A GenBank file of the annotated genome design is available in Supplementary Data 2 of ref[284]. The sequence is also available at GenBank accession CP040347.1. A list of all codon substitutions is provided in Supplementary Data 3 of ref[284].

5.3.2 BAC assembly and delivery by electroporation

REXER selection constructs for BAC assembly (3.2.2) were cloned into pSC101-based plasmids by Gibson assembly of two overlapping fragments. **Figure 5.2** shows an example of a selection construct for a BAC containing *sacB-cat* for flanking synthetic DNA, plus an *rpsL* marker in the backbone. Sequences of selection constructs are available at Genbank accession numbers MN927220 and MN927221. The selection constructs were released by BsaI digestion followed by purification with a Zymogen (USA) DNA Clean and Concentrator Kit.

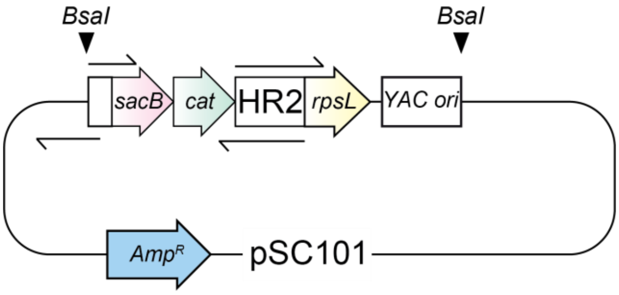


Figure 5.2 - Construction REXER selection constructs. Primers are represented by black arrows. BsaI restriction sites are shown. The white block upstream *sacB* corresponds to a homology region to the last synthetic DNA stretch in BAC assembly.

Synthetic DNA stretches were designed to contain 50-80 bp of homology to their neighbouring fragments. They were obtained from GENEWIZ (USA) in pSC101 or pST vectors flanked BsaI, AvrII, SpeI or XbaI restriction sites. The synthetic stretches did not contain recognition sequences for the restriction enzymes that flanked them. Stretches were released from their harbouring vectors by restriction digestion and purified as above. The sequences of each stretch are detailed in Supplementary Data 2 and 5 of ref[284].

BAC vectors were amplified with suitable homologies by PCR, using a BAC/YAC plasmid generated previously as a template[185], and purified as above. Primers used for amplifying BAC vectors are detailed Supplementary Data 9 in ref[284].

All DNA pieces in the reaction were mixed in equimolar amounts (30-50 fmol), and co-transformed into *S. cerevisiae* spheropasts prepared as in 5.1.3. Assembly proceeded as in Kouprina et al. [ref[277]]. The resulting yeast colonies were picked and resuspended in 30 µL of Milli-Q H_2O. 6 µL of colony resuspension were mixed with 14 µL of 0.03 M NaOH, and incubated at 95 ºC for 5 minutes. 1 µL of the resulting lysates was used as a template for colony PCR at the relevant junctions.

From yeast clones that genotyped positively at all tested junctions, we extracted total DNA with the Gentra Puregene Yeast/Bact. Kit (QIAgen), as per manufacturer's instructions. 10-20 µL of purified DNA were electroporated into 100 µL of competent cells of the relevant strain for REXER, normally containing a helper plasmid plus a genomically integrated double selection cassette (*rpsL-kanR* or *sacB-cat*; see 5.1.5)

5.3.3 BAC delivery by conjugation

An *oriT-apmR* cassette was integrated into BACs by lambda-red recombination (5.1.2) in the donor strains. The position and sequence of the cassette is annotated in accession numbers MK809149.1 and MK809150.1. Colonies were selected in LB agar supplemented with 50 µg/mL apramycin. Donor strains were subsequently transformed with pJF146 (MK809154.1) by electroporation. Recipient cells were transformed with a helper plasmid by electroporation, if they did not contain one already.

5 ml of donor and recipient cultures were grown to saturation overnight in selective LB medium supplemented with appropriate antibiotics to select for:

- **Donor:** pJF146 (50 µg/mL blasticidin in low-salt LB) and the positive marker in the BAC (50 µg/mL kanamycin or 20 µg/mL chloramphenicol)
- **Recipient:** Helper plasmid (5-10 µg/mL tetracycline)

Cultures were pelleted by centrifugation at 5000 x g for 5 minutes, and subsequently washed 3 times with 2 mL of LB medium without antibiotics, and resuspended in 200 µL of LB. The donor and recipient cell suspensions were combined in a 4:1 ratio, spotted on TYE agar plates in 10-20 µL droplets and incubated for 1h at 37°C. Cells were washed off the plate with 2 mL of LB and spread in serial dilutions on LB agar plates supplemented with 2% glucose, 5 µg/ml tetracycline, and 50 µg/mL kanamycin or 20 µg/mL chloramphenicol depending on the cassette present in the BAC. From the resulting colonies, successful transfer of the BAC was confirmed by colony PCR of the BAC-vector insert junctions.

5.3.4 BAC editing

sacB loss-of-function mutations in REXER BACs for fragments 2, 6, 8, 10, 12 and 16 were repaired by replacing *sacB* with *pheS*-hygR*, by lambda-red recombination. *pheS*-hygR* cassettes were amplified with 50 bp of homology to either side of the *sacB* locus in the REXER BAC, and lambda-red recombination proceeded as in 5.1.2, selecting in LB agar supplemented with 5 µg/mL tetracycline and 200 µg/mL hygromycin B. Replacement was confirmed by selective growth assay and colony PCR of the new *pheS*-hygR* locus in the BAC. Primers are detailed in Supplementary Data 9 of ref[284].

Note: BAC editing below was performed by Julius Fredens and Louise Funke, and is described here for completeness.

BAC editing to change recoded codons (e.g. in fragments 37a, 1, and 9, described in Chapter 3) was performed by two successive rounds of lambda-red recombination, as in 5.1.2. *pheS*-hygR* cassettes amplified by PCR to contain 50 bp of homology to either side of the loci of interest were integrated into the BAC. Because BACs shared high levels of homology with the host genomes, a fraction of the resulting clones had integrated *pheS*-hygR* in the genome. Genomic and BAC integration events were distinguished by stamping the cells in selective LB agar medium containing i) 200 µg/mL hygromycin B + 2.5 mM 4-chloro-phenylalanine and ii) 200 µg/mL hygromycin + 7.5% sucrose. Clones that survived condition i) were discarded. Clones that survived condition ii) were further characterized by colony PCR at the relevant locus.

The correct clones were subject to a second round of lambda-red recombination by electroporation of a ssDNA or dsDNA that contained the fix of interest plus 50 bp of homology to either side of *pheS*-hygR*, and spread on 5 µg/mL tetracycline + 7.5% sucrose + 2.5 mM 4-chloro-phenylalanine. The target locus was again characterised by colony PCR, and the sequence of the entire BAC was verified by next-generation sequencing. Primers for mutating BACs are detailed in Supplementary Data 14 of ref[284].

5.3.5 Assembling a synthetic genome from recoded sections

Sections 3.2.3 and 3.2.4 describe the general strategy for preparing donor and recipient strains for conjugation. *oriT* sequences for genomic integration were most often provided in a construct linked to gentamycin resistance gene *gmR* (accession MK809155.1), although one experiment was performed with an *rpsL-hygR-oriT* cassette. These cassettes were amplified by PCR and integrated in the relevant genomes by lambda-red recombination as in 5.1.2; the primers and location of integration for each are described in Supplementary Data 16 of ref[284]. Donor cells were further prepared by curing of the helper plasmid by passaging (as confirmed by selective growth assay in tetracycline) and electroporation of F' plasmid pJF146.

Separately, prior to each conjugation a double selection cassette (often *pheS*-hygR*) was integrated 3 kbp downstream of locus^0 in the donor. This provided a genomic DNA template for amplification of a recoded homology overlap coupled to a double selection cassette. The resulting PCRs were then integrated in the recipient genomes by lambda-red recombination as in section 5.1.2, replacing the cassette that was originally present at locus^0 (**Figure 3.12**, Approach 1). When following the converse Approach 2, the *oriT* cassette was first integrated 3 kbp upstream of locus^0 in the recipient, and this served as a template to PCR-amplify *oriT + 3* kbp of recoded homology in the donor.

For conjugation, donor and recipient strain were grown to saturation overnight in LB medium supplemented with:

- **Donor:** 2% glucose, 50 µg/ml apramycin and 50 µg/ml kanamycin or 20 µg/mL chloramphenicol.
- **Recipient:** 2% glucose, 5 µg/ml tetracycline, and either 50 µg/ml kanamycin, 20 µg/ml chloramphenicol or 200 µg/mL hygromycin B.

The overnight cultures were diluted 1:10 in the same selective LB medium and grown to OD600 = 0.5. 50 ml of both donor and recipient culture were washed 3 times with 5 mL LB medium with 2% glucose and then each resuspended in 400 µL LB medium with 2% glucose. 320 µL of donor was mixed with 80 µL of recipient, spotted on TYE agar plates and incubated at 37°C. The incubation time depended on the length of transferred synthetic DNA and the apparent fitness of the recipient strain, and varied from 1h to 3h. Cells were washed off the plate and transferred into 100 mL LB medium with 2% glucose and 5 µg/ml tetracycline and incubated at 37 °C for 2h with shaking at 200 rpm. Subsequently 50 µg/ml kanamycin or 20 µg/ml chloramphenicol (selecting for the positive selection marker of the donor) was added, followed by another 2 h incubation at 37 °C. The culture was spun down for 10 mins at 4000 x g, resuspended in 4 mL Milli-Q filtered water and spread in serial dilutions on LB agar supplemented with 2% glucose, 5 µg/mL tetracycline, 2.5 mM 4-chloro-phenylalanine and 50 µg/mL kanamycin or 20 µg/ml chloramphenicol. For each experiment, successful DNA transfer and recombination were determined by colony PCR at locus^0 and locus^1 as well as selective growth assay in the relevant selective agents.

5.3.6 Microscopy and cell size measurement

Note: These methods correspond to experiments performed by Yonka Christova and Nick Barry, and are described here for completeness.

Cells were grown with shaking in LB supplemented with 100 µg/mL streptomycin to approximately OD600=0.2. A thin layer of bacteria was sandwiched between an agarose pad and a coverslip. A standard microscope slide was prepared with a 1% agarose pad (Sigma-Aldrich A4018-5G). A sample of 2 µl to 4 µl of bacterial culture was dropped onto the top of the pad. This was covered by a #1 coverslip supported on either side by a glass spacer matched to the ~1 mm height of the pad. Samples were imaged on an upright Zeiss Axiophot phase contrast microscope using a 63X 1.25NA Plan Neofluar phase objective (Zeiss UK, Cambridge, UK). Images were taken using an IDS ueye monochrome camera under control of ueye cockpit software (IDS Imaging Development Systems GmbH, Obersulm, Germany). 10 fields were taken of each sample. Images were loaded in to Nikon NIS Elements software for further quantitation (Nikon Instruments Surrey UK). The General analysis tool was used to apply an intensity threshold to segment the bacteria. A one micron lower size limit was imposed to remove background particulates and dust. Length measurements were subsequently made on the segmented bacteria using the General Analysis quantification tools.

5.3.7 Proteomic analysis

Note: These methods correspond to experiments performed by Julius Fredens with support from the MRC LMB mass spectrometry facility, and are described here for completeness.

Three biological replicates were performed for each strain. Proteins from each *Escherichia coli* lysates were solubilized in a buffer containing 6 M urea in 50 mM ammonium bicarbonate, reduced with 10 mM DTT, and alkylated with 55 mM iodoacetamide. After alkylation, proteins were diluted to 1 M urea with 50 mM ammonium bicarbonate, digested with Lys-C (Promega, UK) at a protein to enzyme ratio of 1:50 for 2 hours at 37 °C, followed by digestion with Trypsin (Promega, UK) at a protein to enzyme ratio of 1:100 for 12 hours 37 °C. The resulting peptide mixtures were acidified by the addition formic acid to a final concentration of 2% v/v. The digests were analysed in duplicate (1 µg initial protein/injection) by nano-scale capillary LC-MS/MS using a Ultimate U3000 HPLC (ThermoScientific Dionex, San Jose, USA) to deliver a flow of approximately 300 nL/min. A C18 Acclaim PepMap100 5 µm, 100 µm x 20 mm nanoViper (ThermoScientific Dionex, San Jose, USA), trapped the peptides prior to separation on a C18 Acclaim PepMap100 3 µm, 75 µm x 250 mm nanoViper (ThermoScientific Dionex, San Jose, USA). Peptides were eluted with a 100 minute gradient of acetonitrile (2% to 60%). The analytical column outlet was directly interfaced via a nano-flow electrospray ionisation source, with a hybrid dual pressure linear ion trap mass spectrometer (Orbitrap Velos, ThermoScientific, San Jose, USA). Data dependent analysis was carried out, using a resolution of 30,000 for the full MS spectrum, followed by ten MS/MS spectra in the linear ion trap. MS spectra were collected over a m/z range of 300–2000. MS/MS scans were collected using a threshold energy of 35 for collision induced dissociation. All raw files were processed with MaxQuant 1.5.5.111 using standard settings and searched against an *Escherichia coli* strain K-12 with the Andromeda search engine12 integrated into the MaxQuant software suite. Enzyme search specificity was Trypsin/P for both endoproteinases. Up to two missed cleavages for each peptide were allowed. Carbamidomethylation of cysteines was set as fixed modification with oxidized methionine and protein N-acetylation considered as variable modifications. The search was performed with an initial mass tolerance of 6 ppm for the precursor ion and 0.5 Da for CID MS/MS spectra. The false discovery rate was fixed at 1% at the peptide and protein level. Statistical analysis was carried out using the Perseus (version 1.5.5.3) module of MaxQuant. Prior to statistical analysis, peptides mapped to known contaminants, reverse hits and protein groups only identified by site were removed. Only protein groups identified with at least two peptides, one of which was unique and two quantitation events were considered for data analysis. For proteins quantified at least once in each strain, the average abundance of each protein across replicates of Syn61 was divided by the abundance in MDS42 replicates, and then log2-transformed. A P-value for the difference in abundance between strains was calculated by two-sample T-test (Perseus).

5.3.8 Toxicity of CYPK incorporation using orthogonal aminoacyl-tRNA synthetases

Electrocompetent MDS42 and Syn61 cells were transformed with plasmid pKW1_MmPylRS_PylTXXX (as in ref[78]), for expression of *M. mazei* PylRS and $\text{tRNA}^{\text{Pyl}}_{\text{XXX}}$, where XXX is the indicated anticodon. Three variants of this plasmid were used, with the anticodon of tRNA^{Pyl} mutated to CGA (pKW1_MmPylRS_PylTCGA), UGA (pKW1_MmPylRS_PylTUGA) or GCU (pKW1_MmPylS_PylTGCU). Cells were grown over night in LB medium with 75 μg/ml spectinomycin. Overnight cultures were diluted 1:100 into LB supplemented with Nε-(((2-methylcycloprop-2-en-1-yl) methoxy) carbonyl)-L-lysine (CYPK) at 0 mM, 0.5 mM, 1 mM, 2.5 mM and 5 mM, in a 96 well plate. Growth of the cultures was measured in a Tecan plate reader. "% Max Growth" was determined as the final OD_{600} in the presence of the indicated concentration of CYPK divided by the final OD_{600} in the absence of CYPK. Final OD_{600}'s were determined after 600 min.

5.3.9 Deletion of *prfA*, *serU* and *serT*

A derivative of the helper plasmid was generated by removal of Cas9 and tracrRNA, replacement of lambda-red components with recoded versions, and replacement of the tetracycline resistance gene with a recoded apramycin resistance gene, by Gibson assembly. This plasmid, pKW20_Ara, was used in lambda-red recombinations for deletions in this section.

tRNA and RF-1 deletions were performed with recoded double selection cassettes, where annotated instances of TCG, TCA and TAG are replaced by their synonyms as described in the text. These cassettes were synthesised as gBlocks from IDT (Integrative DNA Technologies). Cassettes were PCR-amplified with primers that contained 50 bp overhangs to either side of the *serU* or *serT* gene, or the *prfA* CDS. The resulting products were used for lambda-red recombination as in 5.1.2. All recombinations were performed in 2xTY instead of LB.

In order to remove the double selection cassettes at the tRNA loci without replacing a tRNA (clean deletions), we PCR-amplified ~250 bp either side of each tRNA gene in such a way that the two products contained 20-40 bp of mutual homology. These two pieces were stitched together by overlap-extension PCR. The resulting product was used for lambda-red recombination as in 5.1.2, counter-selecting with 100 µg/mL streptomycin.

5.3.10 Construction of plasmids for sense codon reassignment

Recoded, His_6-tagged wild-type sperm-whale myoglobin[51] and T4 Lysozyme[287] were obtained from IDT as gBlocks, and cloned by Gibson assembly into plasmids downstream of a pBAD promoter, together with a p15A origin of replication, a recoded apramycin resistance gene, and a recoded *araC* gene. The resulting plasmids are named pMyo and pT4L. These were derivatized to contain TCG codons at multiple positions by site-directed mutagenesis.

pJF170 is a pCDF-based plasmid containing *alaT*$_{\text{CGA}}$ embedded within a native *serT* regulatory context. It was constructed by Gibson assembly with i) a PCR of *alaT*, ii) a PCR of recoded apramycin (template obtained as a gBlock), and iii) two ~250 bp PCRs of the genomic region immediately upstream and downstream of *serT*. The resulting plasmid was modified by site-directed mutagenesis to introduce a CGA anticodon in *alaT*, yielding pJF170. pJF171 is analogous to pJF170 but lacks the *alaT*$_{\text{CGA}}$ gene, and was constructed similarly. Both plasmids were modified by Gibson assembly to replace the recoded apramycin resistance gene with a recoded spectinomycin resistance gene, yielding pJF170_Spec and pJF171_Spec.

A pMB1-based plasmid containing a *M. barkeri* PylRS under control of the GlnRS promoter and a tRNA$^{\text{Pyl}}_{\text{CGA}}$ under control of an *lpp* promoter[78], termed pKW_PylRS-PylTCGA (as in section 5.3.8), was used for BocK incorporation experiments. This plasmid confers spectinomycin resistance.

5.3.11 Expression of myoglobin and T4Lysozyme

Prior to expression experiments, the deletion strains of interest were passaged to cure the pKW20_Ara plasmid. Absence of the plasmid was verified by parallel stamping of colonies in LB agar supplemented with 50 µg/mL apramycin, and plain LB. Resulting plasmid-free colonies were made electrocompetent as in 5.1.1, and co-transformed with i) a pMyo or pT4L derivative, and ii) pJF170, pJF171 or pKW_PylRS-PylTCGA. Cells were recovered for 1:30 hours in 1 mL SOC, then transferred to 20 mL of 2xTY supplemented with 50 µg/mL apramycin and 75 µg/mL spectinomycin. Cultures were incubated until dense at 37 °C, shaking at 220 rpm. Cultures were diluted 1/50 in 5 mL to 1 L of 2xTY supplemented with 50 µg/mL apramycin, 75 µg/mL spectinomycin, 0.2% L-arabinose. For experiments carried out with pKW_PylRS-PylTCGA, 2-5 mM BocK was added to the medium, as indicated in the text. Expression cultures were incubated for 18-24 hours a 37 °C with shaking at 220 rpm. Expression in cultures harbouring pJF171 was induced 6 hours before cultures harbouring pJF170. Both cultures were grown to saturation and harvested at the same time. Depending on the type of analysis, cells were treated as described in 5.3.12.

5.3.12 Western of myoglobin and T4Lysozyme

In order to perform OD_{600}-normalized western analysis, an equivalent of 75 µL of culture at OD_{600} = 1 (in a 0.2 cm pathlength) was harvested for each sample by centrifugation at 5,000 x g for 5 minutes. Samples were resuspended in 100 µL of 1X NuPage LDS buffer with 5% β-mercaptoethanol, and boiled for 5 minutes at 95 °C. The samples were then vortexed for ~1 minute to shear DNA and reduce viscosity. Equal amounts were loaded into a 4-12% Bis-Tris NuPage gel, transferred to a PVDF membrane and analysed by western blot with an anti-His_6 antibody (Cell Signalling Technologies 27E8, or Abcam ab18184).

In order to perform western blot of crude lysates, entire cultures were harvested by centrifugation at 5,000 x g for 5 minutes. Pellets were resuspended in 1/10th the volume of lysis buffer (1X BugBuster Protein Extraction Reagent, 1X PBS, 100 µg/mL lysozyme, 100 µg/mL DNAse I, 1 mM PMSF and 20 mM imidazole), and incubated at room temperature for 1 hour while rotating in a Hula Shaker. Samples were centrifuged for 20 minutes at 30,000 x g, and the supernatant transferred to a fresh tube. The protein concentration in the supernatant was estimated by absorbance at 280nm or BCA assay. Equal protein masses were loaded into a 4-12% Bis-Tris NuPage gel and analysed by western blot as above. The same sample amounts were loaded into a separate gel, for analysis by SDS-PAGE and Coommassie staining as a loading comparison.

5.3.13 Protein purification and ESI-MS spectrometry

The clarified lysates from 5.3.12 were mixed with 100 µL of Ni^{2+}NTA beads (QIAgen), and incubated at room temperature for 1 hour while rotating in a hula shaker. The lysate-bead mixtures were transferred to columns and washed 3 times with 1mL of 1X PBS + 25 mM imidazole. Proteins were eluted in 250 µL of 1X PBS + 300 mM imidazole.

Protein samples were diluted 1/4 in water and subject to LC-MS analysis. Briefly, proteins were separated on a C4 BEH 1.7µm, 1.0 x 100mm UPLC column (Waters, UK) using a modified nanoAcquity (Waters, UK) to deliver a flow of approximately 50 µl/min. The column was developed over 20 minutes with a gradient of acetonitrile (2% v/v to 80% v/v) in 0.1% v/v formic acid. The analytical column outlet was directly interfaced via an electrospray ionisation source, with a hybrid quadrupole time-of-flight mass spectrometer (Xevo G2, Waters, UK). Data was acquired over a m/z range of 300–2000, in positive ion mode with a cone voltage of 30V. Scans were summed together manually and deconvoluted using MaxEnt1 (Masslynx, Waters, UK). The theoretical average molecular weights of the proteins were calculated using the ExPASY online tool (https://web.expasy.org/compute_pi). The calculated mass of proteins containing BocK was manually adjusted to account for this.

Note: ESI-MS experiments were performed by the MRC Laboratory of Molecular Biology Mass Spectrometry facility; they provided experimental details.

5.4 – Procedures specific for Chapter 4

5.4.1 Strains and plasmids used in this Chapter

All cloning procedures described in this section were performed in *E. coli* DH10b, which carries an *rpsL*K43R mutation that confers resistance to streptomycin.

Construction of fission BACs

Fission BACs were constructed by 4-piece Gibson assembly of:

- *luxABCDE* PCR product.
- BAC vector PCR product, which contains *sacB-cat* and BAC replication/partitioning elements (*repE*, *sopA/B/C*), derived from accession MN226640.1.
- An *rpsL* construct flanked by HR1 and HR4 plus 20 bp of homology to the 5’ of *luxABCDE* and the 3’ of the BAC vector, synthesized as a gBlock (Integrated DNA Technologies).
- An *rpsL* construct flanked by HR2 and HR3 plus 20 bp of homology to the 3’ of luxABCDE and the 5’ of the BAC vector, synthesized as a gBlock (Integrated DNA Technologies).

See GenBank accession MN226640.1 for a template BAC sequence. Homology regions HR1-HR4 in the different fission BACs are defined in Table S2 of ref[278].

pSC101_*oriT-pheS*-kanR*

Fusion and conjugation experiments were performed with a BAC backbone in which the *sacB-cat* cassette is replaced by an *oriT-pheS*-kanR* cassette (see below). To generate this variant, plasmid pSC101_*oriT-pheS*-kanR* was constructed by Gibson assembly of 4 sequences amplified with homologous ends:

- pSC101 backbone PCR product
- *oriT* sequence PCR product
- EM7-*pheS** PCR product
- *kanR* PCR product.

This plasmid served as a template for amplification of the *oriT-pheS*-kanR* cassette in downstream recombinations (see below). The sequence of the *oriT-pheS*-kanR* cassette is detailed in GenBank accession VMQZ00000000.

Fission experiments were performed in the *E. coli* strain MDS42 (AP012306.1), where the *rpsL*K43R mutation had been introduced by recombineering, as above, to confer resistance to streptomycin, but sensitivity in the presence of an additional wild-type *rpsL* copy.

Fission of recoded section C and subsequent fusion experiments were performed in a partially recoded MDS42$^{\textit{rpsL}\text{K43R}}$ derivative, MDS42_SynC, generated by GENESIS in Chapter 3. The sequence is detailed in GenBank accession CP042184; the recoded region spans nucleotides 1,454,024 and 1,979,777. A strain with recoded section A (MDS42_SynA) was generated in a similar fashion (CP042183), where recoding spans 436,924-939,332. This strain was used for fission of wild-type section C in chromosomal transplant and chimeric genome generation (see below). As a result of GENESIS, MDS42_SynC originally contained an *rpsL-kanR* cassette at position 1,979,778- 1,979,783, and MDS42_SynA contained *sacB-cat* at 1,050,809-1,050,814. Both of these markers were removed by lambda-red recombination as in 5.1.2, providing wild-type dsDNA as a recombination template and selecting with 100 μg/mL streptomycin or 7.5% sucrose, respectively. The marker-free strains were then used in fission and fusion experiments.

5.4.2 Genome fission

To perform fission, the strains of interest were electroporated sequentially, first with the helper plasmid, and then with the corresponding fission BAC. Cells harbouring both of these plasmids were made competent with 0.5% L-arabinose induction for 1 hour, as described in 5.1.2, to express Cas9 and the lambda-red recombination machinery. 100 μL of induced electrocompetent cells were electroporated with ~8 μg of the corresponding spacer plasmid in pKW3 (MN226641.1) or pKW5 (MN226642.1) format, encoding spacer RNAs for the 6 necessary Cas9 cleavages. Cells were recovered in 4 mL of SOB shaking at 37 °C for 1 hour. 80 μL of 25% L-arabinose were then added for a final concentration of 0.5%, and incubated for another hour. Cells were subsequently transferred to 100 mL of LB + 5 μg/mL tetracycline + 20 μg/mL chloramphenicol + 100 μg/mL ampicillin (for fissions performed with pKW3) or 50 ng/μL kanamycin (for fissions performed with pKW5), and incubated shaking at 37 °C for 4 hours. Cells were then pelleted by centrifugation for 10 min at 4000 x g, resuspended in Milli-Q filtered water and spread in serial dilutions in LB agar plates supplemented with 5 μg/mL tetracycline + 100 μg/mL streptomycin + 20 μg/mL chloramphenicol and either 100 μg/mL ampicillin (for fissions performed with pKW3) or 50 ng/μL kanamycin (for fissions performed with pKW5). From the resulting colonies, clones were identified that had some luminescent signal, but where this signal was lower than that of a pre-fission control. Candidate clones were assessed phenotypically by resuspending colonies in 50 μL of Milli-Q water and stamping them on the selection plates as indicated throughout the text. From overnight cultures of candidate clones, genomic DNA extractions were performed with the QIAgen Blood and Tissue kit, as per manufacturer's instructions. From the genomic DNA template, PCR reactions were performed with primers targeting the hypothetical new junctions, flanking i) either side of linker 1 (*luxABCDE* operon) in Chr. 1, and ii) either side of linker 2 (BAC backbone) in Chr. 2.

5.4.3 Construction of watermarked Chr. 2 bearing *oriT-pheS*-kanR*

The partially recoded MDS42_SynC strain (CP042184) contains a recoded section C (1,454,024 to 1,979,777). Fission was used to partition this recoded section into Chr. 2. The *sacB-cat* in the linker 2 of Chr. 2 was subsequently replaced by an *oriT-pheS*-kanR* using DOSER[185]. Briefly, post-fission cells harbouring the helper plasmid were made electrocompetent with 0.5% L-arabinose induction for 1 hour, as in 5.1.2. 100 µL of electrocompetent cells were electroporated with ~8 µg of *oriT-pheS*-kanR* PCR product, amplified from pSC101_*oriT-pheS*-kanR* generated in 5.4.1, with primers containing homology to the flanking regions of *sacB-cat*. The sequences are detailed in Table S6 of ref[278]. Cells were recovered for 4 hours in SOB at 37 ºC, while shaking at 200 rpm. Cells were then harvested by centrifugation for 10 minutes at 4000 x g, resuspended in 1 mL of Milli-Q water and spread on LB agar plates containing 7.5% sucrose and 50 µg/mL kanamycin. Clones that had undergone the replacement were identified by colony PCR flanking the *oriT-pheS*-kanR* integration site in Chr. 2, followed by Sanger sequencing. The phenotype of the correct clones was validated by stamping cell suspensions on 20 ng/µL chloramphenicol, 7.5% sucrose, 2.5 mM 4-chloro-phenylalanine, 50 ng/µL kanamycin or a combination of these. The sequence of the resulting strain is detailed in Genbank accession VMQZ00000000.

5.4.4 Genome fusion

Prior to genome fusion, a fusion sequence (*pheS*-hygR*) was integrated in Chr. 1 at several positions depending on the experiment. Integrations were performed by lambda-red recombination as in 5.1.2, providing *pheS*-hygR* PCR products amplified with the primers detailed in Table S6 of ref[278]. Cells were spread in LB agar supplemented with 5 µg/mL tetracycline and 200 µg/mL hygromycin B. Correct integrations were verified by colony PCR followed by Sanger sequencing.

After introduction of a fusion sequence at the desired locus, cells were made electrocompetent with L-arabinose induction, as described in 5.1.2. 100 µL of electrocompetent cells were electroporated with ~8 µg of the corresponding pKW3 fusion plasmid, encoding spacer RNAs for 4 Cas9 cleavages that initiate fusion. Cells were recovered for 2 hours in 4 mL of SOB at 37 °C with shaking at 220 rpm, with addition of 0.5% L-arabinose after 1 hour (as in 5.4.4). Cells were then transferred to

100 mL LB supplemented with 5 µg/mL tetracycline + 100 µg/mL ampicillin and recovered for 4 hours at 37 °C with shaking at 220 rpm. Cells were then pelleted by centrifugation for 10 minutes at 4000 x g, resuspended in 1 mL Milli-Q water and spread on LB agar plates containing 5 µg/mL tetracycline + 100 ng/µL ampicillin + 2.5 mM 4-chloro-phenylalanine. Plates were incubated overnight at 37 °C. The phenotype of the resulting colonies was verified by resuspending them in 50 µL of Milli-Q water and stamping on selective LB agar plates as described in the text. Genomic DNA was prepared with the QIAgen Blood and Tissue kit as above, and PCR reactions were performed with primers flanking either side of the two new junctions generated by fusion as described in the relevant text.

5.4.5 Chromosomal transplant

In this experiment, the donor cell is MDS42_SynC post-fission of section C after replacement of *sacB-cat* in Chr. 2 by *oriT-pheS*-kanR*, as described in 5.4.3. The recipient is MDS42_SynA after fission of wild-type section C.

First, the linker 1 in Chr. 1 of the recipient cell was replaced with a *pheS*-hygR* cassette (fusion sequence), as in 5.4.4. Then, the donor and recipient strains were respectively passaged for 1-3 days in LB supplemented with 50 µg/mL kanamycin or 18 µg/mL chloramphenicol, in the absence of tetracycline or ampicillin to isolate clones cured of both the pKW3 fission spacer plasmid and the helper plasmid. After streaking, plasmid-free clones were identified by stamping in LB agar with and without tetracycline or ampicillin. Post-fission donor cells were subsequently electroporated with F' pJF146, selecting in 50 µg/mL apramycin. A clone resulting from this transformation was grown overnight in 25 µg/mL apramycin as the donor. The recipient was grown overnight in 200 µg/mL hygromycin B.

5 mL of cell suspension of OD_{600} = 2.4 were spun down for 10 mins at 4000 x g and resuspended in 2 mL LB + 2% glucose, and this wash was repeated twice. Cells were then resuspended in 200 µL LB + 2% glucose. 100 µL of donor cell suspension and 100 µL of recipient suspension were gently mixed by pipetting, and the mixture spotted in 10-20 µL droplets on 37 °C pre-warmed TYE agar plates. The spots were left to dry and then incubated for 2 hours at 37 °C. Cells were washed off the plate with LB, diluted in 200 mL LB + 50 µg/mL kanamycin and 200 µg/mL hygromycin B,

and incubated while shaking at 37 °C for 3 hours. The entire culture was pelleted by centrifugation for 10 minutes at 4000 x g, resuspended in 1 mL Milli-Q water and spread in serial dilutions in LB agar supplemented with 50 ng/µL kanamycin + 200 ng/µL hygromycin B + 7.5% sucrose. The resulting colonies were resuspended in 50 µL of Milli-Q and stamped on relevant selection plates, as described in the text.

5.4.6 Nanopore sequencing

E. coli overnight cultures were purified with the PureGene Yeast/Bacteria Kit (QIAgen) as per manufacturer's instructions, taking care to avoid vigorous shaking so as to minimise DNA shearing. Sequencing libraries were prepared with the Rapid Barcoding Kit (SQK-RBK004) with the following modifications. Purified DNA samples were diluted 1:10 in water and quantified using a Qubit Fluorometer with a dsDNA High Sensitivity assay. For a given sequencing run, equal masses of different undiluted DNA samples (at least 600 ng) were subject to fragmentation with 3.5 µL of barcoded fragmentation mix, adding water to a total volume of 20 µL, and incubated for 1min 30s at 30 °C followed by 1 minute at 80 °C. AmpureXP bead clean-up was omitted and the barcoded libraries were pooled directly in equimass ratios, mixed with an equal volume of AMPure XP beads, washed twice with 500 µL of 70% EtOH and eluted in 11 µL of 10 mM Tris-HCl (pH 8.0) + 50 mM NaCl. 1 µL of pooled library was quantified in a Qubit Fluorometer with a dsDNA High Sensitivity assay, and 1 µL of RAP was added to the remaining 10 µL. The library loading mix was prepared with 11 µL of library, 34 µL of SQB, 25.5 µL of loading beads and 4.5 µL of nuclease-free water. The typical total input into the flowcell was ~1 µg of DNA. Sequencing was performed on a MinION Mk1B, using the MinKNOW software with standard 48 hour protocols . Raw data has been deposited in SRA, accession numbers are detailed in **Appendix C.2**.

5.4.7 *De novo* genome assembly

Illumina reads were trimmed with Trimmomatic to remove adapter sequences[288]. Oxford Nanopore reads were basecalled using Guppy basecaller v2.3.1, with a mean quality cut-off filter of 6. Reads that passed the quality filter were subject to adapter trimming and demultiplexing with Porechop v0.2.4 (available at https://github.com/rrwick/Porechop), discarding the reads that contain internal adapters. Demultiplexed reads derived from the same physical DNA sample but obtained in different sequencing runs were in some instances pooled together to increase coverage of ONT reads. Hybrid *de novo* assemblies combining Illumina and ONT data were performed using Unicycler[262] with bridge application set to 'normal', except for Fission G which was performed in 'bold'. Assemblies were visualised using Bandage[265]. The assembly graphs shown in this work are a direct output of Unicycler and Bandage and were not subject to manual refinement.

5.4.8 Pulse-field gel electrophoresis analysis of fission strains

Clones resulting from fission experiments were grown overnight at 37 °C in 3 mL of LB medium supplemented with appropriate antibiotics; 50 µg/mL of kanamycin for cells with Chr. 2 harbouring a *pheS*-kanR* selection cassette, 20 µg/mL chloramphenicol for *sacB-cat*, and 100 µg/mL streptomycin for the pre-fission MDS42 control. Genomic DNA agarose plugs were prepared from each strain using the BioRad CHEF Genomic DNA Kit, as per manufacturer's instructions. Plugs were stored at 4 °C until used.

Plugs were cut in half (~50 µL per half plug), and subject to a restriction digest with AvrII. For this, half-plugs were rinsed in 0.1X BioRad CHEF Wash Buffer and equilibrated twice in 1X New England Biolabs Cutsmart buffer for 20 minutes at room temperature. The buffer was aspirated and each half-plug was then treated with 25U of AvrII in 100 µL of 1X NEB Cutsmart buffer, and incubated for 2 hours at 37 °C. Following restriction digestion the buffer was aspirated, 150 µL of BioRad CHEF Proteinase K solution was added to each half-plug, and plugs were incubated for another 30 minutes at 37 °C. Plugs were then rinsed with 800 µL of 0.5x TBE, and loaded into a 1% agarose pulse-field gel. Electrophoresis was conducted in a BioRad CHEF-DR-III pulse-field electrophoresis system, in 0.5x TBE, for 22 hours at 14 °C, with 20-120s switch times, 120° angle and 6 V/cm. Gels were stained with 1X SybrSafe and visualized in an Amersham Typhoon Biomolecular Imager.

5.4.9 Analysis of genetic stability of fission products

Cells were grown at 37 °C in 24-well plates, in 3 mL of LB medium supplemented with the appropriate antibiotic (e.g. 20 µg/mL of chloramphenicol for strains with Chr. 2

harbouring a *sacB-cat* selection cassette or 50 μg/mL of kanamycin for strains with Chr. 2 harbouring a *pheS*-kanR* selection cassette) while shaking at 200 rpm. The cultures were passaged by diluting 1/1500x once every twelve hours for a total of 5 days, which yielded a total estimated number of generations of 105[289]. To assess the genetic stability of post-passaging cells, the resulting cultures were used directly for the preparation of genomic DNA agarose plugs for pulse-field gel electrophoresis, as described in 5.4.6.

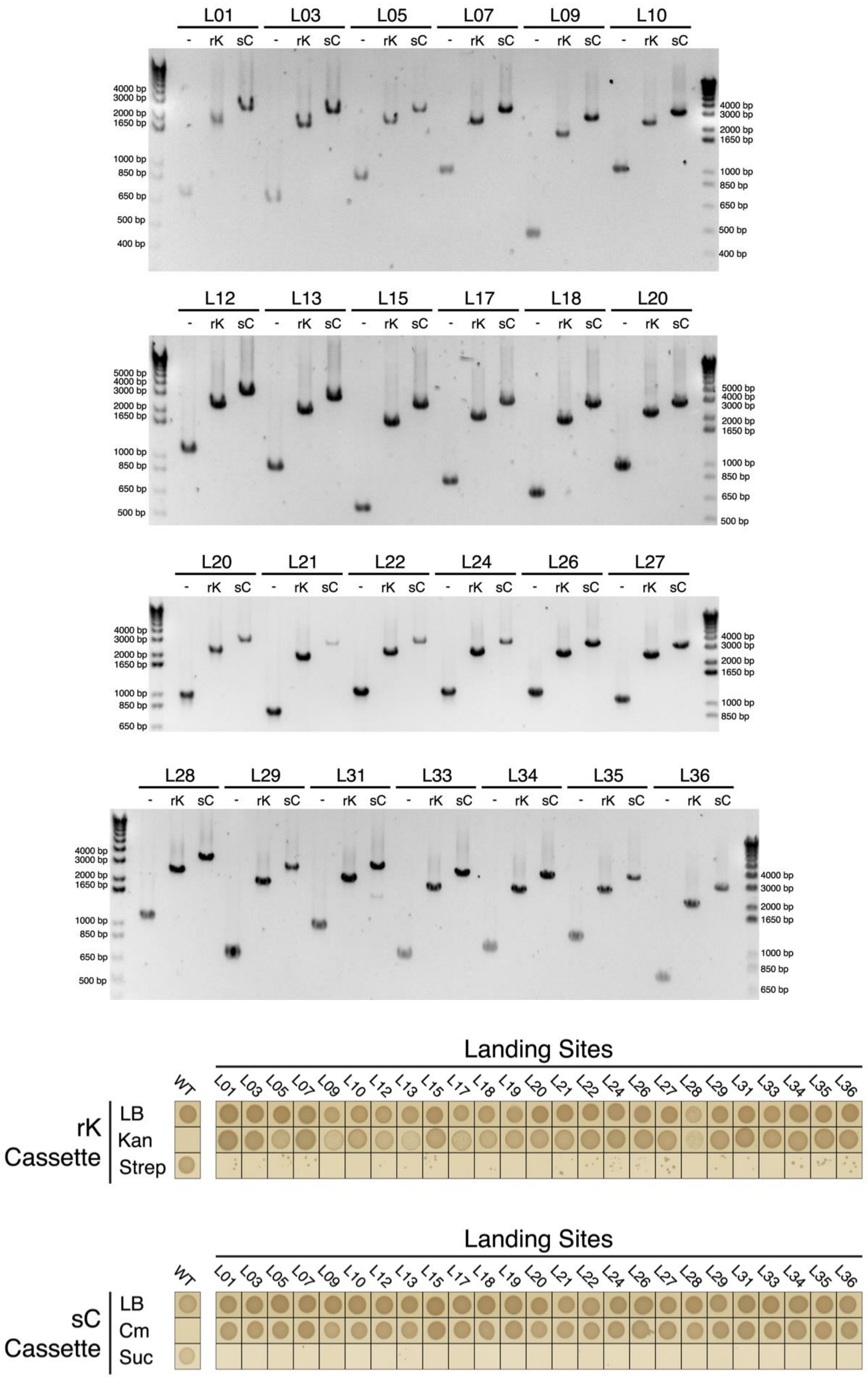
Appendix

Appendix A corresponds to Chapter 2

Appendix B corresponds to Chapter 3

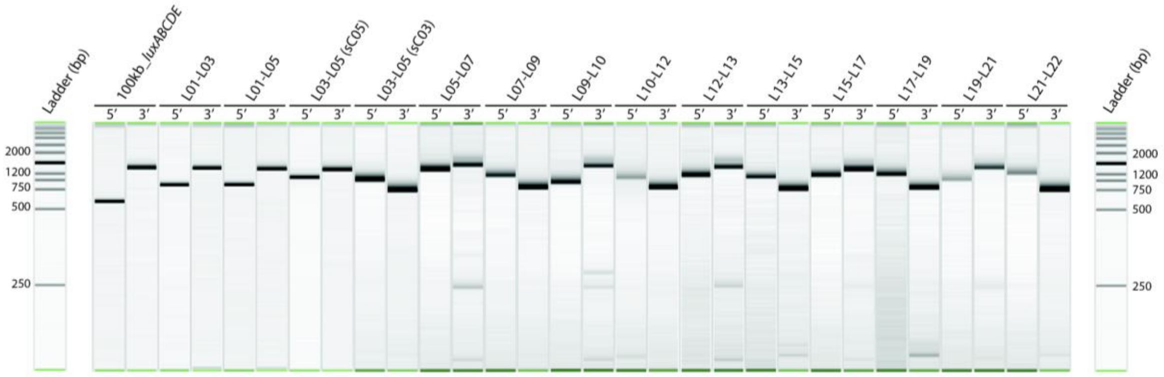
Appendix C corresponds to Chapter 4

A.1 – Integration of selection cassettes at DGF-327 landing site loci



Locus-specific PCR and selective growth assay of DGF-327 strains containing *rpsL-kanR* (rK) or *sacB-cat* (sC) selection cassettes at landing site loci (labelled Lxx, **Figure 2.2**b). '-' stands for wild-type DGF-327. 'Suc' stands for 7.5% sucrose, 'Cm' is 20 µg/mL chloramphenicol, 'Kan' is 50 µg/mL kanamycin, 'Strep' is 100 µg/mL streptomycin.

A.2 – Genotyping vector-insert junctions in construction of DGF-327 REXER BACs



Representative PCR products of the 5’ and 3’ vector-insert junctions for DGF-327 REXER BACs harbouring the indicated genomic regions. PCR products were analysed by capillary gel electrophoresis in a QIAxcel Advanced.

A.3 – Oligos and spacers for Cas9 excision of DGF-327 genomic regions into BACs

Universal gRNA backbone:

5' aaaagcaccgactcggtgccactttttcaagttgataacggactagccttatttaaacttgctatgctgtttccagcat agctcttaaac 3'

Spacer-specific oligo:

5’ GATCACTAATACGACTCACTATA(GG)N_{20}GTTTAAGAGCTATGCTGGAAACAGC

Where N_{20} corresponds to the spacer sequence in the table below. One or two of the G’s in brackets are removed if the first one or two nucleotides of the spacer sequence are G.

Step	Spacer 1 (5'-3')	Spacer 2 (5'-3')
100 kb *luxABCDE*	GAACCGCGTAAGCCAAAACC	GACCAGACAGTCACACACAG
L01-L03	CGTTATTGTTGCAGGCAGTT	GGAAAAGCGAAATTTAAAAG
L03-L05	CTGACTACAACGGTTGGGTT	CTACTATACGCGCGCGATTG
L05-L07	CCGCAGAAGGTTCTTTGGCA	ATGCACTTCTGGCGCACAAG
L07-L09	CTCGCAAATCGTATCGAAAG	ATACTCGACACCTGCTTTAC
L09-L10	TTTTAATCTGCACAGCTTCC	GCGCAGAAAAAGCCTGCCAG
L10-L12	AAGACCGCGAATCGGTTTAA	TGGAACATAGTGCGCAGATC
L12-L13	ATCACCGTAATATTGCCGGA	AACGAAATTATCTTTCAGCA
L13-L15	AACTCTTTCTGCGCTAACTA	CGGAAACGCATAATCCCTCA
L15-L17	AAAATGAGAGGGAATGCTTT	TATAGCCGCTAAGATATTAA
L17-L19	CGCTGGCGCAAAGCCAGTAA	ATACCCGCTCAGAGAATATG
L19-L21	ATGTGGCGGTAACAATCTAC	TTTATTTCTCACATTGATGA
L21-L22	ATTTTCAACATTGTTGCAGC	TGGATATCTGGGGCATGACA

A.4 – Spacer sequences for analysis of DGF-327 REXER BAC integrity and REXER

Step	Spacer 1 (5'-3')	Spacer 2 (5'-3')
L01-L03	GCGCGCTGGCCGTGACCAAACTGCCTGCCT	ATTCAGGAAAAGCGAAATTTAAAAGAGCCT
L03-L05	GGTTTGGGGTTTATATTCACACCCAACCCT	CTTGCCCTCATTCCCAAACCTCAATCGCCT
L05-L07	GCAAGCAATGGGCAAAAAATCCCTTGCCCT	TACGGGCGGTTAAGGTGCCTCTTGTGCCCT
L07-L09	TCATTTACCTGATTAATTGTTCCGCTTCCT	ACCTGCTTTACGGGTGAAAAAAATCAACCT
L09-L10	GCTCCCACCACCGCCACCAGGAAGCTGCCT	AGCCTGCCAGGGGAGAAATCGCAACTGCCT
L10-L12	ATCAAGACCGCGAATCGGTTTAATGGTCCT	CACATCTATCCGGATCTGCGCACTATGCCT
L12-L13	CCGGATTGCGCGTAATCGTCACCATCCCCT	TTCAGCAAGGAGCTGTGAAAATGTCTCCCT
L13-L15	GATCGCGATTTTCCTTAGTTAGCGCAGCCT	TTTTTATTTCCACCGTGAGGGATTATGCCT
L15-L17	TTCAACAATAAAAATGAGAGGGAATGCCCT	ATTAAAGGATGTGTCAAAGATGCATACCCT
L17-L19	ACTCACTGGTCGCTGGCGCAAAGCCAGCCT	GCTAAATCCTTACTTCCGCATATTCTCCCT
L19-L21	CGCGTGTTTATCGCGATAGCAATCGACCCT	ATTTCTCACATTGATGACGGTCGCATGCCT
L21-L22	CATTTTCAACATTGTTGCAGCTGGCAGCCT	GGATATCTGGGGCATGACATGGAAGACCCT

When used for cleaving BACs for integrity analysis by PFGE, the last 20 nt of each sequence were used to design oligonucleotides with structure described in A.3.

B.1 – Efficiency of BAC assembly

		Yeast			*E. coli*
Section	**Fragment**	**# of 10kb stretches**	**# of junctions genotyped**	**Genotyped clones (correct/total)**	**Sequence verified BACs (correct/total)**
H	1	10	8	4/4	4/4
	2	12	7	17/23	5/11
	3	13	0	1/1	1/1
A	4	10	11	7/30	2/3
	5	10	5	23/24	2/4
	6	11	7	6/15	1/4
	7	10	2	16/24	1/4
	8	9	6	13/15	1/6
B	9	12			5/8
	10	10	5	9/22	1/4
	11	10	6	8/8	1/4
	12	11	12	3/4	1/3
	13	11	6	11/22	6/11
C	14	12	7	12/12	4/4
	15	11	7	11/12	4/4
	16	11			4/4
	17	10	6	9/15	3/4
	18	10	11	7/8	1/7
D	19	11	12	4/24	1/3
	20	9			1/3
	21	11	12	3/16	3/3
	22	11	10	3/24	2/3
	23	10	11	4/11	2/4
E	24	10	11	11/11	3/4
	25	10	10	5/24	1/3
	26	12	11	6/7	4/4
	27	12	5	8/24	3/5
	28	13	9	4/24	1/4
F	29	12	13	8/24	1/8
	30	12	9	6/22	1/1
	31	12	12	7/8	6/8
	32	9	9	8/24	1/4
G	33	12	13	6/32	2/4
	34	11	12	8/24	3/5
	35	12	7	5/24	2/3
	36	13	14	4/48	1/1
H	37	14	1	0/56	
	37a	7	7	10/16	3/3
	37b	7	7	1/16	1/1

B.2 – Efficiency of individual REXERs

Section	Frag.	Fully recoded	Spacers				Markers	
			Lin.	pKW1 REXER2	pKW3 REXER2	pKW3 REXER 4	3' to synthetic DNA	on BAC
H	1	2/7*			x		*sC*	*rpsL*
	2	1/5	x				*rK*	*pH*
	3	1/1	x				*sC*	*rpsL*
A	4	1/6		x			*rK*	*sacB*
	5	3/6		x			*sC*	*rpsL*
	6	*N/A*					*rK*	*pH*
	7	3/6	x				*sC*	*rpsL*
	8	*N/A*					*rK*	*pH*
B	9	*N/A*					*sC*	*rpsL*
	10	*N/A*					*rK*	*pH*
	11	1/2	x				*sC*	*rpsL*
	12	2/4	x				*rK*	*pH*
	13	2/4	x				*sC*	*rpsL*
C	14	5/8	x				*rK*	*sacB*
	15	*N/A*					*sC*	*rpsL*
	16	*N/A*					*rK*	*pH*
	17	*N/A*					*sC*	*rpsL*
	18	1/2	x				*rK*	*sacB*
D	19	7/9	x				*sC*	*rpsL*
	20	*N/A*					*rK*	*sacB*
	21	3/5	x				*sC*	*rpsL*
	22	6/6	x				*rK*	*pH*
	23	6/6	x				*sC*	*rpsL*
E	24	2/7	x				*rK*	*pH*
	25	1/3	x				*sC*	*rpsL*
	26	2/3	x				*rK*	*pH*
	27	1/8	x				*sC*	*rpsL*
	28	2/7	x				*rK*	*pH*
F	29	6/6	x				*sC*	*rpsL*
	30	*N/A*					*rK*	*pH*
	31	2/5	x				*sC*	*rpsL*
	32	*N/A*					*rK*	*pH*

G	33	4/8	x				*sC*	*rpsL*
	34	3/5	x				*rK*	*pH*
	35	*N/A*					*sC*	*rpsL*
	36	*N/A*					*rK*	*pH*
H	37a	1/7)			x		*sC*	*rpsL*
	37b	3/6	x				*rK*	*pA*

* indicates efficiency after fixing the fragment, as described in section 3.3. The markers are *rpsL-kanR* (rK), *sacB-cat* (sC), *pheS*-hygR* (pH) and *pheS*-apmR* (pA).

B.3 – Efficiency of REXERs for constructing sections

Section	Frag.	Fully recoded	Spacers				Markers	
			Lin.	pKW1 REXER2	pKW3 REXER2	pKW3 REXER 4	3' to synthetic DNA	on BAC
H	1	*N/A*					*sC*	*rpsL*
	2	2/7			x		*rK*	*pH*
	3	3/5			x		*sC*	*rpsL*
A	4	*N/A*					*rK*	*sacB*
	5	3/6	x				*sC*	*rpsL*
	6	4/6	x				*rK*	*pH*
	7	5/8	x				*sC*	*rpsL*
	8	3/6	x				*rK*	*pH*
B	9	4/5*	x				*sC*	*rpsL*
	10	1/8	x				*rK*	*pH*
	11	2/6	x				*sC*	*rpsL*
	12	*N/A*	-	-	-	-	*rK*	*pH*
	13	7/8	x				*sC*	*rpsL*
C	14	*N/A*					*rK*	*sacB*
	15	3/5	x				*sC*	*rpsL*
	16	4/9	x				*rK*	*pH*
	17	4/8	x				*sC*	*rpsL*
	18	5/10	x				*rK*	*sacB*
D	19	*N/A*					*sC*	*rpsL*
	20	3/4				x	*rK*	*sacB*
	21	1/7	x				*sC*	*rpsL*
	22	6/6	x				*rK*	*pH*
	23	4/6	x				*sC*	*rpsL*
E	24	*N/A*					*rK*	*pH*
	25	2/6	x				*sC*	*rpsL*
	26	4/6	x				*rK*	*pH*
	27	3/6	x				*sC*	*rpsL*
	28	3/8		x			*rK*	*pH*
F	29	*N/A*					*sC*	*rpsL*
	30	2/3	x				*rK*	*pH*
	31	2/10	x				*sC*	*rpsL*

	32	4/4	x				*rK*	*pH*
G	33	*N/A*					*sC*	*rpsL*
	34	1/8	x				*rK*	*pH*
	35	6/6	x				*sC*	*rpsL*
	36	3/7	x				*rK*	*pH*
H	37a	*N/A*						*rpsL*
	37b	3/5	x				*rK*	*pA*

* indicates efficiency after fixing the fragment, as described in section 3.3. The markers are *rpsL-kanR* (rK), *sacB-cat* (sC), *pheS*-hygR* (pH) and *pheS*-apmR* (pA).

The efficiency for REXERs at the beginning of sections in indicated in Appendix B.2. Fragment 12 was not introduced by REXER, but conjugation.

B.4 – Deviations from original design

Design optimisations						
Section	**Frag.**	**Position***	**Original design**	**Final genome**	**Consequence**	**Origin**
H	37a	16,213	AGT	AGC	Viable recoding of S70 in *yaaY*	
H	1	88,037	1nt+TAA +20nt	4nt+TGA +182nt	Viable separation of *ftsI* and *murE*	
H	1	178,509	AGT	TCT	Viable recoding of S4 in *map*	
B	9	976,671	AGC	TGA	Disruption of pseudogene *yceQ* to preserve vialbe expression of *rne*	
		976,686	AGT	TCA		
		976,710	AGT	TCA		
		976,836	AGC	TCG		
		976,899	AGT	TCA		
Non-programmed mutations						
H	37b	53,145	G	A	Intergenic region	In DNA synthesis or BAC assembly
C	15	1,579,495	C	T	D434D in *sdaA* (non-essential gene)	In DNA synthesis or BAC assembly
D	21	2,288,863	T	-	Deletion in *yfiL* (non-essential gene)	During recoding
E	27	2,885,875	A	G	T369A in *acrF* (non-essential gene)	In transfer from DH10b to MDS42
E	28	3,031,081	C	A	S119I in *gntK* (non-essential gene)	In DNA synthesis or BAC assembly
F	30	3,252,858	T	C	S10S in *gmK* (essential gene)	During recoding
F	30	3,252,920	A	G	Y31C in *gmK* (essential gene)	During recoding
F	30	3,319,703	A	G	Intergenic region	During recoding

The position indicates the coordinates in the genome design. The sequence is in Supplementary Data 2 of ref[284].

C.1 – Efficiency of different fission experiments

Experiment	Fraction of 200mL culture plated	Total colonies on plate	Colonies too luminescent	Colonies correct luminescence	Colonies no luminescence	Selective growth correct	PCR across junctions correct	De novo assembly correct
Fission 0.56 Mb	0.001	86	1	64	21	16 out of 16	15 out of 16	1 out 1
Fission 2.1 Mb	0.1	192	158	27	7		4 out of 8	0 out of 2
Fission 0.53 Mb (section D)	0.01	67	11	7	49	5 out of 7	5 out of 5	1 out of 1
Fission 0.59 Mb (section E)	0.1	164	54	109	1	9 out of 9	5 out of 5	1 out of 1
Fission 0.44 Mb (section G)	0.1	365	125	13	227	6 out of 8	5 out of 5	1 out of 1
Fission 0.48 Mb (section H)	0.01	114	36	60	18	9 out of 9	5 out of 5	1 out of 1
Fission 1.55 Mb (section ABC)	0.01	81	2	76	3	32 out of 32	16 out of 16	1 out of 1
Fission 0.54 Mb (watermarked Section C)	0.01	47	13	25	9	27 out of 28	10 out of 10	1 out of 1
Fission 0.54 Mb (non-watermarked Section C)	1	35	10	22	3	18 out of 20	10 out of 10	1 out of 1
Fusion regeneration (Section C)	0.1	Thousands	NA	NA	NA	8 out of 8	8 out of 8	1 out of 1
Translocation 700 kb away	0.1	Thousands	NA	NA	NA	8 out of 8	8 out of 8	1 out of 1
Translocation 500 kb away	0.1	Thousands	NA	NA	NA	8 out of 8	8 out of 8	1 out of 1
Translocation 1.8 Mb away	0.1	50	NA	NA	NA	8 out of 8	8 out of 8	0 out of 1
Inversion	0.1	Thousands	NA	NA	NA	8 out of 8	8 out of 8	1 out of 1

C.2 – Deposition of de genome assemblies and raw sequencing data

Experiment	Genbank #	SRA #
Fission 0.56 Mb	VMRF00000000	SRR9671404, SRR9661817
Fission 2.1 Mb clone 1	CP041993	SRR9671398, SRR9661793
Fission 2.1 Mb clone 2	CP041992	SRR9671397, SRR9661791
Fission 0.53 Mb (section D)	VMRE00000000	SRR9671403, SRR9661819
Fission 0.59 Mb (section E)	VMRD00000000	SRR9671406, SRR9661813
Fission 0.44 Mb (section G)	VMRC00000000	SRR9671405, SRR9661815
Fission 0.48 Mb (section H)	VMRB00000000	SRR9671400, SRR9661811
Fission 1.55 Mb (section ABC)	VMRA00000000	SRR9671399, SRR9661799
Fission 0.54 Mb (watermarked Section C)	VMQZ00000000	SRR9671402, SRR9661797
Fission 0.54 Mb (non-watermarked Section C)	VMQY00000000	SRR9671401, SRR9661795
Fusion regeneration (section C)	CP041991	SRR9671412, SRR9661808
Translocation 700 kb away	VMQX00000000	SRR9671410, SRR9661804
Translocation 500 kb away	CP041989	SRR9671409, SRR9661802
Translocation 1.8 Mb away	N/A	SRR9661776
Inversion	CP041990	SRR9671411, SRR9661806
Post-transplant of watermarked Section C into strain with watermarked Section A	VMQW00000000	SRR9671408, SRR9661810
Post-fusion to combine watermarked Sections A and C	CP041988	SRR9671407, SRR9661789

Genbank files containing annotated sequences of the strains resulting from fission and fusions experiments were deposited, with the Genbank identifier number (Genbank #) of each indicated strain provided. Raw sequencing data from both Illumina and Oxford Nanopore Technologies NGS platforms was deposited on the sequence read archive (SRA) (https://wwww.ncbi.nlm.nih.gov/sra), with the SRA identifier number (SRA #) provided for each indicated strain.

C.3 – Copy numbers of post-fission chromosomes

Experiment	Chr. 2 Relative copy number
Fission 0.56 Mb	1.21x
Fission 0.53 Mb (section D)	1.82x
Fission 0.59 Mb (section E)	1.55x
Fission 0.44 Mb (section G)	1.97x
Fission 0.48 Mb (section H)	1.99x
Fission 1.55 Mb (section ABC)	1.01x
Fission 0.54 Mb (watermarked Section C)	1.47x
Fission 0.54 Mb (non-watermarked Section C)	1.09x

The copy number of Chromosome 2 (Chr. 2) generated by fission, relative to Chromosome 1, is provided for the indicated post-fission strain. The copy number was calculated from the *de novo* assembly data by the program Unicycler[262].

References

1. Crick, F. H. C., Barnett, L., Brenner, S. & Watts-Tobin, R. J. General Nature of the Genetic Code for Proteins. *Nature* **192**, 1227–1232 (1961).

2. Dunill, P. Triplet Nucleotide–Amino-acid Pairing; a Stereo-chemical Basis for the Division between Protein and Non-protein Amino-acids. *Nature* **210**, 1267–1268 (1966).

3. Woese, C. R., Dugre, D. H., Saxinger, W. C. & Dugre, S. A. The molecular basis for the genetic code. *Proc. Natl. Acad. Sci. U. S. A.* **55**, 966–974 (1966).

4. Crick, F. H. C. The origin of the genetic code. *J. Mol. Biol.* **38**, 367–379 (1968).

5. Sonneborn, T. M. Degeneracy of the Genetic Code: Extent, Nature, and Genetic Implications. in *Evolving Genes and Proteins* (eds. Bryson, V. & Vogel, H. J. B. T.-E. G. and P.) 377–397 (Academic Press, 1965). doi:https://doi.org/10.1016/B978-1-4832-2734-4.50034-6.

6. Woese, C. R. On the evolution of the genetic code. *Proc. Natl. Acad. Sci. U. S. A.* **54**, 1546–1552 (1965).

7. Vetsigian, K., Woese, C. & Goldenfeld, N. Collective evolution and the genetic code. *Proc. Natl. Acad. Sci.* **103**, 10696–10701 (2006).

8. Aggarwal, N., Bandhu, A. V. & Sengupta, S. Finite population analysis of the effect of horizontal gene transfer on the origin of an universal and optimal genetic code. *Phys. Biol.* **13**, 036007 (2016).

9. Koonin, E. V & Novozhilov, A. S. Origin and evolution of the genetic code: the universal enigma. *IUBMB Life* **61**, 99–111 (2009).

10. Koonin, E. V & Novozhilov, A. S. Origin and Evolution of the Universal Genetic Code. *Annu. Rev. Genet.* **51**, 45–62 (2017).

11. Barrell, B. G., Bankier, A. T. & Drouin, J. A different genetic code in human mitochondria. *Nature* **282**, 189–194 (1979).

12. Yamao, F. *et al.* UGA is read as tryptophan in Mycoplasma capricolum. *Proc. Natl. Acad. Sci.* **82**, 2306–2309 (1985).

13. Oba, T., Andachi, Y., Muto, A. & Osawa, S. CGG: an unassigned or nonsense codon in Mycoplasma capricolum. *Proc. Natl. Acad. Sci.* **88**, 921–925 (1991).

14. Santos, M. A., Keith, G. & Tuite, M. F. Non-standard translational events in Candida albicans mediated by an unusual seryl-tRNA with a 5′-CAG-3′ (leucine) anticodon. *EMBO J.* **12**, 607–616 (1993).

15. Santos, M. A. S. & Tuite, M. F. The CUG codon is decoded in vivo as serine and not leucine in Candida albicans. *Nucleic Acids Res.* **23**, 1481–1486 (1995).

16. Tourancheau, A. B., Tsao, N., Klobutcher, L. A., Pearlman, R. E. & Adoutte, A. Genetic code deviations in the ciliates: evidence for multiple and independent events. *EMBO J.* **14**, 3262–3267 (1995).

17. Roberts, J. W. & Carbon, J. Molecular mechanism for missense suppression in E. coli. *Nature* **250**, 412–414 (1974).

18. Hirsh, D. Tryptophan transfer RNA as the UGA suppressor. *J. Mol. Biol.* **58**, 439–444 (1971).

19. Andersson, S. G. E. & Kurland, C. G. Genomic evolution drives the evolution of the translation system. *Biochem. Cell Biol.* **73**, 775–787 (1995).

20. Knight, R. D., Freeland, S. J. & Landweber, L. F. Rewiring the keyboard: evolvability of the genetic code. *Nat. Rev. Genet.* **2**, 49–58 (2001).

21. Osawa, S., Jukes, T. H., Watanabe, K. & Muto, A. Recent evidence for evolution of the genetic code. *Microbiol. Rev.* **56**, 229–264 (1992).

22. Santos, M. A. S., Cheesman, C., Costa, V., Moradas-Ferreira, P. & Tuite, M. F. Selective advantages created by codon ambiguity allowed for the evolution of an alternative genetic code in Candida spp. *Mol. Microbiol.* **31**, 937–947 (1999).

23. Schultz, D. W. & Yarus, M. Transfer RNA mutation and the malleability of the genetic code. *J. Mol. Biol.* **235**, 1377–1380 (1994).

24. Santos, M. A. S., Ueda, T., Watanabe, K. & Tuite, M. F. The non-standard genetic code of Candida spp.: an evolving genetic code or a novel mechanism for adaptation? *Mol. Microbiol.* **26**, 423–431 (1997).

25. Wang, K., Schmied, W. H. & Chin, J. W. Reprogramming the genetic code: From triplet to quadruplet codes. *Angew. Chemie - Int. Ed.* **51**, 2288–2297 (2012).

26. Chin, J. W. Expanding and Reprogramming the Genetic Code of Cells and Animals. *Annu. Rev. Biochem.* **83**, 379–408 (2014).

27. Chin, J. W. Expanding and reprogramming the genetic code. *Nature* **550**, 53–60 (2017).

28. Liu, C. C. & Schultz, P. G. Adding New Chemistries to the Genetic Code. *Annu. Rev. Biochem.* **79**, 413–444 (2010).

29. Giegé, R., Sissler, M. & Florentz, C. Universal rules and idiosyncratic features in tRNA identity. *Nucleic Acids Res.* **26**, 5017–5035 (1998).

30. Chin, J. W. Modular approaches to expanding the functions of living matter. *Nat. Chem. Biol.* **2**, 304–311 (2006).

31. Liu, D. R., Magliery, T. J., Pastrnak, M. & Schultz, P. G. Engineering a tRNA and aminoacyl-tRNA synthetase for the site-specific incorporation of unnatural amino acids into proteins in vivo. *Proc. Natl. Acad. Sci.* **94**, 10092–10097 (1997).

32. Santoro, S. W., Wang, L., Herberich, B., King, D. S. & Schultz, P. G. An efficient system for the evolution of aminoacyl-tRNA synthetase specificity. *Nat. Biotechnol.* **20**, 1044–1048 (2002).

33. Chin, J. W. *et al.* Addition of p-Azido-l-phenylalanine to the Genetic Code of Escherichia coli. *J. Am. Chem. Soc.* **124**, 9026–9027 (2002).

34. Wang, L., Brock, A., Herberich, B. & Schultz, P. G. Expanding the Genetic Code of

Escherichia coli. *Science.* **292**, 498–500 (2001).

35. Zinoni, F., Birkmann, A., Stadtman, T. C. & Böck, A. Nucleotide sequence and expression of the selenocysteine-containing polypeptide of formate dehydrogenase (formate-hydrogen-lyase-linked) from Escherichia coli. *Proc. Natl. Acad. Sci.* **83**, 4650–4654 (1986).

36. Zinoni, F., Birkmann, A., Leinfelder, W. & Böck, A. Cotranslational insertion of selenocysteine into formate dehydrogenase from Escherichia coli directed by a UGA codon. *Proc. Natl. Acad. Sci.* **84**, 3156–3160 (1987).

37. Böck, A. *et al.* Selenocysteine: the 21st amino acid. *Mol. Microbiol.* **5**, 515–520 (1991).

38. Forchhammer, K., Leinfelder, W. & Böck, A. Identification of a novel translation factor necessary for the incorporation of selenocysteine into protein. *Nature* **342**, 453–456 (1989).

39. Zinoni, F., Heider, J. & Böck, A. Features of the formate dehydrogenase mRNA necessary for decoding of the UGA codon as selenocysteine. *Proc. Natl. Acad. Sci.* **87**, 4660 LP – 4664 (1990).

40. Low, S. C. & Berry, M. J. Knowing when not to stop: selenocysteine incorporation in eukaryotes. *Trends Biochem. Sci.* **21**, 203–208 (1996).

41. Stadtman, T. C. Selenocysteine. *Annu. Rev. Biochem.* **65**, 83–100 (1996).

42. Bröcker, M. J., Ho, J. M. L., Church, G. M., Söll, D. & O'Donoghue, P. Recoding the genetic code with selenocysteine. *Angew. Chem. Int. Ed. Engl.* **53**, 319–323 (2014).

43. Zhang, Y., Baranov, P. V., Atkins, J. F. & Gladyshev, V. N. Pyrrolysine and selenocysteine use dissimilar decoding strategies. *J. Biol. Chem.* **280**, 20740–51 (2005).

44. Srinivasan, G., James, C. M. & Krzycki, J. A. Pyrrolysine Encoded by UAG in Archaea: Charging of a UAG-Decoding Specialized tRNA. *Science.* **296**, 1459–1462 (2002).

45. Hao, B. *et al.* A New UAG-Encoded Residue in the Structure of a Methanogen Methyltransferase. *Science.* **296**, 1462–1466 (2002).

46. Blight, S. K. *et al.* Direct charging of tRNACUA with pyrrolysine in vitro and in vivo. *Nature* **431**, 333–335 (2004).

47. Polycarpo, C. *et al.* An aminoacyl-tRNA synthetase that specifically activates pyrrolysine. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 12450–12454 (2004).

48. Mukai, T. *et al.* Adding l-lysine derivatives to the genetic code of mammalian cells with engineered pyrrolysyl-tRNA synthetases. *Biochem. Biophys. Res. Commun.* **371**, 818–822 (2008).

49. Kavran, J. M. *et al.* Structure of pyrrolysyl-tRNA synthetase, an archaeal enzyme for genetic code innovation. *Proc. Natl. Acad. Sci.* **104**, 11268–11273 (2007).

50. Nozawa, K. *et al.* Pyrrolysyl-tRNA synthetase–tRNAPyl structure reveals the

molecular basis of orthogonality. *Nature* **457**, 1163–1167 (2009).

51. Neumann, H., Peak-Chew, S. Y. & Chin, J. W. Genetically encoding Nε-acetyllysine in recombinant proteins. *Nat. Chem. Biol.* **4**, 232–234 (2008).

52. Wang, Y.-S., Fang, X., Wallace, A. L., Wu, B. & Liu, W. R. A Rationally Designed Pyrrolysyl-tRNA Synthetase Mutant with a Broad Substrate Spectrum. *J. Am. Chem. Soc.* **134**, 2950–2953 (2012).

53. Tuley, A. *et al.* The genetic incorporation of thirteen novel non-canonical amino acids. *Chem. Commun.* **50**, 2673–2675 (2014).

54. Nguyen, D. P. *et al.* Genetic Encoding and Labeling of Aliphatic Azides and Alkynes in Recombinant Proteins via a Pyrrolysyl-tRNA Synthetase/tRNACUA Pair and Click Chemistry. *J. Am. Chem. Soc.* **131**, 8720–8721 (2009).

55. Kobayashi, T., Yanagisawa, T., Sakamoto, K. & Yokoyama, S. Recognition of Non-α-amino Substrates by Pyrrolysyl-tRNA Synthetase. *J. Mol. Biol.* **385**, 1352–1360 (2009).

56. Wang, L., Magliery, T. J., Liu, D. R. & Schultz, P. G. A New Functional Suppressor tRNA/Aminoacyl−tRNA Synthetase Pair for the in Vivo Incorporation of Unnatural Amino Acids into Proteins. *J. Am. Chem. Soc.* **122**, 5010–5011 (2000).

57. Chin, J. W. *et al.* An Expanded Eukaryotic Genetic Code. *Science.* **301**, 964–967 (2003).

58. Wu, N., Deiters, A., Cropp, T. A., King, D. & Schultz, P. G. A Genetically Encoded Photocaged Amino Acid. *J. Am. Chem. Soc.* **126**, 14306–14307 (2004).

59. Ohno, S. *et al.* Co-Expression of Yeast Amber Suppressor tRNATyr and Tyrosyl-tRNA Synthetase in Escherichia coli: Possibility to Expand the Genetic Code. *J. Biochem.* **124**, 1065–1068 (1998).

60. Pastrnak, M., Magliery, T. J. & Schultz, P. G. A New Orthogonal Suppressor tRNA/Aminoacyl-tRNA Synthetase Pair for Evolving an Organism with an Expanded Genetic Code. *Helv. Chim. Acta* **83**, 2277–2286 (2000).

61. Furter, R. Expansion of the genetic code: Site-directed p-fluoro-phenylalanine incorporation in Escherichia coli. *Protein Sci.* **7**, 419–426 (1998).

62. Hughes, R. A. & Ellington, A. D. Rational design of an orthogonal tryptophanyl nonsense suppressor tRNA. *Nucleic Acids Res.* **38**, 6813–6830 (2010).

63. Chatterjee, A., Xiao, H., Yang, P.-Y., Soundararajan, G. & Schultz, P. G. A Tryptophanyl-tRNA Synthetase/tRNA Pair for Unnatural Amino Acid Mutagenesis in E. coli. *Angew. Chemie Int. Ed.* **52**, 5106–5109 (2013).

64. Nguyen, D. P., Alai, M. M. G., Virdee, S. & Chin, J. W. Genetically Directing ε-N, N-Dimethyl-l-Lysine in Recombinant Histones. *Chem. Biol.* **17**, 1072–1076 (2010).

65. Ai, H., Lee, J. W. & Schultz, P. G. A method to site-specifically introduce methyllysine into proteins in E. coli. *Chem. Commun.* **46**, 5506–5508 (2010).

66. Wang, Y.-S. *et al.* A genetically encoded photocaged Nε-methyl-l-lysine. *Mol. Biosyst.* **6**, 1557–1560 (2010).

67. Neumann, H. *et al.* A Method for Genetically Installing Site-Specific Acetylation in Recombinant Histones Defines the Effects of H3 K56 Acetylation. *Mol. Cell* **36**, 153–163 (2009).

68. Park, H.-S. *et al.* Expanding the Genetic Code of Escherichia coli with Phosphoserine. *Science.* **333**, 1151 LP – 1154 (2011).

69. Rogerson, D. T. *et al.* Efficient genetic encoding of phosphoserine and its nonhydrolyzable analog. *Nat. Chem. Biol.* **11**, 496–503 (2015).

70. Zhang, M. S. *et al.* Biosynthesis and genetic encoding of phosphothreonine through parallel selection and deep sequencing. *Nat. Methods* **14**, 729–736 (2017).

71. Hoppmann, C. *et al.* Site-specific incorporation of phosphotyrosine using an expanded genetic code. *Nat. Chem. Biol.* **13**, 842–844 (2017).

72. Luo, X. *et al.* Genetically encoding phosphotyrosine and its nonhydrolyzable analog in bacteria. *Nat. Chem. Biol.* **13**, 845–849 (2017).

73. Fan, C., Ip, K. & Söll, D. Expanding the genetic code of Escherichia coli with phosphotyrosine. *FEBS Lett.* **590**, 3040–3047 (2016).

74. Davis, L. & Chin, J. W. Designer proteins: applications of genetic code expansion in cell biology. *Nat. Rev. Mol. Cell Biol.* **13**, 168–182 (2012).

75. Uttamapinant, C. *et al.* Genetic Code Expansion Enables Live-Cell and Super-Resolution Imaging of Site-Specifically Labeled Cellular Proteins. *J. Am. Chem. Soc.* **137**, 4602–4605 (2015).

76. Lang, K. & Chin, J. W. Cellular Incorporation of Unnatural Amino Acids and Bioorthogonal Labeling of Proteins. *Chem. Rev.* **114**, 4764–4806 (2014).

77. Aloush, N. *et al.* Live Cell Imaging of Bioorthogonally Labelled Proteins Generated With a Single Pyrrolysine tRNA Gene. *Sci. Rep.* **8**, 14527 (2018).

78. Elliott, T. S. *et al.* Proteome labeling and protein identification in specific tissues and at specific developmental stages in an animal. *Nat. Biotechnol.* **32**, 465–472 (2014).

79. Elliott, T. S., Bianco, A., Townsley, F. M., Fried, S. D. & Chin, J. W. Tagging and Enriching Proteins Enables Cell-Specific Proteomics. *Cell Chem. Biol.* **23**, 805–815 (2016).

80. Krogager, T. P. *et al.* Labeling and identifying cell-specific proteomes in the mouse brain. *Nat. Biotechnol.* **36**, 156–159 (2018).

81. Dieterich, D. C., Link, A. J., Graumann, J., Tirrell, D. A. & Schuman, E. M. Selective identification of newly synthesized proteins in mammalian cells using bioorthogonal noncanonical amino acid tagging (BONCAT). *Proc. Natl. Acad. Sci.* **103**, 9482 LP – 9487 (2006).

82. Arbely, E., Torres-Kolbus, J., Deiters, A. & Chin, J. W. Photocontrol of Tyrosine Phosphorylation in Mammalian Cells via Genetic Encoding of Photocaged Tyrosine. *J. Am. Chem. Soc.* **134**, 11912–11915 (2012).

83. Lemke, E. A., Summerer, D., Geierstanger, B. H., Brittain, S. M. & Schultz, P. G. Control of protein phosphorylation with a genetically encoded photocaged amino acid. *Nat. Chem. Biol.* **3**, 769–772 (2007).

84. Gautier, A. *et al.* Genetically Encoded Photocontrol of Protein Localization in Mammalian Cells. *J. Am. Chem. Soc.* **132**, 4086–4088 (2010).

85. Hayashi, T., Hilvert, D. & Green, A. P. Engineered Metalloenzymes with Non-Canonical Coordination Environments. *Chem. – A Eur. J.* **24**, 11821–11830 (2018).

86. Burke, A. J. *et al.* Design and evolution of an enzyme with a non-canonical organocatalytic mechanism. *Nature* **570**, 219–223 (2019).

87. Wang, K., Neumann, H., Peak-Chew, S. Y. & Chin, J. W. Evolved orthogonal ribosomes enhance the efficiency of synthetic genetic code expansion. *Nat. Biotechnol.* **25**, 770–777 (2007).

88. Rackham, O. & Chin, J. W. A network of orthogonal ribosome·mRNA pairs. *Nat. Chem. Biol.* **1**, 159–166 (2005).

89. Shimizu, Y. *et al.* Cell-free translation reconstituted with purified components. *Nat. Biotechnol.* **19**, 751–755 (2001).

90. Short, G. F., Golovine, S. Y. & Hecht, S. M. Effects of Release Factor 1 on in Vitro Protein Translation and the Elaboration of Proteins Containing Unnatural Amino Acids. *Biochemistry* **38**, 8808–8819 (1999).

91. Agafonov, D. E., Huang, Y., Grote, M. & Sprinzl, M. Efficient suppression of the amber codon in E. coli in vitro translation system. *FEBS Lett.* **579**, 2156–2160 (2005).

92. Sando, S., Ogawa, A., Nishi, T., Hayami, M. & Aoyama, Y. In vitro selection of RNA aptamer against Escherichia coli release factor 1. *Bioorg. Med. Chem. Lett.* **17**, 1216–1220 (2007).

93. Ito, K., Uno, M. & Nakamura, Y. Single amino acid substitution in prokaryote polypeptide release factor 2 permits it to terminate translation at all three stop codons. *Proc. Natl. Acad. Sci.* **95**, 8165–8169 (1998).

94. Johnson, D. B. F. *et al.* RF1 knockout allows ribosomal incorporation of unnatural amino acids at multiple sites. *Nat. Chem. Biol.* **7**, 779–786 (2011).

95. Johnson, D. B. F. *et al.* Release Factor One Is Nonessential in Escherichia coli. *ACS Chem. Biol.* **7**, 1337–1344 (2012).

96. Uno, M., Ito, K. & Nakamura, Y. Functional specificity of amino acid at position 246 in the tRNA mimicry domain of bacterial release factor 2. *Biochimie* **78**, 935–943 (1996).

97. Mukai, T. *et al.* Codon reassignment in the Escherichia coli genetic code. *Nucleic Acids Res.* **38**, 8188–8195 (2010).

98. Mukai, T. *et al.* Highly reproductive Escherichia coli cells with no specific assignment to the UAG codon. *Sci. Rep.* **5**, 9699 (2015).

99. Isaacs, F. J. *et al.* Precise Manipulation of Chromosomes in Vivo Enables Genome-Wide Codon Replacement. *Science.* **333**, 348–353 (2011).

100. Lajoie, M. J. *et al.* Genomically Recoded Organisms Expand Biological Functions. *Science.* **342**, 357–360 (2013).

101. Wang, H. H. *et al.* Programming cells by multiplex genome engineering and accelerated evolution. *Nature* **460**, 894–898 (2009).

102. Ellis, H. M., Yu, D., DiTizio, T. & Court, D. L. High efficiency mutagenesis, repair, and engineering of chromosomal DNA using single-stranded oligonucleotides. *Proceedings of the National Academy of Sciences of the United States of America.* vol. 98 6742–6746 (2001).

103. Amiram, M. *et al.* Evolution of translation components in recoded organisms enables multi-site nonstandard amino acid incorporation in proteins at high yield and purity. *Nat. Biotechnol.* **33**, (2015).

104. Wannier, T. M. *et al.* Adaptive evolution of genomically recoded Escherichia coli. *Proc. Natl. Acad. Sci.* **115**, 3090–3095 (2018).

105. Schmied, W. H., Elsässer, S. J., Uttamapinant, C. & Chin, J. W. Efficient Multisite Unnatural Amino Acid Incorporation in Mammalian Cells via Optimized Pyrrolysyl tRNA Synthetase/tRNA Expression and Engineered eRF1. *J. Am. Chem. Soc.* **136**, 15577–15583 (2014).

106. Reinkemeier, C. D., Girona, G. E. & Lemke, E. A. Designer membraneless organelles enable codon reassignment of selected mRNAs in eukaryotes. *Science.* **363**, eaaw2644 (2019).

107. Wan, W. *et al.* A Facile System for Genetic Incorporation of Two Different Noncanonical Amino Acids into One Protein in Escherichia coli. *Angew. Chemie Int. Ed.* **49**, 3211–3214 (2010).

108. Neumann, H., Wang, K., Davis, L., Garcia-Alai, M. & Chin, J. W. Encoding multiple unnatural amino acids via evolution of a quadruplet-decoding ribosome. *Nature* **464**, 441–444 (2010).

109. Wu, B., Wang, Z., Huang, Y. & Liu, W. R. Catalyst-Free and Site-Specific One-Pot Dual-Labeling of a Protein Directed by Two Genetically Incorporated Noncanonical Amino Acids. *ChemBioChem* **13**, 1405–1408 (2012).

110. Chatterjee, A., Sun, S. B., Furman, J. L., Xiao, H. & Schultz, P. G. A Versatile Platform for Single- and Multiple-Unnatural Amino Acid Mutagenesis in Escherichia coli. *Biochemistry* **52**, 1828–1837 (2013).

111. Wang, K. *et al.* Optimized orthogonal translation of unnatural amino acids enables spontaneous protein double-labelling and FRET. *Nat. Chem.* **6**, 393–403 (2014).

112. Neumann, H., Slusarczyk, A. L. & Chin, J. W. De Novo generation of mutually orthogonal aminoacyl-tRNA synthetase/ tRNA pairs. *J. Am. Chem. Soc.* **132**, 2142–2144 (2010).

113. Willis, J. C. W. & Chin, J. W. Mutually orthogonal pyrrolysyl-tRNA

synthetase/tRNA pairs. *Nat. Chem.* **10**, 831–837 (2018).

114. Italia, J. S. *et al.* Mutually Orthogonal Nonsense-Suppression Systems and Conjugation Chemistries for Precise Protein Labeling at up to Three Distinct Sites. *J. Am. Chem. Soc.* **141**, 6204–6212 (2019).

115. Italia, J. S. *et al.* An orthogonalized platform for genetic code expansion in both bacteria and eukaryotes. *Nat. Chem. Biol.* **13**, 446–450 (2017).

116. Bossi, L. & Roth, J. R. Four-base codons ACCA, ACCU and ACCC are recognized by frameshift suppressor sufJ. *Cell* **25**, 489–496 (1981).

117. Roth, J. R. Frameshift suppression. *Cell* **24**, 601–602 (1981).

118. Bossi, L. & Smith, D. M. Suppressor sufJ: a novel type of tRNA mutant that induces translational frameshifting. *Proc. Natl. Acad. Sci. U. S. A.* **81**, 6105–6109 (1984).

119. Curran, J. F. & Yarus, M. Reading frame selection and transfer RNA anticodon loop stacking. *Science.* **238**, 1545–1550 (1987).

120. Magliery, T. J., Anderson, J. C. & Schultz, P. G. Expanding the genetic code: selection of efficient suppressors of four-base codons and identification of "shifty" four-base codons with a library approach in Escherichia coli. *J. Mol. Biol.* **307**, 755–769 (2001).

121. Anderson, J. C. *et al.* An expanded genetic code with a functional quadruplet codon. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 7566–7571 (2004).

122. Chatterjee, A., Lajoie, M. J., Xiao, H., Church, G. M. & Schultz, P. G. A Bacterial Strain with a Unique Quadruplet Codon Specifying Non-native Amino Acids. *ChemBioChem* **15**, 1782–1786 (2014).

123. Benner, S. A. Understanding Nucleic Acids Using Synthetic Chemistry. *Acc. Chem. Res.* **37**, 784–797 (2004).

124. Hirao, I., Kimoto, M. & Yamashige, R. Natural versus Artificial Creation of Base Pairs in DNA: Origin of Nucleobases from the Perspectives of Unnatural Base Pair Studies. *Acc. Chem. Res.* **45**, 2055–2065 (2012).

125. Feldman, A. W. & Romesberg, F. E. Expansion of the Genetic Alphabet: A Chemist's Approach to Synthetic Biology. *Acc. Chem. Res.* **51**, 394–403 (2018).

126. Malyshev, D. a. *et al.* Efficient and sequence-independent replication of DNA containing a third base pair establishes a functional six-letter genetic alphabet. *Proc. Natl. Acad. Sci.* **109**, 12005–12010 (2012).

127. Yang, Z., Chen, F., Alvarado, J. B. & Benner, S. A. Amplification, Mutation, and Sequencing of a Six-Letter Synthetic Genetic System. *J. Am. Chem. Soc.* **133**, 15105–15112 (2011).

128. Yamashige, R. *et al.* Highly specific unnatural base pair systems as a third base pair for PCR amplification. *Nucleic Acids Res.* **40**, 2793–2806 (2011).

129. Malyshev, D. A. & Romesberg, F. E. The Expanded Genetic Alphabet. *Angew. Chemie Int. Ed.* **54**, 11930–11944 (2015).

130. Bain, J. D., Switzer, C., Chamberlin, R. & Benner, S. A. Ribosome-mediated incorporation of a non-standard amino acid into a peptide through expansion of the genetic code. *Nature* **356**, 537–539 (1992).

131. Hoshika, S. *et al.* Hachimoji DNA and RNA: A genetic system with eight building blocks. *Science.* **363**, 884–887 (2019).

132. Seo, Y. J., Malyshev, D. A., Lavergne, T., Ordoukhanian, P. & Romesberg, F. E. Site-Specific Labeling of DNA and RNA Using an Efficiently Replicated and Transcribed Class of Unnatural Base Pairs. *J. Am. Chem. Soc.* **133**, 19878–19888 (2011).

133. Malyshev, D. A. *et al.* A semi-synthetic organism with an expanded genetic alphabet. *Nature* **509**, 385–388 (2014).

134. Zhang, Y. *et al.* A semi-synthetic organism that stores and retrieves increased genetic information. *Nature* **551**, 644–647 (2017).

135. Zhang, Y. *et al.* A semisynthetic organism engineered for the stable expansion of the genetic alphabet. *Proc. Natl. Acad. Sci.* **114**, 1317–1322 (2017).

136. Ravikumar, A., Arrieta, A. & Liu, C. C. An orthogonal DNA replication system in yeast. *Nat. Chem. Biol.* **10**, 175–177 (2014).

137. Arzumanyan, G. A., Gabriel, K. N., Ravikumar, A., Javanpour, A. A. & Liu, C. C. Mutually Orthogonal DNA Replication Systems In Vivo. *ACS Synth. Biol.* **7**, 1722–1729 (2018).

138. Plotkin, J. B. & Kudla, G. Synonymous but not the same: the causes and consequences of codon bias. *Nat. Rev. Genet.* **12**, 32–42 (2011).

139. Kudla, G., Murray, A. W., Tollervey, D. & Plotkin, J. B. Coding-Sequence Determinants of Gene Expression in Escherichia coli. *Science.* **324**, 255–258 (2009).

140. Bentele, K., Saffert, P., Rauscher, R., Ignatova, Z. & Blüthgen, N. Efficient translation initiation dictates codon usage at gene start. *Mol. Syst. Biol.* **9**, 675 (2013).

141. Bhattacharyya, S. *et al.* Accessibility of the Shine-Dalgarno Sequence Dictates N-Terminal Codon Bias in E. coli. *Mol. Cell* **70**, 894-905.e5 (2018).

142. Presnyak, V. *et al.* Codon Optimality Is a Major Determinant of mRNA Stability. *Cell* **160**, 1111–1124 (2015).

143. Radhakrishnan, A. *et al.* The DEAD-Box Protein Dhh1p Couples mRNA Decay and Translation by Monitoring Codon Optimality. *Cell* **167**, 122-132.e9 (2016).

144. Boël, G. *et al.* Codon influence on protein expression in E. coli correlates with mRNA levels. *Nature* **529**, 358–363 (2016).

145. Sørensen, M. A. & Pedersen, S. Absolute in vivo translation rates of individual codons in Escherichia coli: The two glutamic acid codons GAA and GAG are translated with a threefold difference in rate. *J. Mol. Biol.* **222**, 265–280 (1991).

146. Zhang, G., Hubalewska, M. & Ignatova, Z. Transient ribosomal attenuation

coordinates protein synthesis and co-translational folding. *Nat. Struct. Mol. Biol.* **16**, 274–280 (2009).

147. Walsh, I. M., Bowman, M. A., Soto Santarriaga, I. F., Rodriguez, A. & Clark, P. L. Synonymous codon substitutions perturb cotranslational protein folding in vivo and impair cell fitness. *Proc. Natl. Acad. Sci.* **117**, 3528–3534 (2020).

148. Ikemura, T. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: A proposal for a synonymous codon choice that is optimal for the E. coli translational system. *J. Mol. Biol.* **151**, 389–409 (1981).

149. Sharp, P. M. & Li, W.-H. The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* **15**, 1281–1295 (1987).

150. Welch, M. *et al.* Design parameters to control synthetic gene expression in Escherichia coli. *PLoS One* **4**, e7002–e7002 (2009).

151. Spencer, P. S., Siller, E., Anderson, J. F. & Barral, J. M. Silent Substitutions Predictably Alter Translation Elongation Rates and Protein Folding Efficiencies. *J. Mol. Biol.* **422**, 328–335 (2012).

152. Dana, A. & Tuller, T. The effect of tRNA levels on decoding times of mRNA codons. *Nucleic Acids Res.* **42**, 9171–9181 (2014).

153. Eyre-Walker, A. & Bulmer, M. Reduced synonymous substitution rate at the start of enterobacterial genes. *Nucleic Acids Res.* **21**, 4599–4603 (1993).

154. Tuller, T. *et al.* An Evolutionarily Conserved Mechanism for Controlling the Efficiency of Protein Translation. *Cell* **141**, 344–354 (2010).

155. Goodman, D. B., Church, G. M. & Kosuri, S. Causes and Effects of N-Terminal Codon Bias in Bacterial Genes. *Science.* **342**, 475–479 (2013).

156. Cambray, G., Guimaraes, J. C. & Arkin, A. P. Evaluation of 244,000 synthetic sequences reveals design principles to optimize translation in Escherichia coli. *Nat. Biotechnol.* **36**, 1005–1015 (2018).

157. Rouskin, S., Zubradt, M., Washietl, S., Kellis, M. & Weissman, J. S. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo . *Nature* **505**, 701–705 (2014).

158. Frumkin, I. *et al.* Gene Architectures that Minimize Cost of Gene Expression. *Mol. Cell* **65**, 142–153 (2017).

159. Mittal, P., Brindle, J., Stephen, J., Plotkin, J. B. & Kudla, G. Codon usage influences fitness through RNA toxicity. *Proc. Natl. Acad. Sci.* **115**, 8639–8644 (2018).

160. Michelson, A. M. & Todd, A. R. Nucleotides part XXXII. Synthesis of a dithymidine dinucleotide containing a 3′: 5′-internucleotidic linkage. *J. Chem. Soc.* 2632–2638 (1955) doi:10.1039/JR9550002632.

161. Agarwal, K. L. *et al.* Total Synthesis of the Gene for an Alanine Transfer Ribonucleic Acid from Yeast. *Nature* **227**, 27–34 (1970).

162. Itakura, K. *et al.* Expression in Escherichia coli of a chemically synthesized gene for the hormone somatostatin. *Science.* **198**, 1056 LP – 1063 (1977).

163. Beaucage, S. L. & Caruthers, M. H. Deoxynucleoside phosphoramidites—A new class of key intermediates for deoxypolynucleotide synthesis. *Tetrahedron Lett.* **22**, 1859–1862 (1981).

164. Mandecki, W., Hayden, M. A., Shallcross, M. A. & Stotland, E. A totally synthetic plasmid for general cloning, gene expression and mutagenesis in Escherichia coli. *Gene* **94**, 103–107 (1990).

165. Pan, W. *et al.* Vaccine candidate MSP-1 from Plasmodium falciparum: A redesigned 4917 bp polynucleotide enables synthesis and isolation of full-length protein from Escherichia coli and mammalian cells. *Nucleic Acids Res.* **27**, 1094–1103 (1999).

166. Blight, K. J., Kolykhalov, A. A. & Rice, C. M. Efficient Initiation of HCV RNA Replication in Cell Culture. *Science.* **290**, 1972–1974 (2000).

167. Cello, J., Paul, A. V & Wimmer, E. Chemical Synthesis of Poliovirus cDNA: Generation of Infectious Virus in the Absence of Natural Template. *Science.* **297**, 1016–1018 (2002).

168. Gibson, D. G. *et al.* Complete Chemical Synthesis, Assembly, and Cloning of a Mycoplasma genitalium Genome. *Science.* **319**, 1215–1220 (2008).

169. Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6**, 343–345 (2009).

170. Larionov, V. *et al.* Specific cloning of human DNA as yeast artificial chromosomes by transformation-associated recombination. *Proc. Natl. Acad. Sci.* **93**, 491–496 (1996).

171. Gibson, D. G. *et al.* Creation of a Bacterial Cell Controlled by a Chemically Synthesized Genome. *Science.* **329**, 52–56 (2010).

172. Lartigue, C. *et al.* Genome Transplantation in Bacteria: Changing One Species to Another. *Science.* **317**, 632–638 (2007).

173. Lartigue, C. *et al.* Creating Bacterial Strains from Genomes That Have Been Cloned and Engineered in Yeast. *Science.* **325**, 1693–1696 (2009).

174. Dymond, J. S. *et al.* Synthetic chromosome arms function in yeast and generate phenotypic diversity by design. *Nature* **477**, 471–476 (2011).

175. Richardson, S. M. *et al.* Design of a synthetic yeast genome. *Science.* **355**, 1040–1044 (2017).

176. Mitchell, L. A. *et al.* Synthesis, debugging, and effects of synthetic chromosome consolidation: synVI and beyond. *Science.* **355**, eaaf4831 (2017).

177. Shen, Y. *et al.* Deep functional analysis of synII, a 770-kilobase synthetic yeast chromosome. *Science.* **355**, eaaf4791 (2017).

178. Zhang, W. *et al.* Engineering the ribosomal DNA in a megabase synthetic chromosome. *Science.* **355**, eaaf3981 (2017).

179. Mercy, G. *et al.* 3D organization of synthetic and scrambled chromosomes. *Science.* **355**, eaaf4597 (2017).

180. Wu, Y. *et al.* Bug mapping and fitness testing of chemically synthesized chromosome X. *Science.* **355**, eaaf4706 (2017).

181. Itaya, M., Tsuge, K., Koizumi, M. & Fujita, K. Combining two genomes in one cell: Stable cloning of the Synechocystis PCC6803 genome in the Bacillus subtilis 168 genome. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15971–15976 (2005).

182. Kouprina, N. & Larionov, V. Transformation-associated recombination (TAR) cloning for genomics studies and synthetic biology. *Chromosoma* **125**, 621–632 (2016).

183. Datsenko, K. A. & Wanner, B. L. One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. *Proc. Natl. Acad. Sci.* **97**, 6640–6645 (2000).

184. Murphy, K. C. Use of Bacteriophage λ Recombination Functions To Promote Gene Replacement in Escherichia coli. *J. Bacteriol.* **180**, 2063–2071 (1998).

185. Wang, K. *et al.* Defining synonymous codon compression schemes by genome recoding. *Nature* **539**, 59–64 (2016).

186. Lau, Y. H. *et al.* Large-scale recoding of a bacterial genome by iterative recombineering of synthetic DNA. *Nucleic Acids Res.* **45**, 6971–6980 (2017).

187. Hutchison, C. A. *et al.* Design and synthesis of a minimal bacterial genome. *Science.* **351**, aad6253 (2016).

188. Ostrov, N. *et al.* Design, synthesis, and testing toward a 57-codon genome. *Science.* **353**, 819–822 (2016).

189. Venetz, J. E. *et al.* Chemical synthesis rewriting of a bacterial genome to achieve design flexibility and biological functionality. *Proc. Natl. Acad. Sci.* **116**, 8070–8079 (2019).

190. Calles, J., Justice, I., Brinkley, D., Garcia, A. & Endy, D. Fail-safe genetic codes designed to intrinsically contain engineered organisms. *Nucleic Acids Res.* **47**, 10439–10451 (2019).

191. Schmidt, M. & de Lorenzo, V. Synthetic constructs in/for the environment: Managing the interplay between natural and engineered Biology. *FEBS Lett.* **586**, 2199–2206 (2012).

192. Torres, L., Krüger, A., Csibra, E., Gianni, E. & Pinheiro, V. B. Synthetic biology approaches to biological containment: pre-emptively tackling potential risks. *Essays Biochem.* **60**, 393–410 (2016).

193. Sturino, J. M. & Klaenhammer, T. R. Engineered bacteriophage-defence systems in bioprocessing. *Nat. Rev. Microbiol.* **4**, 395–404 (2006).

194. Pósfai, G. *et al.* Emergent Properties of Reduced-Genome Escherichia coli. *Science.* **312**, 1044–1046 (2006).

195. Hirokawa, Y. *et al.* Genetic manipulations restored the growth fitness of

reduced-genome Escherichia coli. *J. Biosci. Bioeng.* **116**, 52–58 (2013).

196. Albert, H., Dale, E. C., Lee, E. & Ow, D. W. Site-specific integration of DNA into wild-type and mutant lox sites placed in the plant genome. *Plant J.* **7**, 649–659 (1995).

197. Zhang, Z. & Lutz, B. Cre recombinase-mediated inversion using lox66 and lox71: method to introduce conditional point mutations into the CREB-binding protein. *Nucleic Acids Res.* **30**, e90–e90 (2002).

198. Burke, D. T., Carle, G. F. & Olson, M. V. Cloning of Large Segments of Exogenous DNA into Yeast by means of Artificial Chromosome Vectors. *Science.* **236**, 806–812 (1987).

199. Larin, Z., Monaco, A. P. & Lehrach, H. Yeast artificial chromosome libraries containing large inserts from mouse and human DNA. *Proc. Natl. Acad. Sci.* **88**, 4123–4127 (1991).

200. Larionov, V., Kouprina, N., Graves, J. & Resnick, M. A. Highly selective isolation of human DNAs from rodent–human hybrid cells as circular yeast artificial chromosomes by transformation-associated recombination cloning. *Proc. Natl. Acad. Sci.* **93**, 13925 LP – 13930 (1996).

201. Larionov, V., Kouprina, N., Solomon, G., Barrett, J. C. & Resnick, M. A. Direct isolation of human BRCA2 gene by transformation-associated recombination in yeast. *Proc. Natl. Acad. Sci.* **94**, 7384 LP – 7387 (1997).

202. Kouprina, N. *et al.* Functional copies of a human gene can be directly isolated by transformation-associated recombination cloning with a small 3' end target sequence. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 4469–4474 (1998).

203. Yamanaka, K. *et al.* Direct cloning and refactoring of a silent lipopeptide biosynthetic gene cluster yields the antibiotic taromycin A. *Proc. Natl. Acad. Sci.* **111**, 1957–1962 (2014).

204. Li, Y. *et al.* Directed natural product biosynthesis gene cluster capture and expression in the model bacterium Bacillus subtilis. *Sci. Rep.* **5**, 9383 (2015).

205. Kouprina, N. *et al.* Segments missing from the draft human genome sequence can be isolated by transformation-associated recombination cloning in yeast. *EMBO Rep.* **4**, 257–262 (2003).

206. Grimwood, J. *et al.* The DNA sequence and biology of human chromosome 19. *Nature* **428**, 529–535 (2004).

207. Leem, S. H. *et al.* Closing the gaps on human chromosome 19 revealed genes with a high density of repetitive tandemly arrayed elements. *Genome Res.* **14**, 239–246 (2004).

208. Kuspa, A., Vollrath, D., Cheng, Y. & Kaiser, D. Physical mapping of the Myxococcus xanthus genome by random cloning in yeast artificial chromosomes. *Proc. Natl. Acad. Sci.* **86**, 8917–8921 (1989).

209. Gibson, D. G. *et al.* One-step assembly in yeast of 25 overlapping DNA fragments to form a complete synthetic Mycoplasma genitalium genome. *Proc. Natl. Acad.*

Sci. pnas.0811011106 (2008) doi:10.1073/pnas.0811011106.

210. Heuer, T., Bürger, C., Maaß, G. & Tümmler, B. Cloning of prokaryotic genomes in yeast artificial chromosomes: Application to the population genetics of Pseudomonas aeruginosa. *Electrophoresis* **19**, 486–494 (1998).

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

212. Karas, B. J., Tagwerker, C., Yonemoto, I. T., Hutchison, C. A. & Smith, H. O. Cloning the Acholeplasma laidlawii PG-8A Genome in Saccharomyces cerevisiae as a Yeast Centromeric Plasmid. *ACS Synth. Biol.* **1**, 22–28 (2012).

213. Lee, N. C. O., Larionov, V. & Kouprina, N. Highly efficient CRISPR/Cas9-mediated TAR cloning of genes and chromosomal loci from complex genomes in yeast. *Nucleic Acids Res.* **43**, e55 (2015).

214. Schwartz, D. C. & Cantor, C. R. Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis. *Cell* **37**, 67–75 (1984).

215. Jiang, W. *et al.* Cas9-Assisted Targeting of CHromosome segments CATCH enables one-step targeted cloning of large gene clusters. *Nat. Commun.* **6**, 8101 (2015).

216. Kouprina, N., Noskov, V. N. & Larionov, V. Selective Isolation of Large Chromosomal Regions by Transformation-Associated Recombination Cloning for Structural and Functional Analysis of Mammalian Genomes. *YAC Protocols* (Humana Press, 2006). doi:10.1385/1-59745-158-4:85.

217. Shizuya, H. *et al.* Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in Escherichia coli using an F-factor-based vector. *Proc. Natl. Acad. Sci.* **89**, 8794–8797 (1992).

218. Sheng, Y., Mancino, V. & Birren, B. Transformation of Escherichia coli with large DNA molecules by electroporation. *Nucleic Acids Res.* **23**, 1990–1996 (1995).

219. Tao, Q. & Zhang, H.-B. Cloning and stable maintenance of DNA fragments over 300 kb in Escherichia coli with conventional plasmid-based vectors. *Nucleic Acids Res.* **26**, 4901–4909 (1998).

220. Deshpande, A. M. & Newlon, C. S. The ARS consensus sequence is required for chromosomal origin function in Saccharomyces cerevisiae. *Mol. Cell. Biol.* **12**, 4305–4313 (1992).

221. Nieduszynski, C. A., Knox, Y. & Donaldson, A. D. Genome-wide identification of replication origins in yeast by comparative genomics. *Genes Dev.* **20**, 1874–1879 (2006).

222. Wyrick, J. J. *et al.* Genome-Wide Distribution of ORC and MCM Proteins in S. cerevisiae: High-Resolution Mapping of Replication Origins. *Science.* **294**, 2357–2360 (2001).

223. Dershowitz, A. & Newlon, C. S. The effect on chromosome stability of deleting replication origins. *Mol. Cell. Biol.* **13**, 391–398 (1993).

224. Stinchcomb, D. T., Thomas, M., Kelly, J., Selker, E. & Davis, R. W. Eukaryotic DNA

segments capable of autonomous replication in yeast. *Proc. Natl. Acad. Sci. U. S. A.* **77**, 4559–4563 (1980).

225. Hill, C. W. & Harnish, B. W. Inversions between ribosomal RNA genes of Escherichia coli. *Proc. Natl. Acad. Sci. U. S. A.* **78**, 7069–7072 (1981).

226. Kothapalli, S. *et al.* Diversity of genome structure in Salmonella enterica serovar Typhi populations. *J. Bacteriol.* **187**, 2638–2650 (2005).

227. Hanahan, D., Jessee, J. & Bloom, F. R. B. T.-M. in E. Plasmid transformation of Escherichia coli and other bacteria. in *Bacterial Genetic Systems* vol. 204 63–113 (Academic Press, 1991).

228. de Lorenzo, V., Marlière, P. & Solé, R. Bioremediation at a global scale: from the test tube to planet Earth. *Microb. Biotechnol.* (2016) doi:10.1111/1751-7915.12399.

229. Ravikumar, A. & Liu, C. C. Biocontainment through Reengineered Genetic Codes. *ChemBioChem* **16**, 1149–1151 (2015).

230. Reis, M. dos, Savva, R. & Wernisch, L. Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res.* **32**, 5036–5044 (2004).

231. Pundir, S., Martin, M. J. & O'Donovan, C. UniProt Protein Knowledgebase BT - Protein Bioinformatics: From Protein Modifications and Networks to Proteomics. in *Protein Bioinformatics* (eds. Wu, C. H., Arighi, C. N. & Ross, K. E.) 41–55 (Springer New York, 2017). doi:10.1007/978-1-4939-6783-4_2.

232. Fellner, L. *et al.* Phenotype of htgA (mbiA), a recently evolved orphan gene of Escherichia coli and Shigella, completely overlapping in antisense to yaaW. *FEMS Microbiol. Lett.* **350**, 57–64 (2014).

233. Baba, T. *et al.* Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol. Syst. Biol.* **2**, 2006.0008 (2006).

234. Goodall, E. C. A. *et al.* The Essential Genome of Escherichia coli K-12. *MBio* **9**, e02096-17 (2018).

235. Claverie-Martin, F., Diaz-Torresgli, M. R., Yancey, S. D. & Kushner, S. R. Analysis of the altered mRNA stability (ams) gene from Escherichia coli. *J. Biol. Chem.* **266**, 2843–2851 (1991).

236. Ow, M. C. *et al.* RNase E levels in Escherichia coli are controlled by a complex regulatory system that involves transcription of the rne gene from three promoters. *Mol. Microbiol.* **43**, 159–171 (2002).

237. Jain, C. & Belasco, J. G. RNase E autoregulates its synthesis by controlling the degradation rate of its own mRNA in Escherichia coli: Unusual sensitivity of the rne transcript to RNase E activity. *Genes Dev.* **9**, 84–96 (1995).

238. Diwa, A., Bricker, A. L., Jain, C. & Belasco, J. G. An evolutionarily conserved RNA stem–loop functions as a sensor that directs feedback regulation of RNase E gene expression. *Genes Dev.* **14**, 1249–1260 (2000).

239. Schuck, A., Diwa, A. & Belasco, J. G. RNase E autoregulates its synthesis in

Escherichia coli by binding directly to a stem-loop in the rne 5′ untranslated region. *Mol. Microbiol.* **72**, 470–478 (2009).

240. Sakai, Y., Miyauchi, K., Kimura, S. & Suzuki, T. Biogenesis and growth phase-dependent alteration of 5-methoxycarbonylmethoxyuridine in tRNA anticodons. *Nucleic Acids Res.* **44**, 509–523 (2015).

241. Sokabe, M. *et al.* The structure of alanyl-tRNA synthetase with editing domain. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 11028–11033 (2009).

242. Young, T. S., Ahmad, I., Yin, J. A. & Schultz, P. G. An Enhanced System for Unnatural Amino Acid Mutagenesis in E. coli. *J. Mol. Biol.* **395**, 361–374 (2010).

243. Brustad, E. M., Lemke, E. A., Schultz, P. G. & Deniz, A. A. A General and Efficient Method for the Site-Specific Dual-Labeling of Proteins for Single Molecule Fluorescence Resonance Energy Transfer. *J. Am. Chem. Soc.* **130**, 17664–17665 (2008).

244. Schinn, S.-M. *et al.* Rapid in vitro screening for the location-dependent effects of unnatural amino acids on protein expression and activity. *Biotechnol. Bioeng.* **114**, 2412–2417 (2017).

245. Carlsson, A.-C. C. *et al.* Increasing Enzyme Stability and Activity through Hydrogen Bond-Enhanced Halogen Bonds. *Biochemistry* **57**, 4135–4147 (2018).

246. Kamio, Y., Lin, C. K., Regue, M. & Wu, H. C. Characterization of the ileS-lsp operon in Escherichia coli. Identification of an open reading frame upstream of the ileS gene and potential promoter(s) for the ileS-lsp operon. *J. Biol. Chem.* **260**, 5616–5620 (1985).

247. Burt, D. W. Origin and evolution of avian microchromosomes. *Cytogenet. Genome Res.* **96**, 97–112 (2002).

248. Giannuzzi, G. *et al.* Hominoid fission of chromosome 14/15 and the role of segmental duplications. *Genome Res.* **23**, 1763–1773 (2013).

249. Cooper, V. S., Vohr, S. H., Wrocklage, S. C. & Hatcher, P. J. Why Genes Evolve Faster on Secondary Chromosomes in Bacteria. *PLOS Comput. Biol.* **6**, e1000732 (2010).

250. Escudero, J. A. & Mazel, D. Genomic Plasticity of Vibrio cholerae. *Int. Microbiol.* **20**, 138–148 (2017).

251. Luo, J., Sun, X., Cormack, B. P. & Boeke, J. D. Karyotype engineering by chromosome fusion leads to reproductive isolation in yeast. *Nature* **560**, 392–396 (2018).

252. Shao, Y. *et al.* Creating a functional single-chromosome yeast. *Nature* **560**, 331–335 (2018).

253. Surosky, R. T., Newlon, C. S. & Tye, B. K. The mitotic stability of deletion derivatives of chromosome III in yeast. *Proc. Natl. Acad. Sci.* **83**, 414–418 (1986).

254. Murray, A. W., Schultes, N. P. & Szostak, J. W. Chromosome length controls

mitotic chromosome segregation in yeast. *Cell* **45**, 529–536 (1986).

255. Ueda, Y. *et al.* Large-scale genome reorganization in Saccharomyces cerevisiae through combinatorial loss of mini-chromosomes. *J. Biosci. Bioeng.* **113**, 675–682 (2012).

256. Itaya, M. & Tanaka, T. Experimental surgery to create subgenomes of Bacillus subtilis 168. *Proc. Natl. Acad. Sci.* **94**, 5378–5382 (1997).

257. Ausiannikava, D. *et al.* Evolution of Genome Architecture in Archaea: Spontaneous Generation of a New Chromosome in Haloferax volcanii. *Mol. Biol. Evol.* **35**, 1855–1868 (2018).

258. Yamaichi, Y. & Niki, H. migS, a cis-acting site that affects bipolar positioning of oriC on the Escherichia coli chromosome. *EMBO J.* **23**, 221–233 (2004).

259. Liang, X., Baek, C.-H. & Katzen, F. Escherichia coli with Two Linear Chromosomes. *ACS Synth. Biol.* **2**, 734–740 (2013).

260. Deneke, J., Ziegelin, G., Lurz, R. & Lanka, E. The protelomerase of temperate Escherichia coli phage N15 has cleaving-joining activity. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 7721–7726 (2000).

261. Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Completing bacterial genome assemblies with multiplex MinION sequencing. *Microb. Genomics* **3**, (2017).

262. Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLOS Comput. Biol.* **13**, e1005595 (2017).

263. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).

264. Walker, B. J. *et al.* Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLoS One* **9**, e112963 (2014).

265. Wick, R. R., Schultz, M. B., Zobel, J. & Holt, K. E. Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics* **31**, 3350–3352 (2015).

266. Duggin, I., Wake, R., Bell, S. & Hill, T. The replication fork trap and termination of chromosome replication. *Mol. Microbiol.* **70**, 1323–1333 (2008).

267. Esnault, E., Valens, M., Espéli, O. & Boccard, F. Chromosome Structuring Limits Genome Plasticity in Escherichia coli. *PLOS Genet.* **3**, e226 (2007).

268. Niki, H., Yamaichi, Y. & Hiraga, S. Dynamic organization of chromosomal DNA in Escherichia coli . *Genes Dev.* **14**, 212–223 (2000).

269. Valens, M. *et al.* Macrodomain organization of the Escherichia coli chromosome. *EMBO J.* **23**, 4330–4341 (2004).

270. Galli, E. *et al.* Replication termination without a replication fork trap. *Sci. Rep.* **9**, 8315 (2019).

271. Wendel, B. M., Courcelle, C. T. & Courcelle, J. Completion of DNA replication in

Escherichia coli. *Proc. Natl. Acad. Sci.* **111**, 16454–16459 (2014).

272. Lesterlin, C., Barre, F.-X. & Cornet, F. Genetic recombination and the cell cycle: what we have learned from chromosome dimers. *Mol. Microbiol.* **54**, 1151–1160 (2004).

273. Pérals, K., Cornet, F., Merlet, Y., Delon, I. & Louarn, J.-M. Functional polarization of the Escherichia coli chromosome terminus: the dif site acts in chromosome dimer resolution only when located between long stretches of opposite polarity. *Mol. Microbiol.* **36**, 33–43 (2000).

274. Koch, B., Ma, X. & Løbner-Olesen, A. Replication of Vibrio cholerae Chromosome I in Escherichia coli: Dependence on Dam Methylation. *J. Bacteriol.* **192**, 3903–3914 (2010).

275. Messerschmidt, S. J., Kemter, F. S., Schindler, D. & Waldminghaus, T. Synthetic secondary chromosomes in Escherichia coli based on the replication origin of chromosome II in Vibrio cholerae. *Biotechnol. J.* **10**, 302–314 (2015).

276. Schmied, W. H. *et al.* Controlling orthogonal ribosome subunit interactions enables evolution of new function. *Nature* **564**, 444–448 (2018).

277. Kouprina, N., Noskov, V. N. & Larionov, V. *YAC Protocols*. (2006).

278. Wang, K., de la Torre, D., Robertson, W. E. & Chin, J. W. Programmed chromosome fission and fusion enable precise large-scale genome rearrangement and assembly. *Science.* **365**, 922–926 (2019).

279. Strand, T. A., Lale, R., Degnes, K. F., Lando, M. & Valla, S. A New and Improved Host-Independent Plasmid System for RK2-Based Conjugal Transfer. *PLoS One* **9**, e90372 (2014).

280. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

281. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

282. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–192 (2013).

283. Sharan, S. K., Thomason, L. C., Kuznetsov, S. G. & Court, D. L. Recombineering: a homologous recombination-based method of genetic engineering. *Nat. Protoc.* **4**, 206–223 (2009).

284. Fredens, J. *et al.* Total synthesis of Escherichia coli with a recoded genome. *Nature* **569**, 514–518 (2019).

285. Guo, D. *et al.* Online High-throughput Mutagenesis Designer Using Scoring Matrix of Sequence-specific Endonucleases. *J. Integr. Bioinform.* **12**, 35–48 (2015).

286. Shao, Y. *et al.* CRISPR / Cas-mediated genome editing in the rat via direct injection of one-cell embryos. *Nat. Protoc.* **9**, 2493–2512 (2014).

287. Nguyen, D. P., Elliott, T., Holt, M., Muir, T. W. & Chin, J. W. Genetically Encoded 1,2-Aminothiols Facilitate Rapid and Site-Specific Protein Labeling via a Bio-orthogonal Cyanobenzothiazole Condensation. *J. Am. Chem. Soc.* **133**, 11418–11421 (2011).

288. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

289. Choe, D. *et al.* Adaptive laboratory evolution of a genome-reduced Escherichia coli. *Nat. Commun.* **10**, 935 (2019).