LoxTnSeq: Transposon mutagenesis coupled with ultra-sequencing to study large random genome reductions

Daniel Shaw[1], Samuel Miravet-Verde[1], Carlos Pinero-Lambea[1,2], Luis Serrano[1,3,4]* and Maria Lluch-Senar[1,2]*

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

[2]Pulmobiotics SL, Carrer del Dr. Aiguader, 88, 08003 Barcelona, Spain

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

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

*Corresponding authors: Maria.lluch@crg.eu (M.L), luis.serrano@crg.eu (L.S.)

Summary

Rational engineering in synthetic biology requires preliminary knowledge of which genomic regions are dispensable. Typically, these efforts are guided by transposon mutagenesis studies, coupled to ultra-sequencing (TnSeq) which determine single gene essentiality. However, epistatic interactions can rapidly alter these profiles post deletion, leading to the redundancy of these 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, with ultra-sequencing. When LoxTnSeq was applied to the genome reduced bacterium *Mycoplasma pneumoniae*, we obtained a mutant pool containing 285 unique deletions. These deletions spanned from >50 bp to 28Kb, which represents 21% of the total genome. LoxTnSeq also highlighted large regions of non-essential genes that could be removed simultaneously, and other similar regions that could not, providing a guide for future genome reductions.

Introduction

One of the core principles behind synthetic biology is the rational engineering of an organism to elicit a modification in phenotype orientated for a specific application (D'Halluin and Ruiter, 2013; Esvelt and Wang, 2013; Mol et al., 2018). The introduction of new genes, allowing for

the generation of new proteins within a system, is well documented and has 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 healthcare 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 wild-type cells. 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 strain. 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).

Fortunately, these two processes can work synergistically, with large scale genome reductions showing that the removal of superfluous genes can both increase protein production and develop beneficial characteristics for a cell. Sequential genome reduction has been an engineering tool in *Bacillus subtilis,* creating a species with a boosted protein production phenotype by systematically removing genes that hinder protein production, or divert energy and resources to less functionally useful applications (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 (Annaluru et al., 2015; Cameron et al., 2014; Hsu et al., 2014). One of the grand long-term goals in systems and synthetic biology is the generation of a minimal chassis cell (Chi et al., 2019; Sung et al., 2016). While the definition of a true 'minimal cell' is almost impossible to define (Choe et al., 2016; Glass et al., 2017; Koonin, 2000), 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 (Hutchison et al., 2016; Juhas et al., 2011; Zhang, 2010).

While naturally occurring 'minimal' bacteria do exist to some degree (Fadiel et al., 2007; Fraser et al., 1995; van Ham et al., 2003), 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 in the region 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).

However, our understanding of epistatic networks, 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 JCVI-Syn2.0 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 8 sections, and each section had their NE genes removed independently, while the other 7 sections were kept intact. Despite removing only NE genes, only 1/8 configurations resulted in a viable cell (Hutchison et al., 2016).

This is also exemplified by 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 essential. 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). Fitness genes, also known as quasi-essential genes (Glass et al., 2006) are nominally NE. However, while their loss is survivable by the bacterium, it imparts a severe growth defect.

This is a key limitation in many genome reduction studies, as rationally designed reduction pathways often rely on data generated before genetic manipulation occurs. As such, any *a priori* assumptions about gene essentiality can be potentially redundant after any reductions have occurred. Screens by transposon mutagenesis after every cycle of reductions can be applied (Hutchison et al., 2016) but it is both time and labour intensive, and relies on the basis that genes act alone, not 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”. In the yeast *Schizosaccharomyces pombe*, 27% of E genes on chromosome II-L could be rendered NE when a different gene is 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 optimal designs, as our knowledge of 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.

Therefore, here we propose a new methodology, LoxTnSeq, to study the concept of genome reduction in an unbiased manner. By combining a randomised genome deletion protocol with DNA-ultra sequencing, we can identify the large putative reductions that are possible within a genome. This protocol allowed us to delete large sections of DNA without biasing the results by any potentially misguided *a priori* assumptions.

We have applied the technique in a genome reduced bacterium *Mycoplasma pneumoniae*, considered a model of minimal cell. However, this protocol 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, pMTnTcLox71 and pMTnCreGm were obtained to generate different libraries of transposon mutants. 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. To properly assess the coverage of the first transformation, pMTnLox66Cm transposon insertion sites were identified by HITS (Gawronski et al., 2009). Figure 1 shows the insertion pattern of the pMTnLox66Cm transposon.

The sequencing revealed 201891 unique insertions sites, which across the 816Kb genome leads to an insertion every ≈4 bases on average, similar to what has previously been described (Lluch-Senar et al., 2015).

Creation of a pool of genome reduced mutants

This first pool of mutants was then transformed with the pMTnLox71Tc vector. After applying HITS to this library the number of unique insertions for the second transposon was 90530, suggesting that the pool of mutants was exhibiting some epistatic effects in regard to changes in essential regions due to previous transposon insertions. However, while there

are less unique inserts in the second round, there is still a coverage of 1 insertion every 9 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. Of note, transposon insertions take place randomly not only in terms of the chromosome position, but also in respect to their orientation. It is known that lox sites orientation is responsible from the action mode of Cre, 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 Figure 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 gentamicin resistance. As *M. pneumoniae* cannot support replicating plasmids, introducing the Cre via transposon ensured constitutive expression, and thus its selective effects were not lost after cell divisions. In previous work we have demonstrated that the action of the Cre recombinase on a single active lox site in *M. pneumoniae* produces a lethal effect on par with a double stranded break in the DNA (Shaw et al., under review). 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 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 or 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 Figure 2.

<u>Identification of random deletions</u>

Genomic DNA from the gentamycin-resistant pool of cells was sequenced via the circularisation protocol described in Methods. A total of 1291712 reads were recovered. Due to the random nature of the circularisation 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, 83%), and those that did not (285, 17%). We found a background of deletions affecting E genes with few reads (41335 reads, 3% of total reads), which range in size from <50bp to spanning over half the entire genome (See Figure 3), which probably are an artefact resulting from the circularization protocol for deep sequencing (see Methods). Those deletions affecting genomic regions with only NE genes were less in number, span smaller regions of the genome, with the longest continuously NE region being ≈30Kb. However, although they also could also contain circularisation artefacts, they accounted for 1250337 reads (97% of total reads, see Figure 3).

Using the read length as the threshold parameter, we performed a Receiver Operating Characteristic (ROC) curve approach to define high-confidence deletions (Figure 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 (Figure 3C). This set was classified as the ‘gold standard’ of deletions, as they had the highest level of reliability. When excluding all deletions that contained an essential gene, 285 unique deletions were identified and mapped to the M129 genome, detailed in Figure 4

below. We considered actual or 'positive' those deletions which strictly covered NE regions, and artefactual or 'negative' when they included E genes.

While the number of deletions within the gold standard is low, representing just 8 deletions (see Suppl. Table 1 and Figure 4), they are all deletions spanning multiple genes, all of which are clearly well tolerated by the cell as they were able to proliferate within the dataset. However, as the data in Figure 3 shows, the vast majority of the reads generated (97%) indicated deleted regions that did not contain an essential gene, and the remaining 3% that did show an even distribution across the genome (see Suppl. Figure 1). Therefore, applying a slightly less stringent test for fidelity, and simply allowing all the deletions that do not contain the removal of a known E gene, we see a much larger coverage of deletions across the genome (see Figure 4).

Of the regions deleted, the largest was 28.7Kb, the smallest <50bp. 147 genes were deleted across the pool, accounting for 171.2Kb (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 were annotated as NE, with the remaining 8 classified as F genes, according to the most recent essentiality data (Lluch-Senar et al., 2015). The full list of deletions can be found in Suppl. Table 1, and the genes deleted in Suppl. Table 2. Of the 259 genes that are classified as non-essential in *M. pneumoniae* (Lluch-Senar et al., 2015), we deleted 56%, with a mean deletion size of 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 Figure 4, yellow bars), with various characteristics. Region A contained seven non-essential genes (*mpn096* to *mpn102*), was ≈10Kb in size, was the most highly represented deletion within the pool accounting for 964628 reads and the only validation within the gold standard. Region B represented a smaller region, containing four non-essential genes (*mpn397* to *mpn400*), a size of ≈5Kb and covered by 593 reads. Region C represented one of the largest available deletions, and contained 19 non-essential genes and 1 fitness gene (*mpn493* to *mpn512*), with a deleted area of ≈25Kb covered by 193

reads. Region D contained 6 NE genes (*mpn368* to *mpn373*) within 7.9Kbs, but was only represented in the population by a single read.

Figure 5 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 are also multiple deletions visible in the deletion amplifications in conditions A, B and D, which were expected from the sequencing data. 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 dataset, 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 Figure 6.

The highlighted regions in Figure 6 show two of the strongest deletion hotspots, located at 120Kb and 610Kb 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, Figure 4 shows a large NE region centred on the 200Kb region of the chromosome. This region contains 13 NE genes (*mpn141* to *mpn153*), yet only 4 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, essential 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.

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 & 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 Suppl. Table 3 for full breakdown).

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 & 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 Suppl. Table 3 for full breakdown).

According to the database of E genes of 47 bacterial species generated by Shaw et al., (under revision), only 7 of the genes deleted were found to be E in any other species, and 98 of the 147 genes deleted had no homologs in the other species analysed. Of the deleted genes, the most highly conserved was *mpn397* (guanosine-3'-5'-bis(diphosphate) 3'-pyrophosphohydrolase *spoT*) which was had a direct homolog in 46/47 species and E in 26% of cases, and the gene that was the most highly E in the database was *mpn324* (Ribonucleoside-diphosphate reductase subunit alpha *nrdE*) which had a direct homolog in 36/47 species, and was essential in 54% of cases.

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).

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 characterised. 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 to 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 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 et al., Under Revision), 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 Figure 6, 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-balance this loss in efficiency.

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 50bp to 28Kb. However, the distribution of these deletions is not uniform, with the six most represented deletions accounting for over 99% of the dataset. 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 indicates 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 529000 and 630000 (Indicated by the blue bars in Figure 4), 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 suppl. Table 2, *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 50bp 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 minimise 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.

This is reflected in the essentiality of the genes that were deleted. Of the 147 genes, only 8 were classified as ‘fitness’ within the original genome. These genes were annotated as such due to the fact that they imparted a severe growth restriction on the cells when lost in a single disruption (Lluch-Senar et al., 2015). However when part of a larger deletion here, the

cells could still compete with those that had only lost NE genes, clearly recovering some level of fitness as a result of the larger deletion than if they had lost the single fitness gene alone. This could be due in part to the retention of non-adherent cells, which was not undertook in previous essentiality studies in *M. pneumoniae* (Lluch-Senar et al., 2015). As such, phenotypes that imparted a loss of fitness in adherence could be well tolerated here. Also, the major adhesin genes are known players in the cell division machinery (Krause and Balish, 2004), and while the P1 and *hmw1* genes are NE, *hmw2* is classified as F and *hmw3* as E (Lluch-Senar et al., 2015). It is possible that the deletion of the large cytadaherance operon between *mpn141* and *mpn153* also affected cell division and was thus not tolerated, while the deletion of other less important cytadherance genes was tolerable, due to the retention of planktonic cells within the protocol.

This implies that there are epistatic effects at work within the genome, as combinatorial gene deletions do not have the same fitness outcome as the sum of their individual deletions (Arnold et al., 2018; Domingo et al., 2019). As shown by the team at JCVI when creating the JCVI-Syn3.0 cell, simply deleting all the genes deemed non-essential from the genome is a non-viable method, as each gene loss has the potential to change the essentiality of multiple other genes, which constantly needs to be re-evaluated (Hutchison et al., 2016). By utilising a randomised large scale deletion protocol, not only can single genes be tested for their updated essentiality, but 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 610Kb and 140Kb most notably. Looking at Figure 4, 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 244Kb, 370Kb and 490Kb. 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 non-essential 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. If a cell is being minimised 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 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 utilised as an attenuation process as well.

In line with this, the protocol 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 utilised within a single cell. This could convert the technique from a screening tool to identify amenable deletions to a self-contained large scale genome minimisation 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 28Kb 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 the loss of multiple genes and the identification of non-essential regions that contain interact epistatically and may cause loss of viability if removed in part of as a whole, despite tier constituent parts being deemed non-essential. We hope that these attributes will make it a useful contribution to the growing synthetic biology tool box.

Methods

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 both Hayflick and Yus et al., (Hayflick, 1965; Yus et al., 2009), supplemented with 100 µg/ml ampicillin, 2 µg/ml tetracycline, 20 µg/ml chloramphenicol, 3.3µg/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) pre autoclave.

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 100µg/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 for 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 Suppl. Tables 4 & 5. The plasmids used in this study were generated as follows:

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

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

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

Random genome reduction protocol

WT M129 cells were transformed according to the protocol outlined by Hedreyda et al., (Hedreyda et al., 1993). Cells were grown to mid-log phase, identified by the Hayflick media changing from red to orange. The media was decanted and the flask was washed 3x with 10ml chilled electroporation buffer (EB: 8mM HEPES, 272nM sucrose, pH 7.4). Cells were scraped into 500µl chilled EB and homogenised via 10x passages through a 25-gauge syringe needle. Aliquots of 50µl of the homogenised cells were mixed with a pre-chilled 30µl EB solution containing the 1pMole 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 500µl Hayflick media and incubated at 37°C for 4 hours. 125µl of transformed cells were then inoculated into T75 cm^2 culture flasks containing 20ml Hayflick and supplemented with 20 µg/ml chloramphenicol.

The transformed cells were grown to mid-log phase, then isolated. To ensure the recovery of planktonic cells, the media was transferred to a 50ml Falcon tube. The flask was then scraped into 500µl 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 500µl chilled EB. The cells were homogenised via 10x passages through a 25-gauge syringe needle. Aliquots of 50µl of the homogenised cells were mixed with a pre-chilled 30µl EB solution containing the 1pMole of pMTnLox71Tc plasmid DNA, and transformed using the previously described settings. Cells were recovered in the same manner, and 125µl were inoculated into a T75 cm^2 flask containing 20ml 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 50µl of the homogenised cells were mixed with a pre-chilled 30µl EB solution containing the 1pMole of pMTnCreGm plasmid DNA, and transformed using the previously described settings. Cells were recovered in the same manner, and 125µl were inoculated into a T75 cm^2 flask containing 20ml Hayflick media supplemented with 200 µg/ml gentamicin.

<u>Quantification of transposon density</u>

WT M129 cells were transformed with 1pMole of the plasmid pMTnLox66Cm, via the electroporation specified above. 125µl of transformed cells were then inoculated into a T75 cm^2 culture flask containing 20ml Hayflick and supplemented with chloramphenicol. The culture was 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 125bp paired-end read library preparation protocol for an Illumina MiSeq. Insertion sites for the transposon were identified using Oligo 19 (Suppl. Table 5), which is bound to the 3' end of the chloramphenicol resistance genes,

directly downstream of the inverted repeat sequence, and identified in the *M. pneumoniae* M129 genome using FASTQINS and ANUBIS (Miravet-Verde et al., under revision).

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 300bp 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 primer mix containing LE-RE_Lox72, loxloop-6, loxloop-7 and loxloop-8 and a phi29 polymerase to amplify DNA containing a deletion scar. This amplified product was then fragmented again using Covaris to 300bp and NEBNext Adaptor for Illumina were annealed to the linearised DNA. This DNA was then sequenced using paired end reads of 150bp in an Illumina Hi-Seq 2500.

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 circularisation 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.

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 100µl 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 200µl of Hayflick in 96 well plates. The cultures were homogenised, and separate 50µl aliquots were inoculated into wells containing 150µl Hayflick supplemented with 25 ng/µl chloramphenicol, 150µl Hayflick with 2.5 ng/µl tetracycline and 150µl plain Hayflick respectively. 150µl Hayflick was also added to the original well to restore it to 200µl. 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 Suppl. Table 5 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.

References

Annaluru, N., Ramalingam, S., Chandrasegaran, S., 2015. Rewriting the blueprint of life by synthetic genomics and genome engineering. Genome Biol. 16, 125. https://doi.org/10.1186/s13059-015-0689-y

Ara, K., Ozaki, K., Nakamura, K., Yamane, K., Sekiguchi, J., Ogasawara, N., 2007. Bacillus minimum genome factory: effective utilization of microbial genome information. Biotechnol. Appl. Biochem. 46, 169–178. https://doi.org/10.1042/BA20060111

Arnold, B.J., Gutmann, M.U., Grad, Y.H., Sheppard, S.K., Corander, J., Lipsitch, M., Hanage, W.P., 2018. Weak Epistasis May Drive Adaptation in Recombining Bacteria. Genetics 208, 1247–1260. https://doi.org/10.1534/genetics.117.300662

Ausländer, S., Wieland, M., Fussenegger, M., 2012. Smart medication through combination of synthetic biology and cell microencapsulation. Metab. Eng. 14, 252–260. https://doi.org/10.1016/j.ymben.2011.06.003

Breuer, M., Earnest, T.M., Merryman, C., Wise, K.S., Sun, L., Lynott, M.R., Hutchison, C.A., Smith, H.O., Lapek, J.D., Gonzalez, D.J., de Crécy-Lagard, V., Haas, D., Hanson, A.D., Labhsetwar, P., Glass, J.I., Luthey-Schulten, Z., 2019. Essential metabolism for a minimal cell. eLife 8. https://doi.org/10.7554/eLife.36842

Cameron, D.E., Bashor, C.J., Collins, J.J., 2014. A brief history of synthetic biology. Nat. Rev. Microbiol. 12, 381–390. https://doi.org/10.1038/nrmicro3239

Cerisy, T., Rostain, W., Chhun, A., Boutard, M., Salanoubat, M., 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., Chen, S., 2019. Engineering and modification of microbial chassis for systems and synthetic biology. Synth. Syst. Biotechnol. 4, 25–33. https://doi.org/10.1016/j.synbio.2018.12.001

Choe, D., Cho, S., Kim, S.C., Cho, B.-K., 2016. Minimal genome: Worthwhile or worthless efforts toward being smaller? Biotechnol. J. 11, 199–211. https://doi.org/10.1002/biot.201400838

Claesen, J., Fischbach, M.A., 2015. Synthetic microbes as drug delivery systems. ACS Synth. Biol. 4, 358–364. https://doi.org/10.1021/sb500258b

Dewall, M.T., Cheng, D.W., 2011. The minimal genome: a metabolic and environmental comparison. Brief. Funct. Genomics 10, 312–315. https://doi.org/10.1093/bfgp/elr030

D'Halluin, K., Ruiter, R., 2013. Directed genome engineering for genome optimization. Int. J. Dev. Biol. 57, 621–627. https://doi.org/10.1387/ijdb.130217kd

Domingo, J., Baeza-Centurion, P., Lehner, B., 2019. The Causes and Consequences of Genetic Interactions (Epistasis). Annu. Rev. Genomics Hum. Genet. 20, 433–460. https://doi.org/10.1146/annurev-genom-083118-014857

Esvelt, K.M., Wang, H.H., 2013. Genome-scale engineering for systems and synthetic biology. Mol. Syst. Biol. 9, 641. https://doi.org/10.1038/msb.2012.66

Fadiel, A., Eichenbaum, K.D., El Semary, N., 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., Fussenegger, M., 2012. Synthetic biology advancing clinical applications. Curr. Opin. Chem. Biol. 16, 345–354. https://doi.org/10.1016/j.cbpa.2012.06.008

Fraser, C.M., Gocayne, J.D., White, O., Adams, M.D., Clayton, R.A., Fleischmann, R.D., Bult, C.J., Kerlavage, A.R., Sutton, G., Kelley, J.M., Fritchman, R.D., Weidman, J.F., Small, K.V., Sandusky, M., Fuhrmann, J., Nguyen, D., Utterback, T.R., Saudek, D.M., Phillips, C.A., Merrick, J.M., Tomb, J.F., Dougherty, B.A., Bott, K.F., Hu, P.C., Lucier, T.S., Peterson, S.N., Smith, H.O., Hutchison, C.A., Venter, J.C., 1995. The minimal gene complement of Mycoplasma genitalium. Science 270, 397–403.

Gawronski, J.D., Wong, S.M.S., Giannoukos, G., Ward, D.V., 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. U. S. A. 106, 16422–16427. https://doi.org/10.1073/pnas.0906627106

Ghosh, K., Van Duyne, G.D., 2002. Cre-loxP biochemistry. Methods San Diego Calif 28, 374–383. https://doi.org/10.1016/s1046-2023(02)00244-x

Gibson, D.G., Glass, J.I., Lartigue, C., Noskov, V.N., Chuang, R.-Y., Algire, M.A., Benders, G.A., Montague, M.G., Ma, L., Moodie, M.M., Merryman, C., Vashee, S., Krishnakumar, R., Assad-Garcia, N., Andrews-Pfannkoch, C., Denisova, E.A., Young, L., Qi, Z.-Q., Segall-Shapiro, T.H., Calvey, C.H., Parmar, P.P., Hutchison, C.A., Smith, H.O., Venter, J.C., 2010. Creation of a bacterial cell controlled by a chemically synthesized genome. Science 329, 52–56. https://doi.org/10.1126/science.1190719

Gibson, D.G., Young, L., Chuang, R.-Y., Venter, J.C., Hutchison, C.A., Smith, H.O., 2009. Enzymatic assembly of DNA molecules up to several hundred kilobases. Nat. Methods 6, 343–345. https://doi.org/10.1038/nmeth.1318

Gil, R., Silva, F.J., Peretó, J., Moya, A., 2004. Determination of the Core of a Minimal Bacterial Gene Set. Microbiol. Mol. Biol. Rev. 68, 518–537. https://doi.org/10.1128/MMBR.68.3.518-537.2004

Glass, J.I., Assad-Garcia, N., Alperovich, N., Yooseph, S., Lewis, M.R., Maruf, M., Hutchison, C.A., Smith, H.O., Venter, J.C., 2006. Essential genes of a minimal bacterium. Proc. Natl. Acad. Sci. U. S. A. 103, 425–430. https://doi.org/10.1073/pnas.0510013103

Glass, J.I., Merryman, C., Wise, K.S., Hutchison, C.A., Smith, H.O., 2017. Minimal Cells-Real and Imagined. Cold Spring Harb. Perspect. Biol. https://doi.org/10.1101/cshperspect.a023861
Goeddel, D.V., Kleid, D.G., Bolivar, F., Heyneker, H.L., Yansura, D.G., Crea, R., Hirose, T., Kraszewski, A., Itakura, K., Riggs, A.D., 1979. Expression in Escherichia coli of chemically synthesized genes for human insulin. Proc. Natl. Acad. Sci. U. S. A. 76, 106–110.
Hayflick, L., 1965. Tissue cultures and mycoplasmas. Tex. Rep. Biol. Med. 23, Suppl 1:285+.
Hedreyda, C.T., Lee, K.K., Krause, D.C., 1993. Transformation of Mycoplasma pneumoniae with Tn4001 by electroporation. Plasmid 30, 170–175. https://doi.org/10.1006/plas.1993.1047
Himmelreich, R., Plagens, H., Hilbert, H., Reiner, B., Herrmann, R., 1997. Comparative analysis of the genomes of the bacteria Mycoplasma pneumoniae and Mycoplasma genitalium. Nucleic Acids Res. 25, 701–712. https://doi.org/10.1093/nar/25.4.701
Hörner, M., Reischmann, N., Weber, W., 2012. Synthetic biology: programming cells for biomedical applications. Perspect. Biol. Med. 55, 490–502. https://doi.org/10.1353/pbm.2012.0042
Hsu, P.D., Lander, E.S., Zhang, F., 2014. Development and applications of CRISPR-Cas9 for genome engineering. Cell 157, 1262–1278. https://doi.org/10.1016/j.cell.2014.05.010
Hutchison, C.A., Chuang, R.-Y., Noskov, V.N., Assad-Garcia, N., Deerinck, T.J., Ellisman, M.H., Gill, J., Kannan, K., Karas, B.J., Ma, L., Pelletier, J.F., Qi, Z.-Q., Richter, R.A., Strychalski, E.A., Sun, L., Suzuki, Y., Tsvetanova, B., Wise, K.S., Smith, H.O., Glass, J.I., Merryman, C., Gibson, D.G., Venter, J.C., 2016. Design and synthesis of a minimal bacterial genome. Science 351, aad6253. https://doi.org/10.1126/science.aad6253
Juhas, M., Eberl, L., Glass, J.I., 2011. Essence of life: essential genes of minimal genomes. Trends Cell Biol. 21, 562–568. https://doi.org/10.1016/j.tcb.2011.07.005
Koonin, E.V., 2000. How many genes can make a cell: the minimal-gene-set concept. Annu. Rev. Genomics Hum. Genet. 1, 99–116. https://doi.org/10.1146/annurev.genom.1.1.99
Krause, D.C., Balish, M.F., 2004. Cellular engineering in a minimal microbe: structure and assembly of the terminal organelle of Mycoplasma pneumoniae. Mol. Microbiol. 51, 917–924.
Land, M., Hauser, L., Jun, S.-R., Nookaew, I., Leuze, M.R., Ahn, T.-H., Karpinets, T., Lund, O., Kora, G., Wassenaar, T., Poudel, S., Ussery, D.W., 2015. Insights from 20 years of bacterial genome sequencing. Funct. Integr. Genomics 15, 141–161. https://doi.org/10.1007/s10142-015-0433-4
Li, J., Wang, H.-T., Wang, W.-T., Zhang, X.-R., Suo, F., Ren, J.-Y., Bi, Y., Xue, Y.-X., Hu, W., Dong, M.-Q., Du, L.-L., 2019. Systematic analysis reveals the prevalence and principles of bypassable gene essentiality. Nat. Commun. 10, 1–15. https://doi.org/10.1038/s41467-019-08928-1
Lluch-Senar, M., Delgado, J., Chen, W.-H., Lloréns-Rico, V., O'Reilly, F.J., Wodke, J.A., Unal, E.B., Yus, E., Martínez, S., Nichols, R.J., Ferrar, T., Vivancos, A., Schmeisky, A., Stülke, J., van Noort, V., Gavin, A.-C., Bork, P., Serrano, L., 2015. Defining a minimal cell: essentiality of small ORFs and ncRNAs in a genome-reduced bacterium. Mol. Syst. Biol. 11, 780. https://doi.org/10.15252/msb.20145558
Martínez-García, E., Jatsenko, T., Kivisaar, M., de Lorenzo, V., 2015. Freeing Pseudomonas putida KT2440 of its proviral load strengthens endurance to environmental stresses. Environ. Microbiol. 17, 76–90. https://doi.org/10.1111/1462-2920.12492
Mol, M., Kabra, R., Singh, S., 2018. Genome modularity and synthetic biology: Engineering systems. Prog. Biophys. Mol. Biol. 132, 43–51. https://doi.org/10.1016/j.pbiomolbio.2017.08.002
Montero-Blay, A., Piñero-Lambea, C., Miravet-Verde, S., Lluch-Senar, M., Serrano, L., 2020. Inferring Active Metabolic Pathways from Proteomics and Essentiality Data. Cell Rep. 31, 107722. https://doi.org/10.1016/j.celrep.2020.107722
Nakane, D., Adan-Kubo, J., Kenri, T., 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., Plotkin, J.B., 2018. Inferring the shape of global epistasis. Proc. Natl. Acad. Sci. U. S. A. 115, E7550–E7558. https://doi.org/10.1073/pnas.1804015115

Overmars, L., van Hijum, S.A.F.T., Siezen, R.J., Francke, C., 2015. CiVi: circular genome visualization with unique features to analyze sequence elements. Bioinformatics 31, 2867–2869. https://doi.org/10.1093/bioinformatics/btv249

Pan, R., Zhang, J., Shen, W.-L., Tao, Z.-Q., Li, S.-P., Yan, X., 2011. Sequential deletion of Pichia pastoris genes by a self-excisable cassette. FEMS Yeast Res. 11, 292–298. https://doi.org/10.1111/j.1567-1364.2011.00716.x

Parrott, G.L., Kinjo, T., Fujita, J., 2016. A Compendium for Mycoplasma pneumoniae. Front. Microbiol. 7, 513. https://doi.org/10.3389/fmicb.2016.00513

Prasher, D.C., Eckenrode, V.K., Ward, W.W., Prendergast, F.G., Cormier, M.J., 1992. Primary structure of the Aequorea victoria green-fluorescent protein. Gene 111, 229–233. https://doi.org/10.1016/0378-1119(92)90691-h

Ruder, W.C., Lu, T., Collins, J.J., 2011. Synthetic biology moving into the clinic. Science 333, 1248–1252. https://doi.org/10.1126/science.1206843

Sailer, Z.R., Harms, M.J., 2017. Detecting High-Order Epistasis in Nonlinear Genotype-Phenotype Maps. Genetics 205, 1079–1088. https://doi.org/10.1534/genetics.116.195214

Sung, B.H., Choe, D., Kim, S.C., Cho, B.-K., 2016. Construction of a minimal genome as a chassis for synthetic biology. Essays Biochem. 60, 337–346. https://doi.org/10.1042/EBC20160024

Tatusov, R.L., Galperin, M.Y., Natale, D.A., 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., Yukawa, H., 2007. Random segment deletion based on IS31831 and Cre/loxP excision system in Corynebacterium glutamicum. Appl. Microbiol. Biotechnol. 74, 1333–1341. https://doi.org/10.1007/s00253-006-0788-5

Van Duyne, G.D., 2015. Cre Recombinase. Microbiol. Spectr. 3, MDNA3-0014–2014. https://doi.org/10.1128/microbiolspec.MDNA3-0014-2014

van Ham, R.C.H.J., Kamerbeek, J., Palacios, C., Rausell, C., Abascal, F., Bastolla, U., Fernández, J.M., Jiménez, L., Postigo, M., Silva, F.J., Tamames, J., Viguera, E., Latorre, A., Valencia, A., Morán, F., Moya, A., 2003. Reductive genome evolution in Buchnera aphidicola. Proc. Natl. Acad. Sci. U. S. A. 100, 581–586. https://doi.org/10.1073/pnas.0235981100

Vickers, C.E., Blank, L.M., Krömer, J.O., 2010. Grand challenge commentary: Chassis cells for industrial biochemical production. Nat. Chem. Biol. 6, 875–877. https://doi.org/10.1038/nchembio.484

Waites, K.B., Talkington, D.F., 2004. Mycoplasma pneumoniae and its role as a human pathogen. Clin. Microbiol. Rev. 17, 697–728, table of contents. https://doi.org/10.1128/CMR.17.4.697-728.2004

Weinreich, D.M., Lan, Y., Wylie, C.S., Heckendorn, R.B., 2013. Should evolutionary geneticists worry about higher-order epistasis? Curr. Opin. Genet. Dev. 23, 700–707. https://doi.org/10.1016/j.gde.2013.10.007

Xiao, L., Ptacek, T., Osborne, J.D., Crabb, D.M., Simmons, W.L., Lefkowitz, E.J., Waites, K.B., Atkinson, T.P., Dybvig, K., 2015. Comparative genome analysis of Mycoplasma pneumoniae. BMC Genomics 16. https://doi.org/10.1186/s12864-015-1801-0

Xin, Y., Guo, T., Mu, Y., Kong, J., 2018. Coupling the recombineering to Cre-lox system enables simplified large-scale genome deletion in Lactobacillus casei. Microb. Cell Factories 17, 21. https://doi.org/10.1186/s12934-018-0872-4

Yus, E., Maier, T., Michalodimitrakis, K., Noort, V. van, Yamada, T., Chen, W.-H., Wodke, J.A.H., Güell, M., Martínez, S., Bourgeois, R., Kühner, S., Raineri, E., Letunic, I., Kalinina, O.V., Rode, M., Herrmann, R., Gutiérrez-Gallego, R., Russell, R.B., Gavin, A.-C., Bork, P., Serrano, L., 2009. Impact of Genome Reduction on Bacterial Metabolism and Its Regulation. Science 326, 1263–1268. https://doi.org/10.1126/science.1177263

Zhang, X., 2010. From synthetic genome to creation of life. Protein Cell 1, 501–502. https://doi.org/10.1007/s13238-010-0071-5

Funding

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

Figure Legends

Figure 1: Transposon insertion points in WT *M. pneumoniae* M129 cells transformed with 1pMole of the pMTnLox66Cm vector, corresponding to the position of known essential (red) and non-essential (green) genes.

Figure 2: Protocol schematic, 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. (D) Effect of the Cre on lox sites integrating in *cis* orientation. Under the action of the Cre, the genomic DNA between them is circularised and removed. Here, an active loxP remains within the genome, initiating a lethal phenotype. (E) Effect of the Cre on lox sites integrating in *cis* orientation, but encompassing an essential gene. The desired inactive lox72 site is left within the genome, but the loss of an essential

gene causes cell death. (F) Effect of the Cre on lox sites integrating in *cis* orientation, containing no essential genes. An inactive lox72 is formed within the genome, impervious to the effects of the Cre, and the NE DNA is circularised and removed, with the cell continuing to survive.

Figure 3: Distribution of sequencing data. (A) Relative abundances of the deletions that contain an essential 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 size of deletion. (C) ROC curve of our 'Gold Standard' delineation of 900 reads compared to the 0 read control.

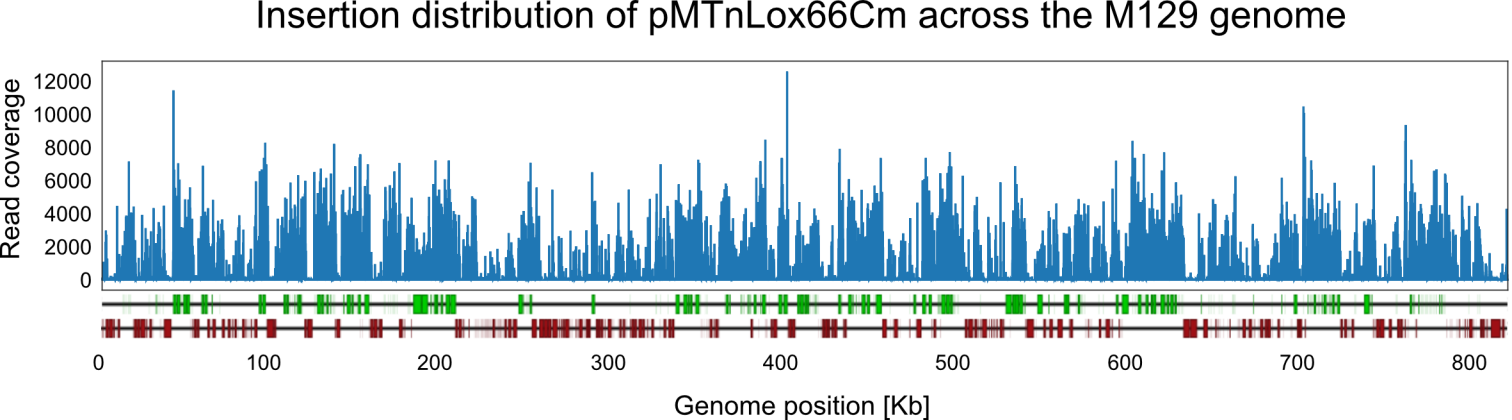
Figure 4: Circos plot showing the locations of the non-essential genes (green, lane 1), essential genes (red, lane 2), deleted regions that were within the gold-standard (purple, lane 3), deleted regions with greater than or equal to 10 reads (dark blue, lane 4) and deleted regions with less than 10 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).

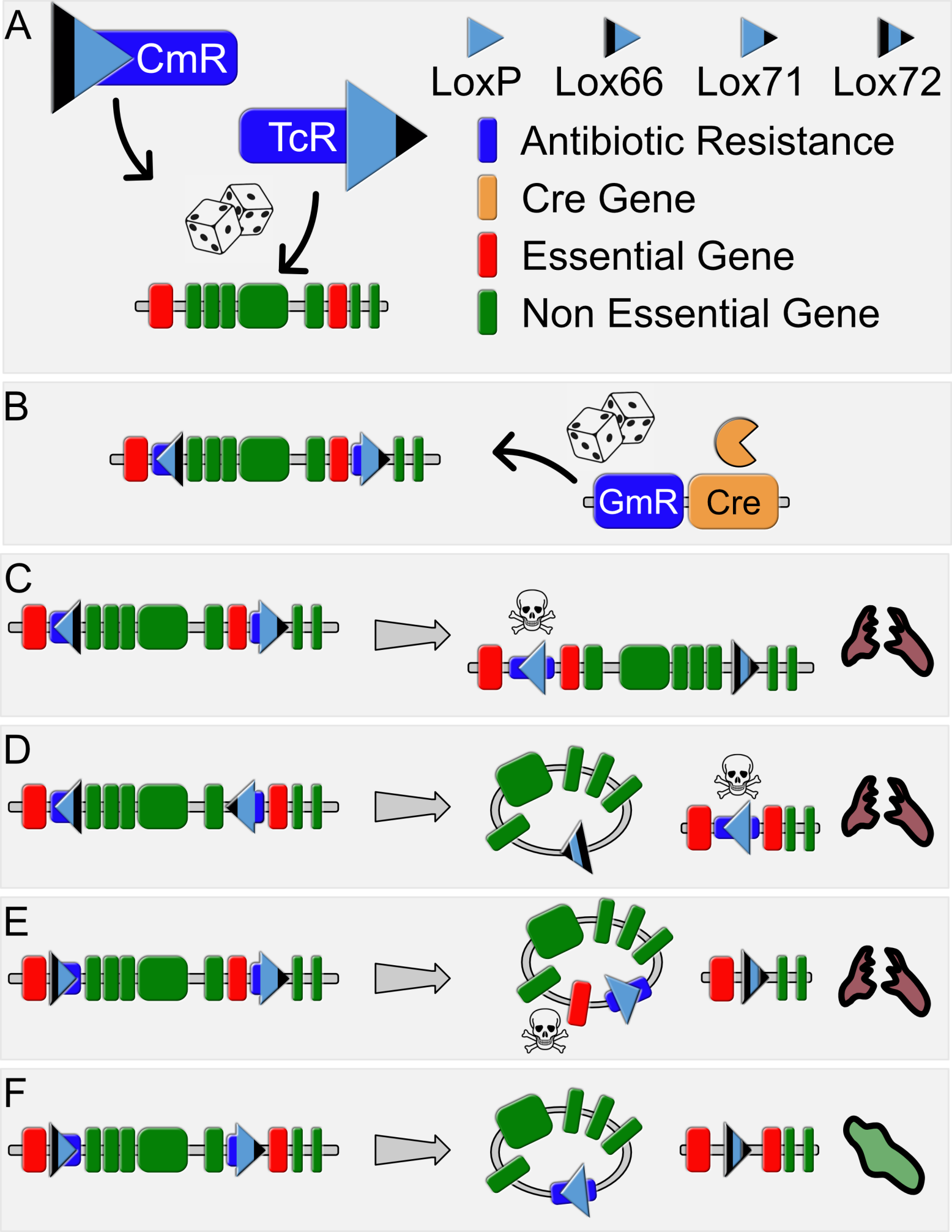
Figure 5: Validation of deletions via PCR. Gel electrophoresis of the four deletions candidates specified in the text above, with the WT on the left and deletions pool on the right in each condition. DNA primers were designed to amplify a 500-1500bp region corresponding to the proposed deletion. Condition A amplified WT genomic DNA and genomic DNA isolated from the pool of deletion cells with oligos 11 & 12. Condition B amplified the same DNAs but using oligos 13 & 14, Condition C used the same DNA but with oligos 15 & 16, and condition D used the same DNAs but with oligos 17 & 18.

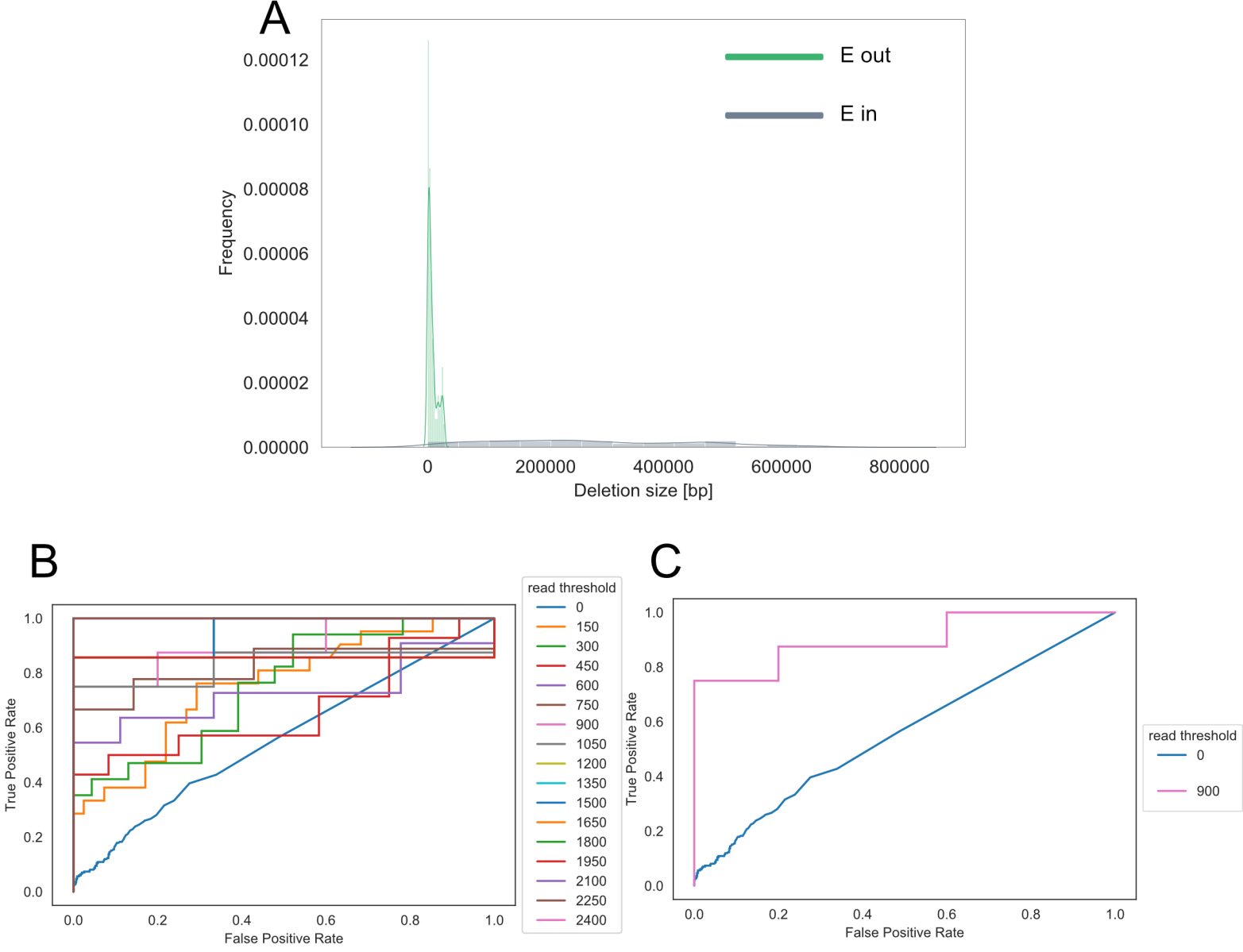
Figure 6: Location of all putative deletions with the M129 genome, with deleted regions shown as black bars. Magnifying glasses show regions with multiple deletions clustering in single regions, with each coloured bar representing a unique deletion within the dataset.

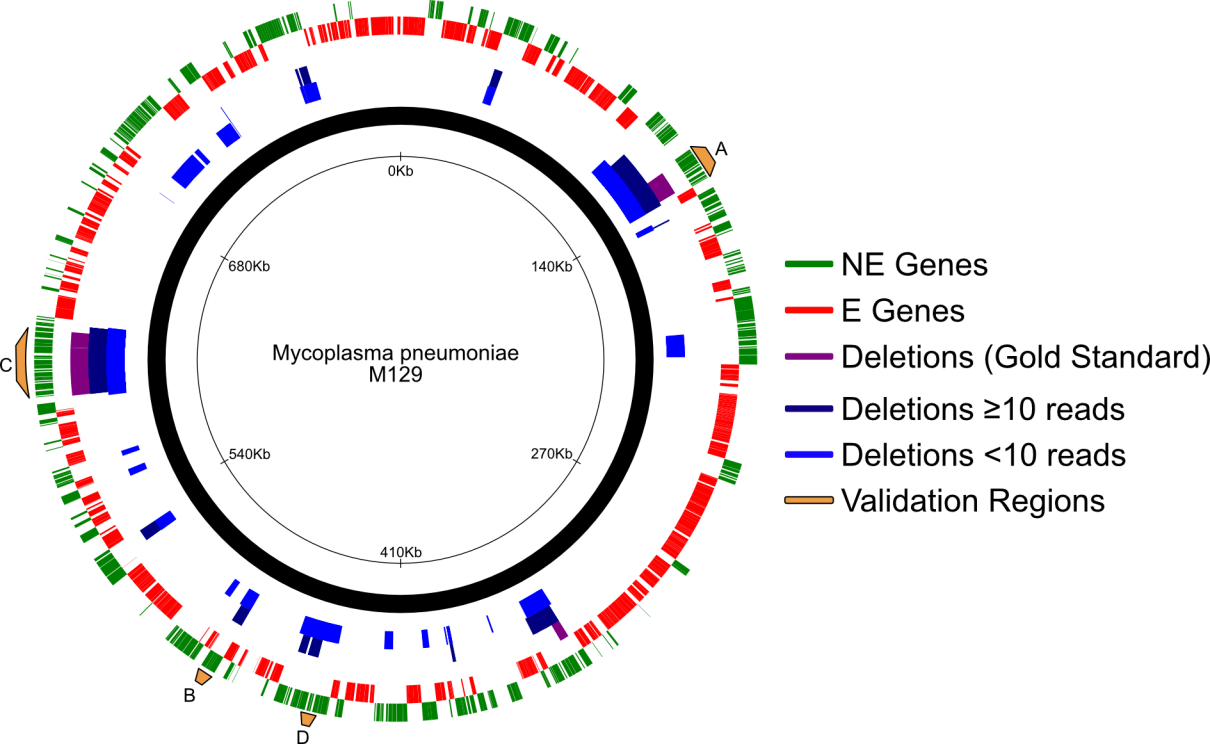
Figure 7: Breakdown of deleted genes via COG categories. [G] Carbohydrate transport and metabolism. [E] Amino acid transport and metabolism. [F] Nucleotide transport and metabolism. [P] Inorganic ion transport and metabolism. [I] Lipid transport and metabolism.

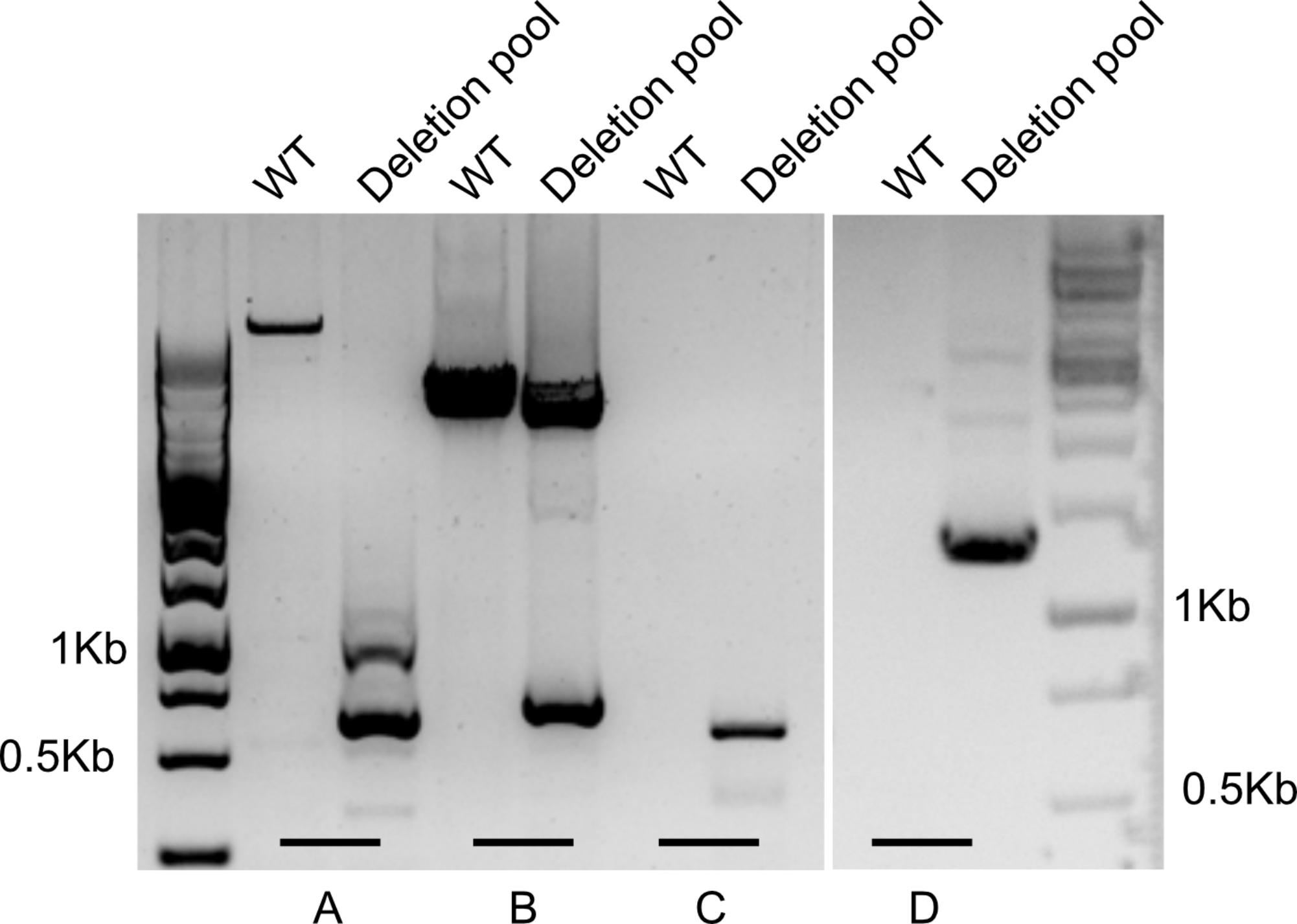
[M] Cell wall/membrane/envelope biogenesis. [V] Defense mechanisms. [N] Cell motility. [O] Post-translational modification, protein turnover, chaperones. [D] Cell cycle control, cell division, chromosome partitioning. [T] Signal transduction mechanisms. [U] Intracellular trafficking, secretion, and vesicular transport. [L] Replication, recombination and repair. [K] Transcription.

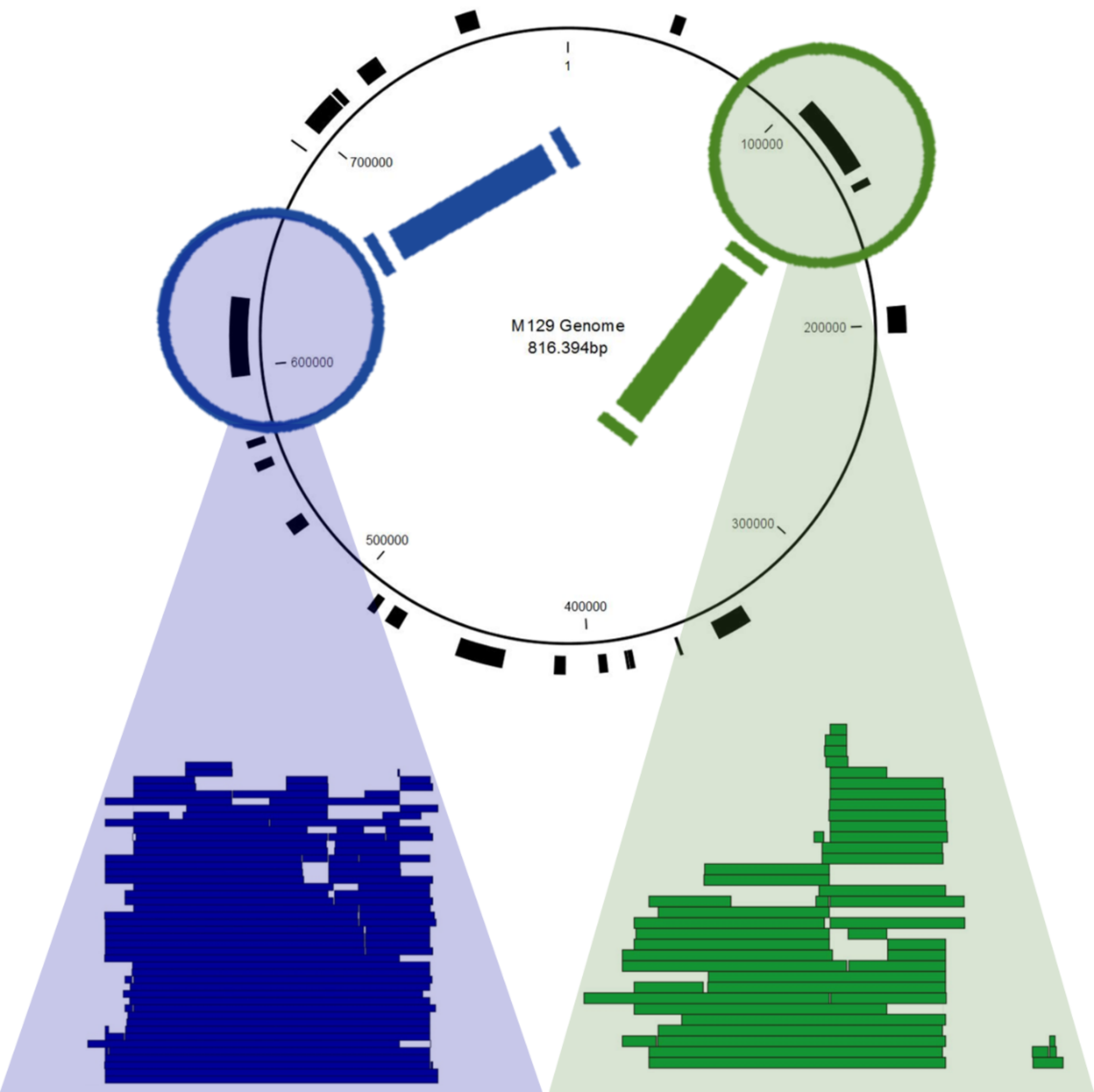

Insertion distribution of pMTnLox66Cm across the M129 genome
Read coverage
12000
10000
8000
6000
4000
2000
0
0
100
200
300
400
500
600
700
800
Genome position [Kb]

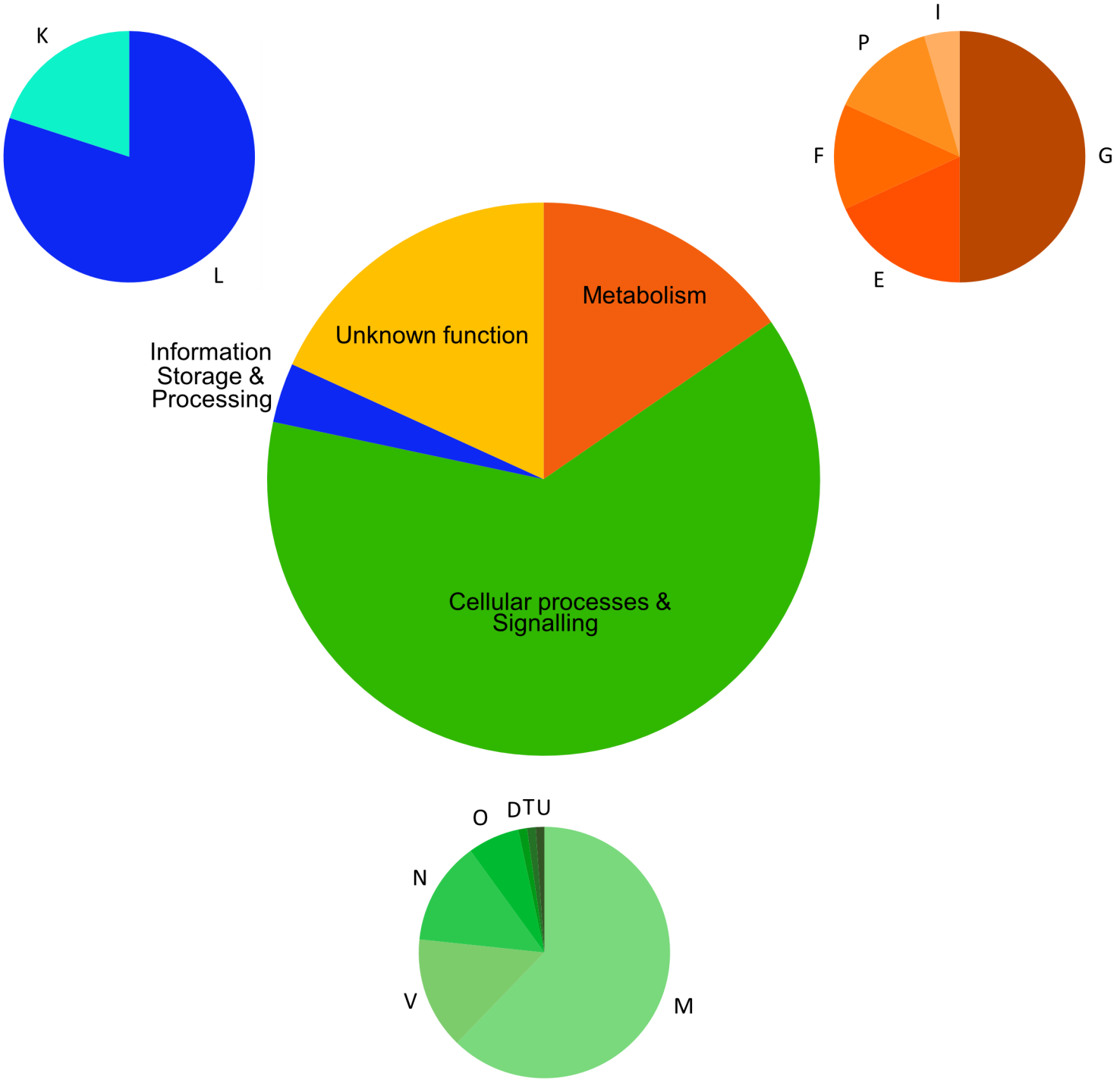

A
CmR
TcR
LoxP
Lox66
Lox71
Lox72
Antibiotic Resistance
Cre Gene
Essential Gene
Non Essential Gene
B
GmR
Cre
C
D
E
F


A
E out
E in
Frequency
0.00012
0.00010
0.00008
0.00006
0.00004
0.00002
0.00000
0
200000
400000
600000
800000
Deletion size [bp]
B
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
read threshold
0
150
300
450
600
750
900
1050
1200
1350
1500
1650
1800
1950
2100
2250
2400
C
True Positive Rate
False Positive Rate
read threshold
0
900


0Kb
140Kb
270Kb
410Kb
540Kb
680Kb
Mycoplasma pneumoniae
M129
A
B
C
D
NE Genes
E Genes
Deletions (Gold Standard)
Deletions ≥10 reads
Deletions <10 reads
Validation Regions


WT
Deletion pool
WT
Deletion pool
WT
Deletion pool
WT
Deletion pool
1Kb
0.5Kb
1Kb
0.5Kb
A
B
C
D


1
100000
200000
300000
400000
500000
600000
700000
M129 Genome
816.394bp


K
L
I
P
F
G
E
Metabolism
Unknown function
Information Storage & Processing
Cellular processes & Signalling
O
DTU
N
V
M