microbial biotechnology

Open Access

LoxTnSeq: random transposon insertions combined with cre/lox recombination and counterselection to generate large random genome reductions

Daniel Shaw,[1] Samuel Miravet-Verde,[1] Carlos Piñero-Lambea,[1,†] Luis Serrano[1,2,3]** and Maria Lluch-Senar[1,5]*

[1] Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr. Aiguader 88, Barcelona, 08003, Spain.

[2] Universitat Pompeu Fabra (UPF), Barcelona, 08002, Spain.

[3] ICREA, Pg. Lluís Companys 23, Barcelona, 08010, Spain.

[5] Basic Sciences Department, Faculty of Medicine and Health Sciences, Universitat Internacional de Catalunya, Sant Cugat del Vallès, 08195, Spain.

Summary

The removal of unwanted genetic material is a key aspect in many synthetic biology efforts and often requires preliminary knowledge of which genomic regions are dispensable. Typically, these efforts are guided by transposon mutagenesis studies, coupled to deepsequencing (TnSeq) to identify insertion points and gene essentiality. However, epistatic interactions can cause unforeseen changes in essentiality after the deletion of a gene, leading to the redundancy of these essentiality maps. Here, we present LoxTnSeq, a new methodology to generate and catalogue libraries of genome reduction mutants. LoxTnSeq combines random integration of lox sites by transposon mutagenesis, and the generation of mutants via Cre recombinase, catalogued via deep sequencing. When LoxTnSeq was applied to the naturally genome reduced bacterium *Mycoplasma pneumoniae*, we obtained a mutant pool containing 285 unique deletions. These deletions spanned from > 50 bp to 28 Kb, which represents 21% of the total genome. LoxTnSeq also highlighted large regions of non-essential genes that could be removed simultaneously, and other non-essential regions that could not, providing a guide for future genome reductions.

Received 10 July, 2020; revised 4 November, 2020; accepted 4 November, 2020.

For correspondence. *E-mail maria.lluch@pulmobio.com (M.L). **E-mail luis.serrano@crg.eu (L.S.).
[†] Present address: Pulmobiotics ltd, Dr. Aiguader 88, Barcelona, 08003, Spain.

Microbial Biotechnology (2021) **14**(6), 2403–2419
doi:10.1111/1751-7915.13714

Introduction

One of the core principles behind synthetic biology is the rational engineering of an organism towards a specific application (D'Halluin and Ruiter, 2013; Esvelt and Wang, 2013; Mol *et al*., 2018). Genetic modifications that allow for the generation of new proteins are well documented and have myriad applications. Famous examples include modifying bacteria to allow us to study their functions in greater detail such as the production fluorescent proteins (Prasher *et al*., 1992) or genetically modified bacteria capable of producing molecules for human use such as insulin (Goeddel *et al*., 1979).

Live biotherapeutics have the potential to become key players in health care over the coming years. Engineered bacteria can be used as drug delivery systems, acting as a chassis from which therapeutic platforms can be plugged into to activate new functions (Ausländer *et al*., 2012; Chi *et al*., 2019; Claesen and Fischbach, 2015; Hörner *et al*., 2012; Vickers *et al*., 2010). A chassis will need to display growth characteristics and phenotypes that fulfil biosafety requirements, yet which may be foreign or counterproductive to cells within their original niche. For example, a limited ability to proliferate *in situ*, or to evade the immune system might be never selected for in the organism's natural niche, but are features that might be of interest for a chassis. To illicit this change in phenotype, allowing the creation of an optimal chassis for a specific application, large-scale changes in genotype will have to occur. This will include both the removal of unwanted or unnecessary genomic regions, and the addition of new genes and functions (Folcher and Fussenegger, 2012; Ruder *et al*., 2011; Sung *et al*., 2016).

The loss of unwanted genetic regions has been shown to have desirable outcomes in bacterial engineering, with large-scale genome reductions showing that the removal of superfluous genes can both increase production of a desired protein and develop beneficial characteristics for a cell (Sung *et al*., 2016; Chi *et al*., 2019; Vernyik *et al*., 2020). For example, sequential genome reduction has been used as an engineering tool in Bacillus subtilis, creating a mutant strain with a boosted capacity for production of heterologous proteins by systematically removing genes that hinder protein production, or divert energy and resources away from optimal protein production (Ara *et al*., 2007). Similarly, the removal of prophage elements in Pseudomonas putida resulted in a strain that demonstrated markedly higher tolerance to DNA damage (Martínez-García *et al*., 2015).

As more and more advanced engineering tools become available to researchers, the scope for genome engineering has similarly increased in scale (Cameron *et al*., 2014; Hsu *et al*., 2014; Annaluru *et al*., 2015). One of the grand long-term goals in systems and synthetic biology is the generation of a chassis with a minimal genome (Sung *et al*., 2016; Chi *et al*., 2019). While the definition of a true 'minimal cell' is almost impossible to define (Koonin, 2000; Choe *et al*., 2016; Glass *et al*., 2017), a general consensus has emerged around a cell with a reduced genome, capable of completing a specific task with as few superfluous functions as possible (Zhang, 2010; Juhas *et al*., 2011; Hutchison *et al*., 2016).

While naturally occurring 'minimal' bacteria do exist to some degree (Fraser *et al*., 1995; van Ham *et al*., 2003; Fadiel *et al*., 2007), with genomes as small as the one of Mycoplasma genitalium coding for only 470 genes (Himmelreich *et al*., 1997), the average gene complement for a bacterial cell is roughly 5000 proteins, though this can vary by two orders of magnitude between the extremes (Land *et al*., 2015). Comparative genomics studies, along with functional considerations give a hypothetical minimal genome consisting of 200-350 genes (Breuer et al., 2019; Dewall and Cheng, 2011; Gil *et al*., 2004; Glass *et al*., 2006; Koonin, 2000), so engineering a bacteria into a 'minimal chassis' requires large levels of genome reduction. On top of this, there is a further layer of small proteins and non-coding RNAs that are often overlooked, all of which have their own essentialities and interactions with the rest of the genome and cell (Lluch-Senar *et al*., 2015; Miravet-Verde *et al*., 2019).

Our understanding of epistatic networks, specifically clusters of genes that contribute directly or indirectly to a single phenotype (Phillips, 2008), and the complex web of interactions between gene circuits, is far from complete (Otwinowski *et al*., 2018; Sailer and Harms, 2017; Weinreich *et al*., 2013). This has big implications for large-scale genome reduction projects, as knowledge gleaned from previous studies identifying non-dispensable or essential (E) and dispensable or non-essential (NE) regions of DNA can become obsolete as soon as alterations are made to the genome. A good example of this can be found in the creation of the reduced-genome JCVI-Syn2.0 from Mycoplasma mycoides by the team at the J. Craig Venter Institute. After the creation of their landmark JCVI-Syn1.0 strain (Gibson *et al*., 2010), random transposon mutagenesis was performed on the organism to identify the NE genes. The genome was then divided into eight sections, and each section had their NE genes removed independently, while the other seven sections were kept intact. Despite removing only NE genes, only 1/8 configurations resulted in a viable cell (Hutchison *et al*., 2016).

A similar example of interdependence was seen when studying gene essentiality and metabolism in *M. pneumoniae* and *Mycoplasma agalactiae*. It was shown that in linear metabolic pathways producing an E metabolite, all genes were E. However, when the E metabolite could be produced by two pathways, often both genes were classified as fitness (F) genes (Montero-Blay *et al*., 2020). Deletion of one pathway will make the genes in the other pathway become E.

This is a key limitation in many genome reduction studies, as rationally designed genome reduction experiments often rely on data generated before genetic manipulations are introduced. As such, any *a priori* assumptions about gene essentiality can be potentially redundant after any reductions have occurred. Screening by transposon mutagenesis after every cycle of reductions can be applied (Hutchison et al., 2016) but such screening efforts are both time- and labour-intensive, and rely on the basis that genes act alone, and not as a part of a larger network. However, the mutation or loss of certain genes has been shown to modulate the essentiality of others, known as 'bypass of essentiality' (LI, 2020). In the yeast Schizosaccharomyces pombe, 27% of E genes on chromosome II-L could be rendered NE in response to a different gene within the genome being deleted, mutated or overexpressed (Li *et al*., 2019).

This in turn demonstrates that reducing genomes based on *a priori* assumptions may not lead to the most optimized genome for the desired trait, as our knowledge on both how gene networks interact with each other on an epistatic level and which genes are truly E or NE under any given configuration is still far from complete. On top of this, the essentiality of a gene is highly dependent on the environment the bacterium inhabits (Sassetti *et al*., 2001; Bloodworth *et al*., 2013), and a change in metabolic function may lead to a change in secreted or imported by-products. This in turn could have knock-on effects on a cell's micro-environment, causing certain genes to change their essentiality in unpredictable ways.

Previous studies have demonstrated that random deletions are feasible in bacteria as a genome streamlining methodology, notably in Pseudomonas putida (Leprince *et al*., 2012) and Escherichia coli (Vernyik *et al*., 2020). However, both have been limited in either the ability to be scaled into a high-throughput screen or by the number of successful deletions reported.

Therefore, here we propose a new methodology, LoxTnSeq, to study the concept of genome reduction in an unbiased manner. Deletions are performed using the cre/lox system, consisting of two 34 bp lox sites which are acted upon by the Cre recombinase. If both lox sites are in a *cis* orientation, the DNA between them is cirularized and removed from the genome (Ghosh and Van Duyne, 2002). We utilize mutant lox sites, specifically lox66 and lox71 (Albert *et al*., 1995), delivered via random transposon mutagenesis to remove large segments of genomic material via the action of the Cre recombinase. As the Cre can recognize and act upon these lox sites, we have designated them as 'active' here. To affect a deletion, the Cre binds to the two active lox sites and recombines them, excising any DNA within them from the genome. As a result, a double mutant lox72 site is generated as a result of this recombination. This sequence is no longer recognized by the Cre (Berzin *et al*., 2012; Van Duyne, 2015) and is therefore 'inactive'. By combining a randomized genome deletion protocol with DNA deep sequencing, we can identify the large putative reductions that are possible within a genome. This methodology allowed us to delete large sections of DNA without biasing the results by any potentially misguided *a priori* assumptions.

We have applied the methodology in the naturally genome reduced bacterium *M. pneumoniae*, considered a model for a minimal cell. However, this methodology could be applied to different bacterial systems in order to obtain different bacterial chassis and thus to develop different applications in the synthetic biology field.

Results

Obtaining a high resolution library of lox mutants

Three different vectors (pMTnLox66Cm, pMTnLox71Tc and pMTnCreGm) were obtained, as described in Methods, to generate different libraries of transposon mutants. The plasmids here are derivatives of the Mini-pMTn4001 (Pour-El *et al*., 2002). Each plasmid contains a Tn4001 transposase outside of the two inverted repeats flanking the cargo region, ensuring that the cargo is inserted without the transposon and is thus stable within the genome. The first library of transposon mutants was obtained by transforming the *M. pneumoniae* M129 strain with the pMTnLox66Cm vector. This vector inserted a lox66 site randomly in the genome. After selection with chloramphenicol, we obtained the first mutant pool.

The sequencing of the first round of transposition was analysed via HITS (high-throughput insertion tracking by deep sequencing) to properly assess the coverage of the first transformation. This methodology employs sequencing the DNA using a known oligo within the transposon that reads outwards, to identify where the transposon inserted within the genome (Gawronski *et al*., 2009). Insertions were mapped to M. pneumoniae (see Methods) revealing 355 319 unique insertions sites (Table S1), which across the 816 Kb genome leads to a genome coverage of 43.5%, increasing to 64.1% when considering only known NE (Lluch-Senar *et al*., 2015). This represented an insertion every ~3 bases, an insertion frequency similar to what has been described in the latest transposition experiments using the same transposase and strain (Miravet-Verde *et al*., 2020). Figure 1 shows the insertion pattern of the pMTnLox66Cm transposon, along with the second pMTnLox71Tc transposon.

Creation of a pool of genome reduced mutants

This first pool of mutants was then transformed with the pMTnLox71Tc vector, and genomic DNA was isolated from the pool of surviving cells. After applying HITS and the same bioinformatic analysis procedure to this library, the number of unique insertions for the second transposon was 187 814. This leads to a genome coverage of 23% (1 insertion every ~4 bases), increasing to 38.2% for known NE genes. When both transposon samples were combined, we found a total of 387 962 unique insertions, corresponding to a 47.5% of genome coverage. Out of the 355 319 unique insertions found in the first transformation with pMTnLox66Cm, 155 170 (43.6%) were also recovered in pMTnLox71Tc. This left 200 149 insertions (56.4%) found in the first sample but not detected in the second. On the other hand, sample pMTnLox71Tc presented 32 644 unique insertions (17.4%) not found within the first sample, with the remaining 155 170 insertions shared between both transformations. As a result, 82.6% of the insertions found in the second transformation were also found in the first.

In terms of read counts per insertion event, pMTnLox66Cm presented an average value of 250 reads per insertion increasing to 386 reads per insertion in pMTnLox71Tc. The locations and relative abundances for the two transformations are shown in Fig. 1 and Table S1.

We explored where the insertions were located at each transformation step by comparing the frequency of insertion for genome bins (Table S2) and between the samples (Table S3). Also, to detect any differential preference compared with a general transposon insertion protocol, we included the comparative the essentiality reference samples for *M.* pneumoniae (Lluch-Senar *et al*., 2015). We observed a correlation between transformation steps of $R^2_{\text{genes}} = 0.86$ for genes and

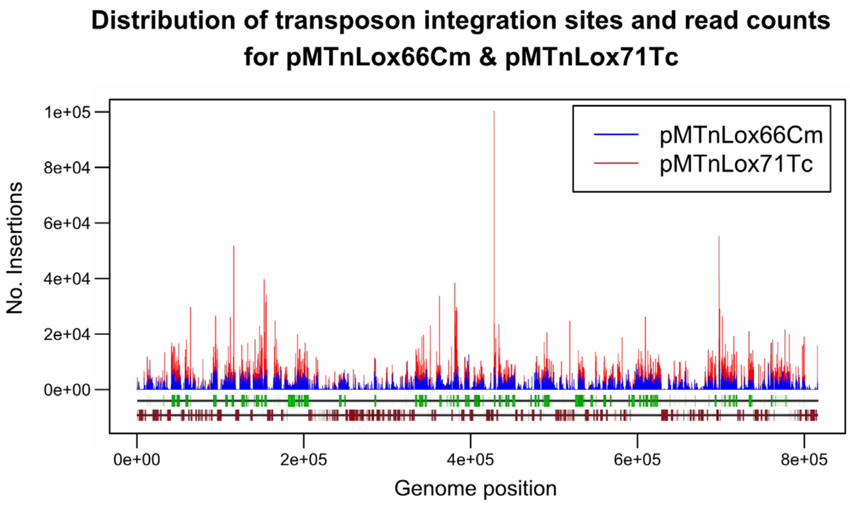

Fig. 1. Transposon insertion points in the *M. pneumoniae* M129 genome, transformed in series with 1 pMole of the pMTnLox66Cm vector and subsequently 1 pMole of the pMTnLox71Tc, corresponding to the position of known E (dark red boxes) and NE (green boxes) genes, as described by Lluch-Senar et al. (2015).

R^2_{bins} = 0.88 for bins. We also observed a better correlation between previous studies and pMTnLox66Cm (R^2_{genes} = 0.88; R^2_{bins} = 0.88) compared with pMTnLox71Tc (R^2 = 0.81; R^2_{bins} = 0.82).

We then explored the source of the differences found by calculating the fold change of the first transformation over the reference samples. We observed that bins with fold change > 2 (i.e. four times more insertions in pMTnLox66Cm compared with Lluch-Senar *et al.*, 2015 data set) were all in E regions, thus with initial low frequency, and the observed enrichment was not significant because the difference in number of insertions was never higher than 50 insertions every 1kb bin. On the other hand, bins with fold change < 2 (i.e. four times less insertions in pMTnLox66Cm) were all NE, and despite presenting less insertions when inserting lox sites, these bin regions were still NE in that condition (Table S2). Considering these observations, we could assume the transformation with lox sites produces libraries comparable to a transposon sequencing regular experiment. When looking at the gene level, we observed a general trend of genes having less insertions in the second transformation compared with the first (Table S4).

Both transposons utilize the same *Tn4001* transposase and were grown under the same conditions (with the exception of selective antibiotic, see Methods), so we must assume that the loss of unique insertions sites can be a result of epistatic interactions from the first round of transposon insertion. This could be the presence of synthetic lethality pairs, or a case of selective pressure towards insertions that have the least fitness cost, due to the already compromised state of the genome. Despite the fewer unique inserts in the second round, there was still a coverage of one insertion every four bases on average, indicating a very large pool of potential deletions. In this way, we generated a population carrying two lox sites, randomly distributed across the whole genome.

Importantly, transposon insertions take place randomly not only in terms of the chromosome position, but also with respect to their orientation. It is known that lox sites orientation is responsible from the action mode of Cre (Van Duyne, 2015), catalysing the excision of the DNA flanked by lox sites when they are in the same orientation, or the inversion of the region if lox sites are placed in opposite directions, as shown in Fig. 2. Therefore, a population of cells carrying pMTnLox66Cm and pMTnLox71Tc insertions would undergo either genome reduction or inversion depending on the particular clone analysed when subjected to the action of Cre recombinase.

Cells containing both lox sites were then transformed with the pMTnCreGm transposon, without lox sites and containing the Cre recombinase and a gene encoding gentamicin resistance. The Cre expression was controlled by the constitutive p438 promoter (Pich *et al.*, 2006; Montero-Blay *et al.*, 2019) and was introduced via transposon to ensure high levels of expression. In previous work, we demonstrated that the action of the Cre on a single active lox site (i.e. one that can be recognized and acted upon by the Cre, unlike lox72) in M. pneumoniae produces a lethal effect on par with a double-stranded break in the DNA (Shaw, 2019). This was corroborated here, as surviving colonies were isolated and grown in media containing either gentamicin, or chloramphenicol and tetracycline, to assay which cells contained

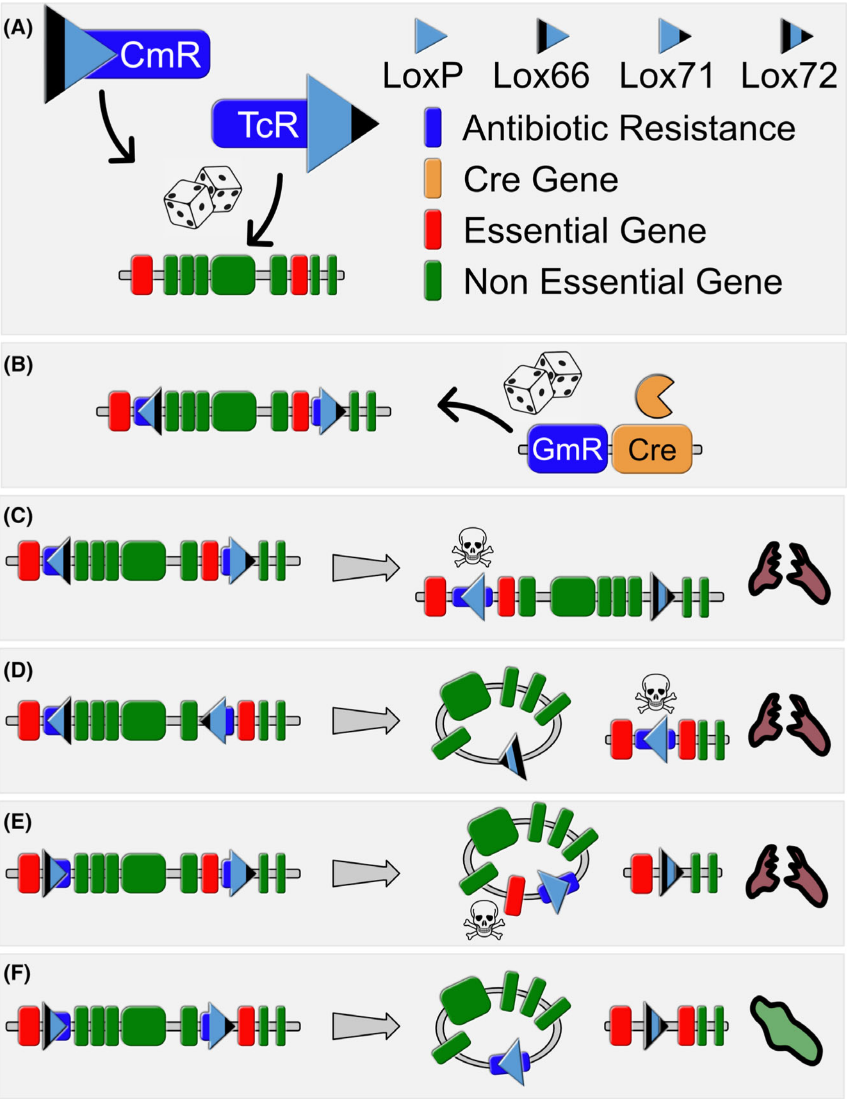

Fig. 2. Schematic protocol and interactions between mutant lox sites allowing for the selection of relevant deletion mutants.
A. Representation of the pMTnLox66Cm and pMTnLox71Tc transposons, along with dummy genomic DNA region and key. Both lox-containing transposons are integrated randomly into the genome.
B. After transformation with the two lox site containing transposons, the third transposon pMTnCreGm containing the Cre recombinase and gentamicin selective marker was added and randomly integrated into the genome.
C. Effect of the Cre on lox sites integrating in *trans* orientation. The genomic DNA between the two lox sites is inverted, leaving a lox72 and loxP site. The active loxP in the presence of constitutively expressed Cre is lethal in *M. pneumoniae*, and thus, the cells containing an inversion are killed, indicated via a lysed purple cell.
D. Effect of the Cre on lox sites integrating in *cis* orientation. Under the action of the Cre, the genomic DNA between them is circularized and removed. Here, an active loxP remains within the genome, initiating a lethal phenotype, indicated via a lysed purple cell.
E. Effect of the Cre on lox sites integrating in *cis* orientation, but encompassing an E gene. The desired inactive lox72 site is left within the genome, but the loss of an E gene causes cell death, indicated via a lysed purple cell.
F. Effect of the Cre on lox sites integrating in *cis* orientation, containing no E genes. An inactive lox72 is formed within the genome, impervious to the effects of the Cre, and the NE genomic regions are circularized and removed, with the green cell continuing to survive.

deletions and which cells contained inversions. 100/100 colonies picked grew in media containing gentamicin, indicating the Cre/gentamicin transposon was present, yet 0/100 colonies grew in media with chloramphenicol and tetracycline, indicating 100% of the colonies contained a reduced genome and could no longer express the resistance genes on the lox transposons and that the inversions were fully removed, as shown in Fig. 2.

Identification of random deletions

Genomic DNA from the gentamicin-resistant pool of cells was sequenced via the circularization protocol described in Methods. This methodology was chosen so that both sides of the deletion could be identified in a single read. Due to the random nature of the deletions, the only ‘known’ region of DNA to sequence from is the lox72 site, which is flanked by two ‘unknown’ genomic regions. Sequencing linear DNA would only reveal one side of the deletion per read, and it would then be impossible to pair the two reads from the same deletion, as all reads would start from the same point. By circularizing the DNA first, a single read can encompass both sides of the deletion. A total of 1 291 712 reads were recovered. Due to the random nature of the circularization protocol, not all reads could reach both inverted repeats of the transposons and give accurate insertion points for each transposon. Therefore, to allow for as accurate mapping as possible for those reads that do not include both insertion points, the genome was split into 50 bp bins, and reads were grouped into these bins. The reads were then filtered by putative deletions that contained a known E gene (1365 unique reads, 83% of all unique reads), and those that did not (285 unique reads, 17% of all unique reads), according to Lluch-senar et al’s previous classification (Lluch-Senar *et al*., 2015). We found a background of deletions affecting E genes with few reads (41 335 reads, 3% of total reads), which range in size from < 50 bp to spanning over half the entire genome (see Fig. 3), which probably are an artefact resulting from the circularization protocol for transposon sequencing (see Methods). Those deletions affecting genomic regions with only NE genes were less in number, spanned smaller regions of the genome, with the longest continuously NE region being ≈30 Kb. However, they also could also contain circularization artefacts. In total, they accounted for 1 250 337 reads (97% of total reads, see Fig. 3). When excluding all deletions that contained an E gene, 285 unique deletions were identified and mapped to the M129 genome (Table S5).

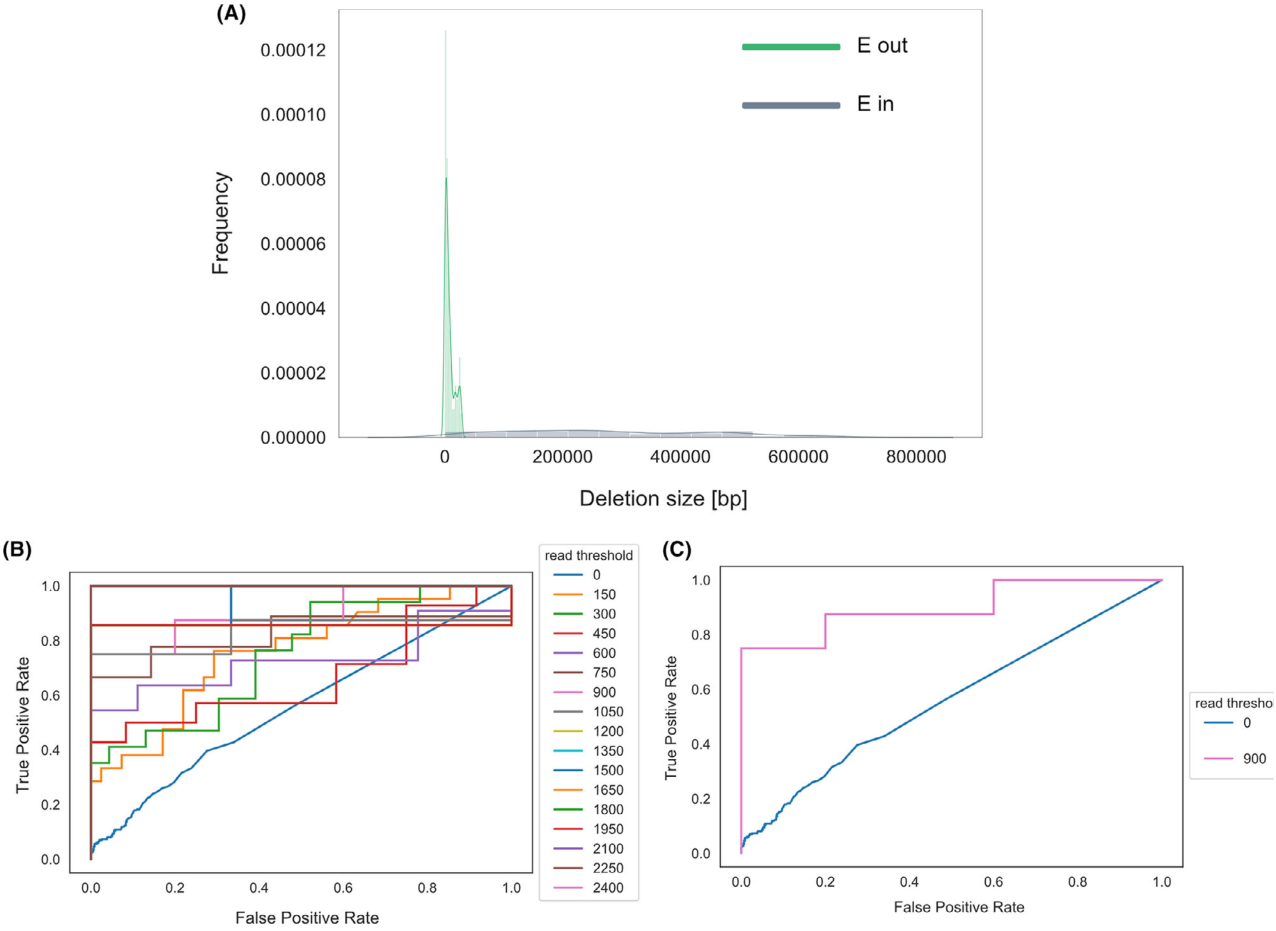

Fig. 3. Distribution of sequencing data.
A. Relative abundances of the deletions that contain an E gene (E in, blue) and those that do not (E out, green), based on the size of the deletion.
B. ROC curves assaying the false- and true-positive ratings of all deletions, depending on the number of reads.
C. ROC curve of our ‘gold standard’ delineation of 900 reads compared with the 0 read control.

Because the circularization protocol involved fragmentation and re-ligation of the DNA, the possibility of spurious ligations occurs. Therefore, we needed a way of discerning which reported deletions were true and which were false positives, created via random DNA ligation. Using the read length as the threshold parameter, we performed a receiver operating characteristic (ROC) curve approach to define high-confidence deletions (Fig. 3B). This methodology allows the definition of a read count threshold maximizing the true-positive rate (percentage of actual deletions properly detected) against the false-positive rate (percentage of artefactual deletions wrongly detected as positive). Due to the fact that most of the positive and negative deletions were represented by very few reads, we observed that no discrimination could be performed without previous filtering based on read count. Thus, we iterated the process along with different pre-filtering and defined as best conditions to filter deletions with read count > 900 that returned 8 deletions with a TPR > 75% and no false positives (Fig. 3C). This set was classified as the 'gold standard' of deletions (Table S5), as they had the highest level of reliability and were mapped to the M. pneumoniae genome.

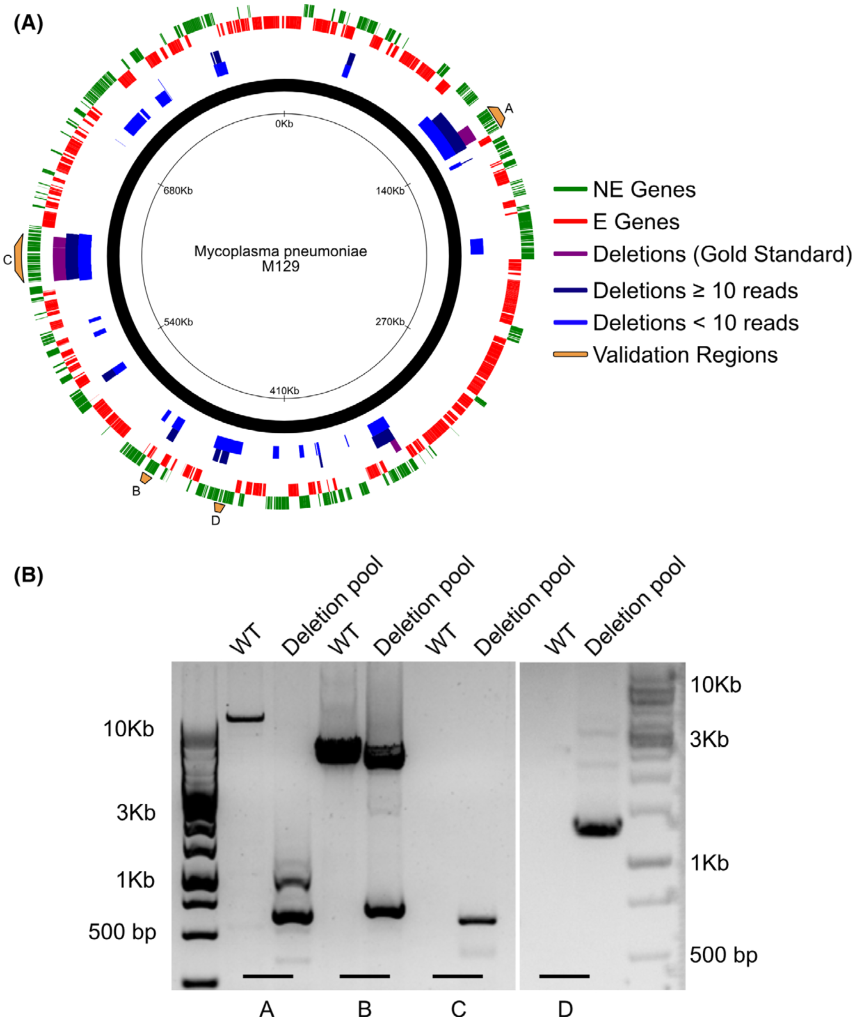

Fig. 4. (A) Circos plot showing the locations of the NE (green, lane 1), E genes (red, lane 2), deleted regions that were within the gold standard (purple, lane 3), deleted regions with greater than or equal to 10 reads (dark blue, lane 4) and deleted regions with less than 10 reads (light blue, lane 5) relative to the *M. pneumoniae* M129 genome. Yellow regions indicate the areas amplified via PCR for validations. Created using CiVi (Overmars *et al.*, 2015).

B. Validation of deletions via PCR. Gel electrophoresis of the four deletions candidates specified in the text above, with the WT on the left and deletions pool on the right in each condition. DNA primers were designed to amplify a 500–1500 bp region corresponding to the proposed deletion. Condition A amplified WT genomic DNA and genomic DNA isolated from the pool of deletion cells with oligos 11 and 12. Condition B amplified the same DNAs but using oligos 13 and 14, Condition C used the same DNA but with oligos 15 and 16, and condition D used the same DNAs but with oligos 17 and 18.

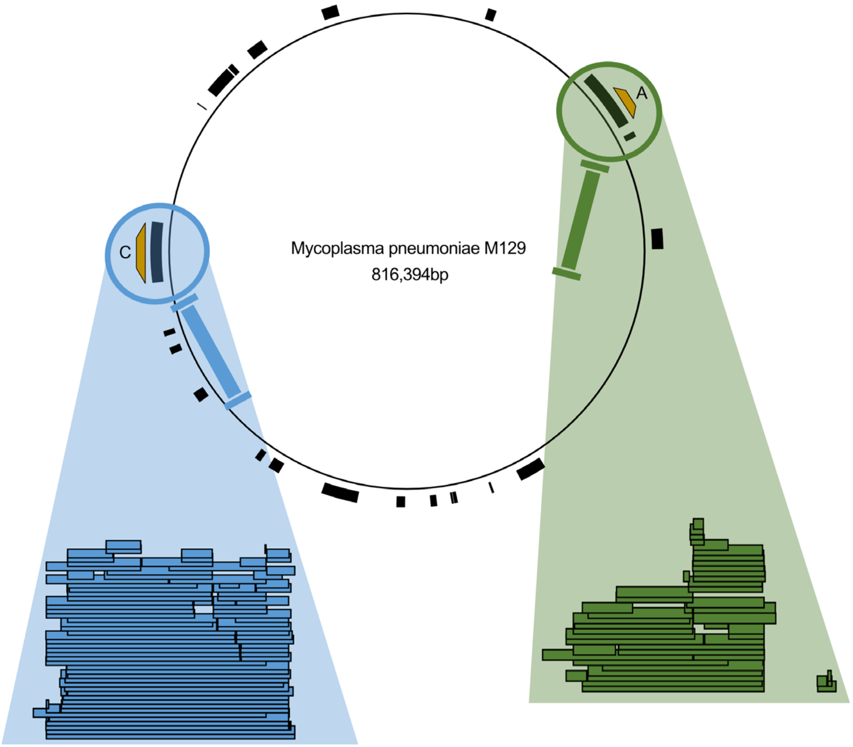

Fig. 5. Location of all putative deletions with the M129 genome, with deleted regions shown as black bars. Magnifying glasses show regions with multiple deletions clustering in single regions, with each coloured bar representing a unique deletion within the data set. Regions A and C from Figure 4A are indicated with the relevant yellow bars

We then split the remaining 285 deletions that did not contain an essential gene via read count, those deletions that had greater than 10 reads and those with less than 10 reads, which were also mapped to the *M. pneumoniae* genome, as shown in Fig. 4A.

While the number of deletions within the gold standard is low, representing just 8 deletions (see Table S5 and Fig. 4A), they are all deletions spanning multiple genes, all of which are clearly well tolerated by the cell as they were highly represented via read numbers within the data set. There are also many similarities between the sets, with similar deletions appearing in both highly and low read sets.

Despite the low number of 'gold standard' genes, we are confident that our sequencing method did not produce a large number of false positives. As the data in Fig. 3 show, the vast majority of the reads generated (97%) indicated deleted regions that did not contain an E gene, and the remaining 3% that did show an even distribution across the genome (see Fig. S1). Therefore, we feel confident applying the slightly less stringent test for fidelity and simply allowing all the deletions that do not contain the removal of a known E gene.

Of the regions deleted, the largest was 28.7 Kb, the smallest $<$ 50 bp. 147 genes were deleted across the pool, accounting for 171.2 Kb (21% of the genome), with the vast majority of the functions unknown. Of the genes deleted, only 29 (19.7%) had an ascribed name and function. 139 genes were annotated as NE, with the remaining 8 classified as F genes, those genes whose loss is not lethal, but imparts a severe growth defect, according to the most recent essentiality data (Lluch-Senar et al., 2015). The full list of all deletions can be found in Table S6, all deletions that did not contain an E gene in Table S5 and the genes deleted in Table S7. Of the 259 genes that are classified as NE in M. pneumoniae (Lluch-Senar et al., 2015), we deleted 56%. The mean size of depletions comprising those NE genes was 7750 bp and median of 4750 bp, indicating the majority of genes deleted were removed as part of a larger region, not as single knockouts.

As validations, 4 regions were chosen from the list of putative deletions, labelled A, B, C and D (see Fig. 4A, yellow bars), with various characteristics. Region A contained seven NE genes (*mpn096* to *mpn102*), ≈10 Kb in size, was the most highly represented deletion within the pool accounting for 964 628 reads and the only validation within the gold standard. Region B represented a smaller region, containing four NE genes (*mpn397* to *mpn400*), a size of ≈5 Kb and covered by 593 reads. Region C represented one of the largest available deletions, and contained 19 NE genes and 1 F gene (*mpn493* to *mpn512*), with a deleted area of ≈25 Kb

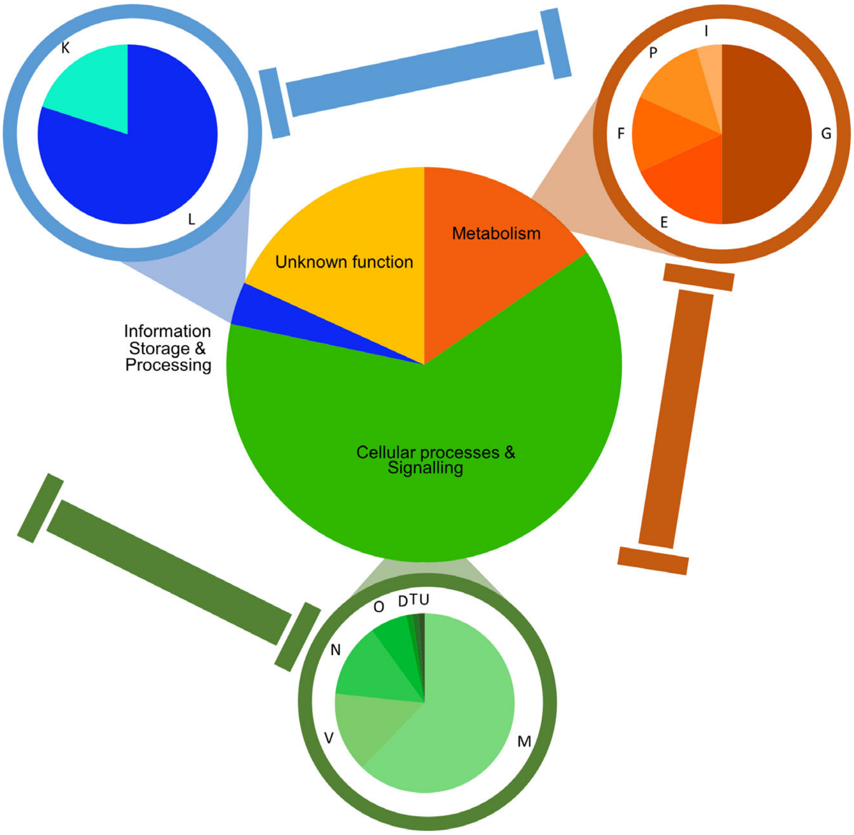

Fig. 6. Breakdown of deleted genes via COG categories. [G] Carbohydrate transport and metabolism. [E] Amino acid transport and metabolism. [F] Nucleotide transport and metabolism. [P] Inorganic ion transport and metabolism. [I] Lipid transport and metabolism. [M] Cell wall/membrane/envelope biogenesis. [V] Defence mechanisms. [N] Cell motility. [O] Post-translational modification, protein turnover and chaperones. [D] Cell cycle control, cell division and chromosome partitioning. [T] Signal transduction mechanisms. [U] Intracellular trafficking, secretion and vesicular transport. [L] Replication, recombination and repair. [K] Transcription.

covered by 193 reads. Region D contained 6 NE genes (*mpn368* to *mpn373*) within 7.9 Kbs, but was represented by just one read in the data set.

Figure 4B shows clearly that the putative deletion in all four regions were represented in the pool of deletions, but not found in the WT cells. There was no amplification in WT C due to its large size (25 Kb) There are also multiple deletions visible in the deletion amplifications in conditions A, B and D, which were expected from the sequencing data due to the multitude of different deletions in that area of the genome. The WT band is not visible in conditions C and D due to the PCR constraints in run length. The most prominent deletion band in the three conditions was isolated and sequenced, and each showed the requisite genomic regions intersected by the lox72 site within the transposon inverted repeats.

As the distribution of read counts was not uniform among the data set, nor was the distribution of deletions uniform across the genome. There were clear hotspots present where multiple different deletions were found across the population, as shown by Fig. 5.

The highlighted regions in Fig. 5 show two of the strongest deletion hotspots, located at 120 Kb and 610 Kb regions respectively. These regions show multiple overlapping unique deletions with slightly different lox insertions, across various sizes and configurations. This indicates that these regions are amenable to multiple different deletions, and there is little epistatic interaction between the genes present.

Contrary to this, Fig. 4A shows a large NE region centred on the 200 Kb region of the chromosome. This region contains 13 NE genes (*mpn141* to *mpn153*), yet only four unique deletions were found within it, all of which were variations of a deletion from *mpn146* to *mpn152*. Despite the majority of the genes within this region having no known function, E or otherwise, deletions within this region appear to be incompatible with cellular survival. The only known functions are linked to cytoadherence, specifically P1 adhesins (Nakane et al., 2011; Xiao et al., 2015). The canonical P1 adhesin (mpn141) was not deleted, nor were any of the other major adhesin genes hmw1 (mpn447), hmw2 (mpn310) and hmw3 (mpn452). However, 15/22 adherence proteins were deleted across the population, so why this section is more essential than the others is unclear (Fig. 6).

Looking at the functions of the deleted genes, they were grouped by COG category (Tatusov et al., 2000). The most commonly deleted functional category was M, genes involved in the composition and biogenesis of the cell membrane, accounting for 39% of deleted genes. Following this was the genes of unknown function (COG category S), with 18%. In total, deleted genes involved in metabolism (COG categories E, F, G, I and P) composed 15% of all deleted genes, information storage and processing genes (COG categories K and L) composed 3%, genes of unknown function 18% and genes involved cellular processes and signalling (COG categories D, M, N, O, T, U and V) 63% (see Table S8 for full breakdown).

Discussion

The cre/lox system has been used as a deletion system on many occasions, due to its ability to act in both prokaryotic and eukaryotic cells. In addition, the usage of mutant lox sites to facilitate deletions that result in an inactive lox has also been demonstrated in bacteria, such as being used to knock out single genes in series (Pan et al., 2011), or to knock out large but targeted genome region using either targetrons (Cerisy et al., 2019) or via recombineering (Xin et al., 2018).

A similar approach to this, using s transposon delivered FLP-FRT system instead of the cre/lox, was undertaken in Pseudomonas putida, creating large random deletions within the genome (Leprince et al., 2012). There, the authors succeeded in performing cyclical deletions within the genome; however, there were three main caveats that our methodology aims to improve upon. First, their system is not self-selective for deletions. Individual mutants need to be screened manually for a loss of the two antibiotic resistances found in the transposons which would indicate a deletion has taken place. This precludes its use as high-throughput assay for surveying all possible deletions. Secondly, the efficacy of transformation was low, with only 255 independent insertions recorded for the second transposon insertion. This drastically lowers the utility in finding all possible deletions. Finally, while their system does allow for multiple rounds of deletion, this is limited by the usage of the FLP-FRT system. After a successful deletion has occurred, the system still contains a FRT site within the genome. Any subsequent insertions of a new FRT will utilize this pre-existing genomic FRT to create the deletion. Therefore, all subsequent deletions are limited to the areas directly adjacent to the original. While our system is not currently optimized for multiple round of deletion, the creation of the lox72 site via our methodology gives the potential for new deletions to attempted across the entire genome, as the lox72 is not recognized by the Cre and thus will not interact with any subsequently added lox sites.

There have also been previous studies using random integration of lox sites to create deletions within a bacterial genome. Multiple random deletions were observed in *Corynebacterium glutamicum* using a similar method, with loxP sites contained within transposons and randomly integrated into the genome and activated via a Cre suicide vector, ranging from 400 bp to 158 Kb in size (Tsuge et al., 2007). However, the system was not self-selective, with the authors commenting that only 1.5% of final colonies contained deletions. The rest of the colonies contained inversions, and only 42 unique deletions were characterized. The authors also note that of the 42 deletion strains they recovered, only 2 had growth rates stronger than the WT strain, and many showed severe fitness defects. It is reasonable to assume that there may have been other deletion strains that could not compete against the large number of cells that had inversions, and therefore increased fitness due to no gene loss, and thus were lost from the pool.

By ensuring that our system leads to full selection against inversions between the lox sites, we can attempt to prevent this issue of competition against quasi-WT cells and thus potentially retain as many different deletions as possible. By looking at them, read numbers of each unique deletion, we can also get an estimate of how many cells were in the population, and thus an estimation of the growth rate compared with the other clones.

This study also highlighted the utility of using the Cre recombinase as its own self-selective marker. The inversion between a left-mutant and right-mutant lox sites necessitated the creation of a loxP site (Ghosh and Van Duyne, 2002; Van Duyne, 2015), and thus, we propose the action of the Cre alone was enough to cause this self-selection and self-retention on only the cells that contained a deletions resulting in an inactivated lox72. We have previously shown that a lethal phenotype is expressed in M. pneumoniae when the Cre recombinase acts upon a lone active lox site (Shaw, 2019), and this worked perfectly as a counter-selective method here. The one caveat to this method being that 50% of potential deletions were removed due to the formation of the loxP instead of the lox72. However, looking at the overall distribution of deletions in Fig. 4A, it is clear that the vast majority of putatively NE regions were deleted somewhere within our pool. Therefore, the high transposons density appears to be able to counter-balanced this loss in efficiency. A further improvement could be to use

PacBio or Oxford Nanopore deep-sequencing technologies to ensure we reach the inverted repeats of both transposons, and therefore, we could map the deletions to a single-base resolution.

Due to this high transposon density in the transformations, we achieved a very high coverage of deletions across the *M. pneumoniae* genome, resulting in the deletion of 21% of the genome across the pool, with deletions ranging from under 50 bp to 28 Kb. However, the distribution of these deletions is not uniform, with the six most represented deletions accounting for over 99% of the data set. Whether this is a true reflection of the ratio of deletion mutants, or an unforeseen by-product of the sequencing and analysis pipeline is unclear. Despite this, we were able to validate deletions with low read numbers. Regions B and C consisted of 0.047% and 0.015% of the total reads, respectively, and were easily identified, and region D was isolated via PCR and sequenced, despite accounting for only a single read. This indicates that the full list of 285 deletions is likely reliable. It also indicates that there was probably a bias within the library preparation for the sequencing protocol that artificially inflated the most common deletions, thus skewing the final deletion ratio.

The high levels of variation within the larger regions also indicate the pool is robust and contains multiple valid deletions, as it shows that the variation of insertion sites for the original lox site transposons is as high as we expected. The concentration of deletions in hotspots is not due to a bottleneck caused by poor transformation efficiency in either of the lox insertion stages, as the variation in the hotspots shows multiple integrations, with a vast range in the size of the deletions across the general region. If the hotspot was caused by the fact that only a small number of transposons were present in one of stages, the vast majority of the deletions would share a common end or starting point, which is not what we observe. Instead, we are probably seeing those regions whose loss imparts the lowest reduction in cellular fitness. Looking at the region between bases 529 000 and 630 000 (indicated by the blue bars in Fig. 4A), we see a very high density of deletions. This region is almost entirely populated by NE genes (24 NE genes, 3 F genes), of which 17 have no clearly defined function (see Table S7, *mpn490* to *mpn513*).

Due to the uncertainty inherent to the data generated from the sequencing protocol, i.e. the majority of reads not containing both inverted repeat regions, we decided that splitting the genome into 50 bp bins gave us specificity enough to map the deletions as accurately as possible. Due to this aggregation method however, there could be many more deletions that are similar to each other by fewer than 50 bp and thus are missed from the analysis by being grouped with the other reads. This could mean, however, that we are greatly underestimating the number of unique deletion events that occurred.

A major consideration with this protocol is its population-based, and thus competitive, nature. The large variation of deletions that are created are growing in direct competition with each other, and thus, this protocol also allows for us to partially select for those cells that are the most robust. While we only allowed for one passage for the cells to grow in an attempt to minimize this as much as possible, the selection for faster growing mutants is inevitable, and it remains an inherent property of bacterial life that the fast-growers will proliferate at the expense of the slow growers. One possibility to have a larger coverage could be to plate the cells in agar and once grown scrape all and sequence. This way since cells will grow as single colonies competition will be minimized.

By utilizing a randomized large-scale deletion protocol, not only can single genes be tested for their updated essentiality, but also whole regions as well. This has the power to greatly increase the scope of genome reduction projects, by quickly identifying the largest areas that are amenable to deletion at any given time and identifying those regions that may not be amenable to deletion despite being counterintuitive. This is shown well by the deletions we see across the population. While the distribution of E and NE genes is fairly even in the *M. pneumoniae* genome, there are some clear islands of non-essentiality. We have shown that large deletions are possible in many of these islands, the regions around the 610 and 140 Kb most notably. Looking at Fig. 4A, the majority of the regions with multiple NE genes contain some level of deletion. However, there are also places where few if any deletions are observed, such as the clusters of NE genes at 244, 370 and 490 Kb. While the loss of any of these genes individually is possible for the cell, their combined loss appears to be lethal. Therefore, this tool can be used not only to find which regions of the genome are most amenable to large-scale deletion, but also to identify which putative NE regions play host to the most epistatic interactions.

Furthermore, the competitive nature of the protocol allows the researcher to not only elucidate the larger regions that can be deleted, but also those with the least fitness expense in any given environment. Similar to the RANDEL protocol outlined by Vernyik et al, we can place emphasis on cell fitness and robustness of growth during selection (Vernyik et al., 2020). As LoxTnSeq does not rely on endogenous DNA repair mechanisms for its deletion however, it has the potential to be applicable in a wider range of organisms.

If a cell is being minimized for a specific application, then genome reduced pools can be grown in the desired condition for as long as required, and the cells with the most viable reductions will outcompete those whose deletions are less viable, and give researchers insights into not only which genes provide a desirable phenotype in new conditions, but also which operons and larger genome areas as well. We found limited examples of this in our own study. Our pool of deletions was grown for just a single passage under standard laboratory conditions, and thus, genes required for the *M. pneumoniae* cells to exhibit pathogenesis were not required. As such, 15 separate adhesion proteins (out of a total of 22) were among those deleted, as were nine restriction putative enzyme proteins. On top of this, the main virulence factor in M. pneumoniae, the CARDS toxin (Parrott et al., 2016; Waites and Talkington, 2004), was also among those genes deleted, along with many of the adhesins that are linked to pathogenesis (Parrott et al., 2016). This indicates that the protocol has the ability to be utilized as an attenuation process as well.

In line with this, the methodology also has potential for conversion into a multi-step protocol. The removal of the constitutive Cre transposon from the genome, or its replacement with a conditionally activated Cre could allow for multiple rounds of the technique to be utilized within a single cell. This could convert the technique from a screening tool to identify amenable deletions to a self-contained large-scale genome minimization technique, capable of deleting as many large and small genomic regions as the cell can endure. Coupled with its random nature, the technique could help researchers avoid unforeseen negative epistatic interactions and delete as much genetic material as is feasible within the cell, potentially paving the way for further elucidation of the minimal machinery needed for a cell to survive. While single rounds of the protocol may only remove relatively small amounts of DNA in this case (our largest deletion of 28 Kb accounts for 3% of the total *M. pneumoniae* genome), its unbiased and competitive nature make LoxTnSeq an attractive prospect.

In conclusion, we present the LoxTnSeq protocol as a multi-purpose tool for synthetic biology. It is capable of deleting large regions of NE genes from a host genome, identifying candidate regions for genome reductions based on the fitness, the deletion imparts on the cell, allowing for more accurate essentiality maps based on the loss of multiple genes. It also allows for the identification of NE regions that contain strong epistatic interactions that may cause a loss of viability if removed together, despite their constituent parts being deemed NE. We hope that these attributes will make it a useful contribution to the growing synthetic biology tool box.

Experimental procedures

Strains and culture conditions

Wild-type *Mycoplasma pneumoniae* strain M129 (ATTC 29342, subtype 1, broth passage no. 35) was used. Cells were cultured in 75cm^2 tissue culture flasks at 37°C in standard Hayflick media, as described by Hayflick (1965) and Yus *et al.,* (2009) (Hayflick, 1965; Yus et al., 2009), supplemented with 100 μg/ml ampicillin, 2 μg/ml tetracycline, 20 μg/ml chloramphenicol, 3.3 μg/ml puromycin and 200 μg/ml gentamicin as appropriate. Hayflick agar plates were created by supplementing the Hayflick with 1% Bacto Agar (BD, Cat. No. 214010) before autoclaving.

NEB 5-alpha Competent *E. coli* cells (New England Biolabs, Catalogue number C2987H) were used for plasmid amplification and cloning. They were grown at 37°C in Lysogeny Broth (LB) at 200RPM or static on LB agar plates, supplemented with 100 μg/ml ampicillin.

Plasmid DNA

All plasmids were generated using the Gibson isothermal assembly method (Gibson *et al.*, 2009). DNA was isolated from NEB 5-alpha Competent *E. coli* cells, and individual clones were selected by using LB + ampicillin plates (100 μg/ml). Correct ligation was confirmed via Sanger sequencing (Eurofins Genomics). A list of all plasmids and primers used in their generation and sequencing can be found in Table S9 and S10. The plasmids used in this study were generated as follows:

pMTnLox71Tc. The plasmid pMTnTetM438 was amplified separately via PCR using oligos 1 and 2, and 3 and 4. The samples were digested with DpnI and isolated via electrophoresis. Bands of approx. 4.2 Kb and 2 Kb, respectively, were isolated and annealed via Gibson assembly.

pMTnLox66Cm. The plasmid pMTnCat was amplified via PCR using oligos 5 and 6. The sample was digested with DpnI and isolated via electrophoresis. A band of approx. 5 Kb was isolated and self-annealed via Gibson assembly.

pMTnCreGm. The plasmid pMTnCat was amplified via PCR using oligos 7 and 8, and the plasmid pGmRCre was amplified via PCR using oligos 9 and 10. The samples were digested with DpnI and isolated via electrophoresis. Bands of approx. 4.2 Kb and 2.6 Kb, respectively, were isolated and annealed via Gibson assembly.

Random genome reduction protocol

WT M129 cells were transformed according to the protocol outlined by Hedreyda et al. (1993). Cells were grown to mid-log phase, identified by the Hayflick media changing from red to orange. The media were decanted, and the flask was washed 3x with 10 ml chilled electroporation buffer (EB: 8mM HEPES, 272nM sucrose, pH 7.4). Cells were scraped into 500 μl chilled EB and homogenized via 10x passages through a 25 gauge syringe needle. Aliquots of 50 μl of the homogenized cells were mixed with a pre-chilled 30 μl EB solution containing the 1 pMole of pMTnLox66Cm plasmid DNA. Samples were then kept on ice for 15 mins. Electroporation was done using a Bio-Rad Gene Pulser set to 1250 V, 25 μF and 100Ω. After electroporation, cells were incubated on ice for 15 mins, then recovered into a total of 500 μl Hayflick media and incubated at 37°C for 4 hours. 125 μl of transformed cells were then inoculated into T75 cm^2 culture flasks containing 20 ml Hayflick and supplemented with 20 μg/ml chloramphenicol.

The transformed cells were grown to mid-log phase and then isolated via pelleting. The described protocol was replicated in duplicate, one of the samples was sequenced to validate the transposon insertion using pMTnLox66Cm (R1B1), while the second was used in the subsequent transformations (R1B3). To ensure the recovery of planktonic cells, the media were transferred to a 50 ml Falcon tube. The flask was then scraped into 500 μl Hayflick, which was added to the Falcon tube with the media. The sample was centrifuged at 10 000 RPM at 4°C for 10 mins to pellet the cells. The supernatant was discarded and the cells resuspended in 500 μl chilled EB. The cells were homogenized via 10x passages through a 25-gauge syringe needle. Aliquots of 50 μl of the homogenized cells were mixed with a pre-chilled 30 μl EB solution containing the 1 pMole of pMTnLox71Tc plasmid DNA, and transformed using the previously described settings. Cells were recovered in the same manner, and 125μl was inoculated into a T75 cm^2 flask containing 20 ml Hayflick media supplemented with 2 μg/ml tetracycline and 20 μg/ml chloramphenicol.

The transformed cells were again grown to mid-log phase and isolated via the centrifugation method described above. Aliquots of 50 μl of the homogenized cells were mixed with a pre-chilled 30 μl EB solution containing the 1 pMole of pMTnCreGm plasmid DNA, and transformed using the previously described settings. Cells were recovered in the same manner, and 125 μl were inoculated into a T75 cm^2 flask containing 20 ml Hayflick media supplemented with 200 μg/ml gentamicin.

Quantification of transposon density

Cultures R1B1 (transformed only with pMTnLox66Cm) and R1B3 (R1B1 transformed with pMTnLox71Tc) were grown to mid-log phase; then, cells were isolated via the centrifugation protocol outlined above. Genomic DNA was isolated via the MasterPure™ genomic DNA purification kit and sent for sequencing using a standard 125 bp paired-end read library preparation protocol for an Illumina MiSeq. The raw data were submitted to the ArrayExpress database (http://www.ebi.ac.uk/arrayexpress) and assigned the accession identifier E-MTAB-9590. Insertion sites for the transposon were identified using oligo 19 (Table S10), which is bound to the 3' end of the chloramphenicol resistance genes, directly downstream of the inverted repeat (IR) sequence, and identified in the M. pneumoniae M129 genome using FASTQINS (Miravet-Verde et al., 2020). This bioinformatic tool allows to identify the point of insertion of the transposon used in the study and quantify, by read counts, how many times that specific insertion is detected by sequencing. This tool was run in both samples R1B1 and R1B3, corresponding to the population transformed with pMTnLox66Cm and then secondly transformed with pMTnLox71Tc respectively (Table S1). These results were later analysed using ANUBIS essentiality framework (Miravet-Verde et al., 2020) to explore the coverage (i.e. percentage of nucleotide bases found inserted for a considered set of positions), considering the genome of M. pneumoniae (816 394 bp) and for each set of known E and NE genes as described by Lluch Senar et al. (Lluch-Senar et al., 2015). For the analysis of the location of the insertions, we considered the transposon density within a gene (i.e. number of insertions normalized by the gene length, Table S4) and the genome of M. pneumoniae at 1kb resolution (bin size of 820bp, Table S2). Pearson's *R* correlations can be found in Table S3.

High-throughput deletion sequencing

From the original isolated cells that survived transformation with the pMTnCreGm, 100μl was isolated and had genetic material isolated via the MasterPure™ genomic DNA purification kit (Lucigen, Cat. No. MC85200). Genomic DNA was fragmented to 300 bp via Covaris sonication. 5' phosphorylation was undertaken to allow for adapter binding; then, 3' overhangs were filled to create blunt ends. These were then ligated using T4 ligase to create circular fragments, and linear DNA was removed via digest with exonuclease I and lambda exonuclease. Circular DNA was then denatured and amplified using an oligo mix containing oligos 20, 21, 22 and 23, and a phi29 polymerase to amplify DNA containing a deletion scar. This amplified product was then fragmented again using Covaris to 300 bp, and NEBNext Adaptor for Illumina was annealed to the linearized DNA. This DNA was then sequenced using paired-end reads of 150 bp in an Illumina Hi-Seq 2500. The raw data were submitted to the ArrayExpress database (http://www.ebi.ac.uk/arrayexpress) and assigned the accession identifier E-MTAB-9582.

From the sequencing data, paired reads were extracted that contained the inverted repeat sequence from the transposons, a sequence of genomic DNA, the adapter sequence from the circularization protocol and a second sequence of genomic DNA. We used basic bash tools to trim the adapter sequence and BlastN to map against *M. pneumoniae* M129 genome (Accession Number: NC_000912). Then, we used custom Python scripts to detect deletion points selecting for reads where the two halves of the read map to two different genomic loci. The scripts required to run these processing steps can be found in https://github.com/CRG-CNAG/Fastq2LoxDel. With this, we extracted a list of genomic positions representing each side of a deletion and a read count value representing how many times they were found (Table S1).

Validation of inversion removal

To ensure that the protocol only selected for cells that had undergone a deletion, the pool of cells containing all three transformations was grown to mid-log phase and isolated via the centrifugation protocol described above. The cells were then serial diluted in Hayflick media to a 10^{-5} concentration, and 100 μl was lawn plated onto freshly prepared Hayflick agar plates. Plates were incubated at 37°C for 10 days. 100 individual colonies were picked and suspended in wells containing 200 μl of Hayflick in 96-well plates. The cultures were homogenized, and separate 50 μl aliquots were inoculated into wells containing 150 μl Hayflick supplemented with 25 ng/μl chloramphenicol, 150 μl Hayflick with 2.5 ng/μl tetracycline and 150 μl plain Hayflick respectively. 150 μl Hayflick was also added to the original well to restore it to 200 μl. The plates were incubated at 37°C for 7 days. After this time, colonies were assayed for their ability to grow in plain Hayflick media vs media containing the antibiotic resistances conferred by the transposons.

Deletion validation

Selected deletions uncovered through the sequencing protocol were validated using PCRs of the genomic DNA from the pooled cells. 100μl of the cells that grew after transformation with the pMTnCreGm plasmid had their genomic DNA isolated via the MasterPure Genomic DNA Extraction Kit (Invitrogen). Oligos described in Table S9 were used to amplify specific regions in both the deletion pool and WT DNA, and differences in band sizes were visualized via gel electrophoresis. Specific bands showing a deletion were cut, and DNA was purified via the Qiagen Gel Purification Kit and sequenced via Sanger sequencing to confirm the presence of the lox72 site and genomic loci.

Statistical analysis

Statistical analysis performed on the transposon libraries was done using ANUBIS (Miravet-Verde *et al.*, 2020) to define the coverage and read count metrics by bins included in the analyses. Correlations between previous and new transposon sequencing samples were evaluated using Pearson's R correlation coefficient from scipy package in Python. Due to the transposon libraries having $n = 1$, further detailed analysis was deemed unfeasible. However, our transposon results show very high levels of correlation with data produced by Lluch Senar et al., (2015), whose data sets had much more statistical power. Relevant comparisons are shown in the Table S1-10. Confidence in the accuracy of the deletion library was tested using receiver operating characteristics (ROC). All results and libraries generated are available in the Table S1-10.

Funding information

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement 634942 (MycoSynVac) and was also financed by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme, under grant agreement 670216 (MYCOCHASSIS). We also acknowledge support of the Spanish Ministry of Economy, Industry and Competitiveness (MEIC) to the EMBL partnership, the Centro de Excelencia Severo Ochoa and the CERCA Programme/Generalitat de Catalunya.

Conflict of interest

None declared.

References

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

Annaluru, N., Ramalingam, S., and Chandrasegaran, S. (2015) Rewriting the blueprint of life by synthetic genomics and genome engineering. *Genome Biol* **16:** 125.

Ara, K., Ozaki, K., Nakamura, K., Yamane, K., Sekiguchi, J., and Ogasawara, N. (2007) Bacillus minimum genome factory: effective utilization of microbial genome information. *Biotechnol Appl Biochem* **46:** 169–178.

Ausländer, S., Wieland, M., and Fussenegger, M. (2012) Smart medication through combination of synthetic biology and cell microencapsulation. *Metab Eng* **14:** 252–260.

Berzin, V., Kiriukhin, M., and Tyurin, M. (2012) Cre-lox66/lox71-based elimination of phosphotransacetylase or acetaldehyde dehydrogenase shifted carbon flux in acetogen rendering selective overproduction of ethanol or acetate. *Appl Biochem Biotechnol* **168:** 1384–1393.

Bloodworth, R.A.M., Gislason, A.S., and Cardona, S.T. (2013) *Burkholderia cenocepacia* conditional growth mutant library created by random promoter replacement of essential genes. *MicrobiologyOpen* **2:** 243–258.

Breuer, M., Earnest, T.M., Merryman, C., Wise, K.S., Sun, L., Lynott, M.R. (2019) Essential metabolism for a minimal cell. *eLife* **8.** https://doi.org/10.7554/eLife.36842

Burgos, R., and Totten, P.A. (2014) Characterization of the operon encoding the Holliday junction helicase RuvAB from *Mycoplasma genitalium* and its role in mgpB and mgpC gene variation. *J Bacteriol* **196:** 1608–1618.

Cameron, D.E., Bashor, C.J., and Collins, J.J. (2014) A brief history of synthetic biology. *Nat Rev Microbiol* **12:** 381–390.

Cerisy, T., Rostain, W., Chhun, A., Boutard, M., Salanoubat, M., and Tolonen, A.C. (2019) A targetron-recombinase system for large-scale genome engineering of clostridia. *mSphere* **4:** https://doi.org/10.1128/mSphere.00710-19

Chi, H., Wang, X., Shao, Y., Qin, Y., Deng, Z., Wang, L., and Chen, S. (2019) Engineering and modification of microbial chassis for systems and synthetic biology. *Synth Syst Biotechnol* **4:** 25–33.

Choe, D., Cho, S., Kim, S.C., and Cho, B.-K. (2016) Minimal genome: worthwhile or worthless efforts toward being smaller? *Biotechnol J* **11:** 199–211.

Claesen, J., and Fischbach, M.A. (2015) Synthetic microbes as drug delivery systems. *ACS Synth Biol* **4:** 358–364.

D'Halluin, K., and Ruiter, R. (2013) Directed genome engineering for genome optimization. *Int J Dev Biol* **57:** 621–627.

Dewall, M.T., and Cheng, D.W. (2011) The minimal genome: a metabolic and environmental comparison. *Brief Funct Genomics* **10:** 312–315.

Van Duyne, G.D. (2015) Cre recombinase. *Microbiol Spectr* **3:** MDNA3-0014–2014.

Esvelt, K.M., and Wang, H.H. (2013) Genome-scale engineering for systems and synthetic biology. *Mol Syst Biol* **9:** 641.

Fadiel, A., Eichenbaum, K.D., El Semary, N., and Epperson, B. (2007) Mycoplasma genomics: tailoring the genome for minimal life requirements through reductive evolution. *Front Biosci J Virtual Libr* **12:** 2020–2028.

Folcher, M., and Fussenegger, M. (2012) Synthetic biology advancing clinical applications. *Curr Opin Chem Biol* **16:** 345–354.

Fraser, C.M., Gocayne, J.D., White, O., Adams, M.D., Clayton, R.A., Fleischmann, R.D., *et al.* (1995) The minimal gene complement of *Mycoplasma genitalium*. *Science* **270:** 397–403.

Gawronski, J.D., Wong, S.M.S., Giannoukos, G., Ward, D.V., and Akerley, B.J. (2009) Tracking insertion mutants within libraries by deep sequencing and a genome-wide screen for Haemophilus genes required in the lung. *Proc Natl Acad Sci USA* **106:** 16422–16427.

Ghosh, K., and Van Duyne, G.D. (2002) Cre-loxP biochemistry. *Methods San Diego Calif* **28:** 374–383.

Gibson, D.G., Glass, J.I., Lartigue, C., Noskov, V.N., Chuang, R.-Y., Algire, M.A., *et al.* (2010) Creation of a bacterial cell controlled by a chemically synthesized genome. *Science* **329:** 52–56.

Gibson, D.G., Young, L., Chuang, R.-Y., Venter, J.C., Hutchison, C.A., and Smith, H.O. (2009) Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat Methods* **6:** 343–345.

Gil, R., Silva, F.J., Peretó, J., and Moya, A. (2004) Determination of the core of a minimal bacterial gene set. *Microbiol Mol Biol Rev* **68:** 518–537.

Glass, J.I., Assad-Garcia, N., Alperovich, N., Yooseph, S., Lewis, M.R., Maruf, M., *et al.* (2006) Essential genes of a minimal bacterium. *Proc Natl Acad Sci USA* **103:** 425–430.

Glass, J.I., Merryman, C., Wise, K.S., Hutchison, C.A., and Smith, H.O. (2017) *Minimal Cells-Real and Imagined*. Cold Spring Harb: Perspect. Biol.

Goeddel, D.V., Kleid, D.G., Bolivar, F., Heyneker, H.L., Yansura, D.G., Crea, R., *et al.* (1979) Expression in *Escherichia coli* of chemically synthesized genes for human insulin. *Proc Natl Acad Sci USA* **76:** 106–110.

van Ham, R.C.H.J., Kamerbeek, J., Palacios, C., Rausell, C., Abascal, F., Bastolla, U., *et al.* (2003) Reductive genome evolution in *Buchnera aphidicola*. *Proc Natl Acad Sci USA* **100:** 581–586.

Hayflick, L. (1965) Tissue cultures and mycoplasmas. *Tex Rep Biol Med* **23:** Suppl 1:285.

Hedreyda, C.T., Lee, K.K., and Krause, D.C. (1993) Transformation of *Mycoplasma pneumoniae* with Tn4001 by electroporation. *Plasmid* **30:** 170–175.

Himmelreich, R., Plagens, H., Hilbert, H., Reiner, B., and Herrmann, R. (1997) Comparative analysis of the genomes of the bacteria *Mycoplasma pneumoniae* and *Mycoplasma genitalium*. *Nucleic Acids Res* **25:** 701–712.

Hörner, M., Reischmann, N., and Weber, W. (2012) Synthetic biology: programming cells for biomedical applications. *Perspect Biol Med* **55:** 490–502.

Hsu, P.D., Lander, E.S., and Zhang, F. (2014) Development and applications of CRISPR-Cas9 for genome engineering. *Cell* **157:** 1262–1278.

Hutchison, C.A., Chuang, R.-Y., Noskov, V.N., Assad-Garcia, N., Deerinck, T.J., Ellisman, M.H., *et al.* (2016) Design and synthesis of a minimal bacterial genome. *Science* **351:** aad6253.

Juhas, M., Eberl, L., and Glass, J.I. (2011) Essence of life: essential genes of minimal genomes. *Trends Cell Biol* **21:** 562–568.

Koonin, E. V. (2000) How many genes can make a cell: the minimal-gene-set concept. *Annu Rev Genomics Hum Genet* **1:** 99–116.

Land, M., Hauser, L., Jun, S.-R., Nookaew, I., Leuze, M.R., Ahn, T.-H., *et al.* (2015) Insights from 20 years of bacterial genome sequencing. *Funct Integr Genomics* **15:** 141–161.

Leprince, A., de Lorenzo, V., Völler, P., van Passel, M.W.J., and Martins dos Santos, V.A.P. (2012) Random and cyclical deletion of large DNA segments in the genome of *Pseudomonas putida*. *Environ Microbiol* **14:** 1444–1453.

Li, J., Wang, H.-T., Wang, W.-T., Zhang, X.-R., Suo, F., Ren, J.-Y., *et al.* (2019) Systematic analysis reveals the prevalence and principles of bypassable gene essentiality. *Nat Commun* **10:** 1–15.

Ll, D. (2020) Resurrection from lethal knockouts: bypass of gene essentiality. *Biochem Biophys Res Commun* **528:** 405–412.

Lluch-Senar, M., Delgado, J., Chen, W.-H., Lloréns-Rico, V., O'Reilly, F.J., Wodke, J.A., *et al.* (2015) Defining a minimal cell: essentiality of small ORFs and ncRNAs in a genome-reduced bacterium. *Mol Syst Biol* **11:** 780.

Mariscal, A.M., González-González, L., Querol, E., and Piñol, J. (2016) All-in-one construct for genome engineering using Cre-lox technology. *DNA Res Int J Rapid Publ Rep Genes Genomes* **23:** 263–270.

Martínez-García, E., Jatsenko, T., Kivisaar, M., and de Lorenzo, V. (2015) Freeing *Pseudomonas putida* KT2440 of its proviral load strengthens endurance to environmental stresses. *Environ Microbiol* **17:** 76–90.

Miravet-Verde, S., Burgos, R., Delgado, J., Lluch-Senar, M., and Serrano, L. (2020) FASTQINS and ANUBIS: two bioinformatic tools to explore facts and artifacts in transposon sequencing and essentiality studies. *Nucleic Acids Res* **48:** e102.

Miravet-Verde, S., Ferrar, T., Espadas-García, G., Mazzolini, R., Gharrab, A., Sabido, E., *et al.* (2019) Unraveling the hidden universe of small proteins in bacterial genomes. *Mol Syst Biol* **15:** e8290.

Mol, M., Kabra, R., and Singh, S. (2018) Genome modularity and synthetic biology: engineering systems. *Prog Biophys Mol Biol* **132:** 43–51.

Montero-Blay, A., Miravet-Verde, S., Lluch-Senar, M., Piñero-Lambea, C., and Serrano, L. (2019) SynMyco transposon: engineering transposon vectors for efficient transformation of minimal genomes. *DNA Res Int J Rapid Publ Rep Genes Genomes* **26:** 327–339.

Montero-Blay, A., Piñero-Lambea, C., Miravet-Verde, S., Lluch-Senar, M., and Serrano, L. (2020) Inferring active metabolic pathways from proteomics and essentiality data. *Cell Rep* **31:** 107722.

Nakane, D., Adan-Kubo, J., Kenri, T., and Miyata, M. (2011) Isolation and characterization of P1 adhesin, a leg protein of the gliding bacterium *Mycoplasma pneumoniae*. *J Bacteriol* **193:** 715–722. https://doi.org/10.1128/JB.00796-10.

Otwinowski, J., McCandlish, D.M., and Plotkin, J. B. (2018) Inferring the shape of global epistasis. *Proc Natl Acad Sci USA.* **115:** E7550–E7558.

Overmars, L., van Hijum, S.A.F.T., Siezen, R.J., and Francke, C. (2015) CiVi: circular genome visualization with unique features to analyze sequence elements. *Bioinformatics* **31:** 2867–2869.

Pan, R., Zhang, J., Shen, W.-L., Tao, Z.-Q., Li, S.-P., and Yan, X. (2011) Sequential deletion of *Pichia pastoris* genes by a self-excisable cassette. *FEMS Yeast Res* **11:** 292–298.

Parrott, G. L., Kinjo, T., and Fujita, J. (2016) A compendium for *Mycoplasma pneumoniae*. *Front Microbiol* **7:** 513.

Phillips, P.C. (2008) Epistasis — the essential role of gene interactions in the structure and evolution of genetic systems. *Nat Rev Genet* **9:** 855–867.

Pich, O.Q., Burgos, R., Planell, R., Querol, E., and Piñol, J. (2006) Comparative analysis of antibiotic resistance gene markers in *Mycoplasma genitalium*: application to studies of the minimal gene complement. *Microbiol Read Engl* **152:** 519–527.

Pour-El, I., Adams, C., and Minion, F.C. (2002) Construction of mini-Tn4001tet and its use in *Mycoplasma gallisepticum*. *Plasmid* **47:** 129–137.

Prasher, D.C., Eckenrode, V.K., Ward, W.W., Prendergast, F.G., and Cormier, M.J. (1992) Primary structure of the *Aequorea victoria* green-fluorescent protein. *Gene* **111:** 229–233.

Ruder, W.C., Lu, T., and Collins, J.J. (2011) Synthetic biology moving into the clinic. *Science* **333:** 1248–1252.

Sailer, Z.R., and Harms, M.J. (2017) Detecting high-order epistasis in nonlinear genotype-phenotype maps. *Genetics* **205:** 1079–1088.

Sassetti, C.M., Boyd, D.H., and Rubin, E.J. (2001) Comprehensive identification of conditionally essential genes in mycobacteria. *Proc Natl Acad Sci USA* **98:** 12712–12717.

Shaw, D. (2019) *Streamlining minimal bacterial genomes : Analysis of the pan bacterial essential genome, and a novel strategy for random genome deletions in Mycoplasma pneumoniae*. TDX Tesis Dr. En Xarxa: Universitat Pompeu Fabra (Ph.D. Thesis).

Sung, B.H., Choe, D., Kim, S. C., and Cho, B.-K. (2016) Construction of a minimal genome as a chassis for synthetic biology. *Essays Biochem* **60:** 337–346.

Tatusov, R.L., Galperin, M.Y., Natale, D.A., and Koonin, E.V. (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* **28:** 33–36.

Tsuge, Y., Suzuki, N., Inui, M., and Yukawa, H. (2007) Random segment deletion based on IS31831 and Cre/loxP excision system in *Corynebacterium glutamicum*. *Appl Microbiol Biotechnol* **74:** 1333–1341.

Vernyik, V., Karcagi, I., Tímár, E., Nagy, I., Györkei, Á., Papp, B., *et al.* (2020) Exploring the fitness benefits of genome reduction in *Escherichia coli* by a selection-driven approach. *Sci Rep* **10:** 7345.

Vickers, C.E., Blank, L.M., and Krömer, J.O. (2010) Grand challenge commentary: chassis cells for industrial biochemical production. *Nat Chem Biol* **6:** 875–877.

Waites, K.B., and Talkington, D.F. (2004) *Mycoplasma pneumoniae* and its role as a human pathogen. *Clin Microbiol Rev* **17:** 697–728.

Weinreich, D.M., Lan, Y., Wylie, C.S., and Heckendorn, R.B. (2013) Should evolutionary geneticists worry about higher-order epistasis? *Curr Opin Genet Dev* **23:** 700–707.

Xiao, L., Ptacek, T., Osborne, J.D., Crabb, D.M., Simmons, W.L., Lefkowitz, E.J., *et al.* (2015) Comparative genome analysis of *Mycoplasma pneumoniae*. *BMC Genom* **16:** https://doi.org/10.1186/s12864-015-1801-0

Xin, Y., Guo, T., Mu, Y., and Kong, J. (2018) Coupling the recombineering to Cre-lox system enables simplified large-scale genome deletion in *Lactobacillus casei*. *Microb Cell Factories* **17:** 21.

Yus, E., Maier, T., Michalodimitrakis, K., van Noort, V., Yamada, T., Chen, W.-H., *et al.* (2009) Impact of genome reduction on bacterial metabolism and its regulation. *Science* **326:** 1263–1268.

Zhang, X. (2010) From synthetic genome to creation of life. *Protein Cell* **1:** 501–502.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Fig. S1. Frequency of deletions by size in our database. (A) Histogram showing the distribution in size of all the deletions by base pairs that do not contain an essential gene (E_Out). (B) Histogram showing the distribution in size by base pairs of all the deletions that do contain an essential gene (E_In).

Table S1-S10.