Title: TraDIS-validate: a method for curating ordered gene-replacement libraries

Running Title: Sequencing the Keio library

Authors: Emily C. A. Goodall¹, Faye C. Morris², Samantha A. McKeand², Rudi Sullivan², Isabel A. Warner¹, Emma Sheehan², Gabriela Boelter², Christopher Icke¹, Adam F. Cunningham³, Jeffrey A. Cole², Manuel Banzhaf², Jack A. Bryant², Ian R. Henderson^{1,2*}

¹Institute for Molecular Biosciences, University of Queensland, Brisbane, 4072, Australia

²Institute of Microbiology and Infection, and ³Institute of Immunology and Immunotherapy, University of Birmingham, Birmingham, B15 2TT, UK

*Corresponding author: Ian R. Henderson, i.henderson@uq.edu.au

Abstract

In recent years the availability of genome sequence information has grown logarithmically resulting in the identification of a plethora of uncharacterised genes. To address this gap in functional annotation, many high-throughput screens have been devised with the goal of uncovering novel gene functions. Gene-replacement libraries are one such tool that can be screened in a high-throughput way to link genotype and phenotype and are key community resources. However, for a phenotype to be attributed to a specific gene, there needs to be confidence in the genotype. Construction of large libraries can be laborious and occasionally errors will arise. Here, we present a rapid and accurate method for the validation of any ordered gene-replacement library. We applied our method (TraDIS-Validate) to the well-known Keio library of *Escherichia coli* gene-deletion mutants. Our method identified 3,718 constructed mutants out of a total of 3,728 confirmed isolates. TraDIS-validate therefore had a success rate of 99.7% for identifying the correct kanamycin cassette position. This dataset provides a benchmark for the purity of the Keio mutants and a screening method for mapping the position of an antibiotic resistance cassette in any ordered library.

Keywords: *E. coli* / gene-knockout / Keio / library validation / sequencing / TraDIS

Introduction

Advances in whole genome sequencing have led to an exponential increase in the number of sequenced genomes, and consequently increased the number of uniquely annotated genes. However, the functional annotations of these genes are vastly lagging behind. Functional annotations are a fundamental resource to the field of biology; uncharacterised genes may encode enzymes of biotechnological importance, disease determinants, antimicrobial targets, or may offer information about strain speciation, evolution and diversity. Hence it is an important task to develop systematic approaches to increase functional annotations to be able to harness these untapped resources.

In recent decades there has been significant improvement of high-throughput approaches that are able to assay thousands of mutants to link phenotype to genotype (Baym et al. 2016). The mutant libraries used in these screens generally fall into one of two categories: pooled libraries or ordered libraries. A pooled library comprises a heterogeneous collection of mutants, whereas an ordered library is a collection of arrayed mutants where each individual mutant is stocked

independently of others (Gray et al. 2015). Screening pooled libraries can offer a much higher level of genetic resolution (Goodall et al. 2018). However, the number of mutants in such libraries can be limiting in experiments with a tight bottleneck (Chaudhuri et al. 2013). Smaller, defined libraries can therefore be preferential in more clinically relevant animal studies. An additional strength of screening ordered libraries is the ability to assay each mutant independently and easily access individual mutants for further study. Ordered libraries now exist for a number of diverse organisms, and these collections serve as a fundamental resource for the research community to conduct further functional assays (Giaever et al. 2002; Jacobs et al. 2003; Kobayashi et al. 2003; Baba et al. 2006; Lane and Rubin 2006; Liberati et al. 2006; Gallagher et al. 2007; De Berardinis et al. 2008; Cameron et al. 2008; Enstrom et al. 2012; Porwollik et al. 2014; Baym et al. 2016; Mesarich et al. 2017). However, in smaller, ordered libraries each mutant is often represented by a small number of independent copies (typically 1-2); this therefore places greater importance on each individual mutant being correct. This is significant when considering such libraries will be used in clinically important phenotypic screens.

While ordered gene-deletion libraries are excellent resources for linking phenotype to genotype, errors during construction of gene-deletion mutants can arise, such as gene-duplication events, secondary suppressor mutations, mis-targeting of the selection cassette or accidental interruption of adjacent genes (Yamamoto et al. 2009; S. Mahalakshmi et al. 2014). In addition, library stocking errors and neighbouring well contamination can arise when creating or using such libraries (Gallagher et al. 2007; Cameron et al. 2008). As such, it is important to verify the accuracy of each when using such large, ordered libraries. Whole genome sequencing of each individual mutant would prove challenging and costly, thus previous attempts to validate ordered libraries have relied on PCR to check the site of insertion of the selection cassette (Yamamoto et al. 2009). While such analyses are useful, they do not define the precise point of insertion or whether an adjacent coding sequence is interrupted. Furthermore, if the amplicons for the mutant and wild-type bacterium are of a similar size it can be difficult to discern if the mutant was constructed correctly. Thus, a method to rapidly validate mutations within gene deletion libraries would be useful.

Recently, we benchmarked TraDIS (Transposon-Directed Insertion-site Sequencing), a method that makes use of a pooled library of transposon-insertion mutants, against an ordered library of gene-deletion mutants within the same strain, the Keio library, as a method of

defining essential genes (Baba et al. 2006; Goodall et al. 2018). After detailed analysis we were unable to resolve some of the discrepancies between the datasets. We hypothesised that by applying the TraDIS method to a gene-deletion library we would be able to rapidly validate the precise position of every insertion within the library and resolve some of the discrepancies observed. Thus, to test our hypothesis, we pooled the Keio collection of mutants and used the TraDIS methodology to identify the position of the selection cassette throughout the genome. We present this method, TraDIS-validate, as a tool for rapidly validating any ordered library.

Results

Identification of the kanamycin resistance cassette insertion sites within the Keio library

During construction of the Keio library, each annotated gene of *E. coli* BW25113 was targeted for replacement with a kanamycin resistance cassette. The intention was to create a single gene-deletion mutant with the start codon and seven terminal codons sandwiching the kanamycin resistance cassette (Fig. 1B; Datsenko and Wanner 2000). To identify the precise positions of the kanamycin resistance cassette, we pooled all the mutants of the Keio library that grew on LB agar plates at 37°C and sequenced the pool following an amended TraDIS protocol (Fig. 1). Genomic DNA was extracted from the pooled mutants and fragmented. Following end repair and adapter ligation, the sample was subjected to PCR amplification of the resistance cassette-genomic DNA (gDNA) junctions and the resulting amplicons were sequenced (Fig. 1 and Supplementary Fig. 1). We obtained ~1.6 million reads (accession no. PRJEB37909). Following confirmation and removal of the kanamycin resistance cassette sequence and removal of short or poor-quality reads, 881,943 reads were successfully mapped to the *E. coli* BW25113 reference genome (Table 1; accession no. CP009273.1). During data processing, the largest number of reads were discarded between the Ktag1 and Ktag2 pattern matching steps (Table 1): ‘Ktag1’ corresponds with the sequence of the K2F forward primer while ‘Ktag2’ corresponds with the sequence of the kanamycin resistance cassette immediately downstream from the K2F primer binding site (Supplementary Fig. 1). The disparity in the number of reads between these two steps is likely a result of mis-priming of the K2F forward primer and amplification of off-target DNA, however, data can also be lost at this step if there are single nucleotide polymorphisms (SNPs) in the cassette sequence resulting in barcode mismatch, discussed later. Following read processing, 881,943 mapped reads resulted in the identification of 4,176 kanamycin resistance cassette insertion loci. The mean number of reads

per insertion site was 211, while the median number of reads per insertion site was 177. To conservatively determine the presence of mutants within the library, we filtered our sequencing data to remove insertion-sites represented by 3 or fewer reads, leaving 4,054 mapped unique insertion sites. Using the *E. coli* BW25113 genome as a reference (CP009273.1; Grenier et al. 2014), 34 insertion sites were identified in intergenic regions and 4,020 insertion sites were identified within a protein coding sequence (CDS). Of the insertions within an annotated CDS, 3,718 insertions were 6 codons upstream from the terminal stop codon, 302 insertions were identified elsewhere within an annotated CDS.

We next compared our insertion data with the reported mutants of the Keio library. A gene was considered correctly disrupted if a cassette insertion site was identified 6 codons upstream from the first stop codon of the targeted open reading frame (ORF; Fig. 2A). The gene names and annotation boundaries used were derived from the *E. coli* BW25113 reference genome. Altogether we identified 3,718 genes with a correctly located kanamycin resistance cassette (Fig. 2C-D; Supplementary Table 1). We identified 148 genes with discrepancies. For 73 genes the annotation of the reference genome differs from their coordinates given within the Keio library and these were excluded from our analysis. These genes fall into three categories: (i) genes disrupted by insertion elements; (ii) genes that are encoded across more than one open reading frame; and (iii) genes that are annotated with more than one stop codon (Fig. 2B). We failed to identify an insertion site for 61 genes, 7 of which had an insertion site represented by ≤ 3 sequence reads and had therefore been filtered by our analysis (discussed later). We identified a nearby but inaccurate insertion site in 14 genes (Fig. 2C-D). We defined an inaccurate mutant as one where an insertion was identified within 3 codons of the correct position. Lastly, there were 23 instances where the identified insertion site does not coincide with an annotated CDS but upon closer inspection, the CDS is disrupted at the correct locus and the CDS annotation prematurely ends before a stop codon, resulting in the recorded insertion site appearing outside of the annotated CDS. These examples are considered to be correctly constructed mutants and the lack of identified CDS disruption is due to genome annotation error in the *E. coli* BW25113 reference genome (Supplementary Table 1).

There are an additional 151 genes annotated within the *E. coli* BW25113 reference genome sequence deposited in 2014 that were unannotated at the time of construction of the Keio library, and as such the corresponding mutants were not reported during the initial construction of the Keio library (Baba et al. 2006). However, we identified that six of the more recently

annotated genes (*ycaZ*, *yoaI*, *yegZ*, *ypdJ*, *ypjK* and *ldrD*) appear to be disrupted in a manner consistent with the Keio mutants (Baba et al. 2006). Surprisingly, we identified two genes (*yhhQ* and *tnaB*) that were reported to be essential in the initial Keio library paper but are correctly disrupted in our library. As these mutants have been constructed, and our previous TraDIS data supports the notion that these genes are not required for viability, we conclude that these genes are not essential (Goodall et al. 2018). Lastly, the disparity in numbers between total number of intragenic insertions detected and the total number of mutants can be attributed to the presence of 2 copies of each mutant: for some mutants we observed both a “correct” copy, and an “inaccurate” copy. A comprehensive list of the status of each mutant, with comments, can be found in Supplementary Table 1 and the processed insertion data can be viewed at our online browser: <https://tradis-vault.qfab.org/>.

Comparison of targeted and identified cassette insertion sites

We failed to identify the presence of a kanamycin resistance cassette within 61 genes; all of these genes were predicted to be non-essential by Baba et al (2006) and are thus expected to be present in the library. At the time of construction of the Keio library there was no sequenced genome available for the *E. coli* BW25113 strain, instead closely related strains *E. coli* MG1655 and W3110 were used as references for primer design (Supplementary Fig. 2A; Bachmann 1972; Blattner et al. 1997; Baba et al. 2006; Hayashi et al. 2006). Although the nucleotide sequences of these strains are highly similar, one possibility for the failure to detect a mutant is if localised regions of sequence disparity result in a failure to construct the mutant. To understand whether the missing mutants could be explained by primer mistargeting, we converted the 4,288 primer sequences used to make the mutants by Baba *et al.* (2006) into FASTQ format and mapped these to the *E. coli* BW25113 genome. The primer corresponding with the 3' end of the target gene and consistent with the end of the cassette we sequenced out from was used. Mapping the primer sequences in this way revealed the targeted insertion sites of the cassette (Fig. 3A), enabling a comparison between expected and identified insertion sites (Fig. 3B). Of the initial 4,288 primer sequences, 4,284 mapped successfully, corresponding with 4,276 unique insertion sites. The difference between the number of successfully mapped primers and identified insertion sites was due to more than one primer mapping to the same site. For example, the primer to construct the *stfR* mutant also has homology to the 3' end of *stfQ*. The four primer sequences (targeting *argI*, *eutA*, *lacI* and *ygcR*) that did not map to the BW25113 genome were due to differences in the genome annotation and/or sequence at the

loci targeted (Supplementary Material). However, despite minor sequence differences there was still sufficient sequence homology for the *argI*, *lacI* and *ygcR* mutants to be constructed (Supplementary Material). There was only one mutant, *eutA*, for which we did not identify an insertion in the initial screen and the primer to construct the mutant did not map to the *E. coli* BW25113 reference genome. The unique example of *eutA* is discussed later. Overall, the high success rate of primer mapping suggests the failure to identify the 61 mutants cannot be attributed to primer mistargeting.

However, mapping the mutant-construction primers did reveal incorrectly targeted mutants (Fig. 4). Four examples of this are presented below. The first of these is the *flu* mutant, for which the primer position is incorrect by one nucleotide, which is reflected in the constructed mutant. The Keio mutants were constructed in such a way that the kanamycin resistance cassette can be excised via recombination of the Flp recombination target (FRT) sites within the cassette (using Flp recombinase) to result in the native gene start codon fused in frame with a 34 amino acid scar and the native terminal 7 codons of the gene. In the example of the *flu* mutant, while the mis-localisation of the cassette by 1 nucleotide is relatively minor, this has the unintended consequence of creating an out-of-frame C-terminal fusion mutant when the FRT-sites are recombined, so that the native *flu* stop codon is no longer read by the ribosome (Fig. 4A). The same is applicable for all eight mutants where the cassette is out of frame by 1-2 nucleotides (Supplementary Table 1). For each of these examples, during excision of the cassette the resulting mutant will encode a fusion protein spliced from two reading frames. In this scenario, it is possible that synthesis of an artefact protein could have unintended polar effects. The most significant example of this is the *yedZ* mutant, as the excision of the kanamycin resistance cassette within this mutant will theoretically result in a fused peptide comprised of the *yedZ* start codon, the 34-residue cassette scar peptide and a 51 amino acid C-terminal peptide, resulting in an 86-residue artefact protein (Supplementary Fig. 3).

In some examples the primer maps to a substantially distant (>9 nucleotides) and incorrect position within the CDS. An example of this is *mltC*: the downstream reverse primer maps to a position in the middle of the CDS (Fig. 4B), which will result in only partial deletion of the *mltC* gene when the mutant is constructed. However, both the truncated *mltC* mutant and the correct *mltC* deletion mutant appear present within the Keio collection (Fig. 4B). Dense transposon mutagenesis screens have shown that provided there is translational read-through or read-out from the inserted cassette, disruptions to the 5' end of a CDS can still produce a

functional protein (Goodall et al. 2018). Partial gene deletions can be problematic if the CDS is functional and confers an alternative phenotype to the WT allele or complete gene deletion (Ma et al. 2011). The *mltC* gene encodes an OM-bound lytic murein transglycosylase (Dijkstra and Keck 1996). In the truncated Keio *mltC* mutant the transglycosylase portion of the coding sequence is still intact, however it is out of frame and therefore should not be expressed (Supplementary Material). There were at least 4 other instances where the original reported primers would result in construction of a truncated gene-deletion mutant, but both the incorrect and correct mutants are present within the glycerol stocks (Supplementary Table 1).

Conversely, the reverse primer to construct the *csiE* mutant is downstream from the correct target site (Fig. 4C). As a result, deletion of *csiE* will also truncate the converging neighbouring gene *hcaT*. While only ~8% of the *hcaT* gene would be deleted, partial deletion of a neighbouring gene can account for an observed phenotype (Bosl and Kersten 1994). Deletion of the 3' end of *hcaT* would result in the loss of the last (12th) transmembrane (TM) domain and disruption of the penultimate TM domain (Supplementary Material), which might affect the localisation, folding, stability, or function of the HcaT protein. A *csiE* mutant constructed with the incorrectly localised *csiE* primer site was identified within the pool (Fig. 4C, indicated by an asterisk), however an insertion site was also identified 6 codons upstream from the *csiE* stop codon (Fig. 4C). As such, both the correct and incorrect *csiE* mutants appear present within the library. Finally, the gene *rzoR* is nested within *rzpR*. However, the primers designed to construct each mutant are identical and do not correspond with the appropriate binding sites to correctly construct either mutant (Fig. 4D). It is impossible to distinguish between each respective mutant from the insertion data alone, however, a cassette was also identified 6 codons upstream from the *rzoR* stop codon.

Although these examples represent a very small proportion of the library, each of these mutants are representative of distinct scenarios where incorrect construction of the mutant might result in off target phenotypic effects. Furthermore, for these mutants, validation of the cassette position by a standard PCR size check would result in these errors going undetected. However, each of these examples can be identified by pooled amplicon sequencing, enabling rapid curation of an ordered library. When working with these Keio isolates, confirmation of the mutation by sequencing is advised.

Validation of Keio mutants by PCR and Sanger sequencing

There were 61 genes for which we did not detect a cassette insertion site. First, we re-evaluated our insertion data. For 7 of these genes (*rnt*, *pykF*, *guaA*, *rlmE*, *rtcA*, *atpC* and *yccA*) an insertion site was identified in the correct location, however there were ≤ 3 reads that matched the insertion site and therefore these did not meet our stringency threshold in our initial analysis. Two of these genes, *rnt* and *guaA*, were reported to be essential within a dense transposon library dataset (Goodall et al. 2018). As such, the validity of these mutants is inconclusive as the insertion site was in the correct location but with significantly fewer reads than the median number (177). One possible explanation for this is that the glycerol stock of each of these mutants is a mixed population and the correct mutant is in low abundance, and the sequencing method is sensitive enough to identify the correct mutant amongst the noise. Alternate explanations for a disproportionately low read count include if there was a plating error or if mutants have a slow growth phenotype and form pin-prick colonies. However, we revisited the raw data, this time extracting the sequence data by pattern matching for the expected gene rather than the kanamycin cassette. We identified mutations within the kanamycin cassette sequence upstream of four genes: *rtcA*, *atpC*, *pykF* and *rlmE*, resulting in these data being filtered during the kanamycin cassette pattern matching steps ('ktag1' and 'ktag2').

Of the remaining 54 genes where no insertion was detected, with the exception of *eutA*, the primer to construct each of these mutants correctly maps to the expected location within the reference genome. Therefore, to check whether these isolates are indeed incorrect, one copy of each mutant was checked by PCR to map the genomic position of the kanamycin resistance cassette. Briefly, a single colony was isolated and the cassette-gDNA junction was amplified using a primer specific for the kanamycin resistance cassette and a reverse degenerate primer, using a protocol adapted from Levano-Garcia *et al.* (2005). The amplified product was sequenced using Sanger sequencing and a nested primer specific for the cassette. We were unable to check nine isolates as our lab stock of the Keio collection was purchased in 2007 and only includes plates 1-90; more recent versions of the library include plates 91-94. Another explanation for the failure to identify an insertion site is if the mutant failed to grow. In our initial screen, we only identified 4 mutants that did not grow (Supplementary Table 2), however, during purification for colony PCR, one mutant (*araE*) did not grow upon re-streaking and we propose this isolate was missed in the initial TraDIS-validate screen. For one isolate, *gatA*, although the gene was targeted for deletion and the primer for construction of the

mutant maps to the correct location, we did not identify this mutant in our screen. On closer inspection, this gene is not reported as essential in the original paper, however we could not find a record of the *gatA* mutant within 3 independent databases that describe the well positions of each mutant, despite searching for the gene name ('*gatA*'), the ECK number (ECK2087), or b number (b2094). We therefore conclude that although this gene was targeted for deletion, this mutant is not present within the Keio library. Of the 43 isolates we were able to validate, 10 were correctly disrupted but due to nucleotide differences in either the kanamycin resistance cassette or the genome, reads for these data were either binned during processing or did not map to the reference genome (Supplementary Table 3; Supplementary Material). Of note, when the position of the kanamycin resistance cassette within the *eutA* mutant was validated by PCR and Sanger sequencing, the cassette was found in the correct locus and is immediately upstream of *eutB*. Both *E. coli* MG1655 and the sequenced *E. coli* BW25113 genome contain a CPZ-55 cryptic prophage inserted into the ethanolamine biosynthesis operon between genes *eutA* and *eutB* (Grenier et al. 2014), which is absent in *E. coli* W3110. However, the primer to construct the *eutA* mutant has homology for the *eutAB* junction, not the neighbouring prophage reported for *E. coli* BW25113 (Supplementary Fig. 2B). As such, there is no prophage within the ethanolamine biosynthesis operon in the Keio BW25113 *eutA* mutant; this localised sequence difference explains why both the primer and the sequence data did not map to the reference genome.

We identified the kanamycin resistance cassette elsewhere in the genome for the remaining 33 isolates. Three of these (*ybeY*, *dnaT* and *thyA*) were reported to be essential by TraDIS analysis (Goodall et al. 2018), therefore for these isolates we checked the position of the kanamycin resistance cassette in the second copy of each mutant. The second copy of *ybeY* did not grow, while the cassette was also incorrect in the additional *dnaT* and *thyA* isolates. We propose that these three genes are essential (Supplementary Table 3).

For the 33 isolates where the kanamycin resistance cassette was identified elsewhere in the genome, we noted that the cassette was not randomly distributed throughout the genome, but appeared 6 codons before a stop codon, consistent with Keio mutants. To determine whether these mutants might be cross-contaminants from neighbouring wells, we cross-referenced the identified mutant with neighbouring mutants within the same plate. We found that 25 of the mis-identified mutants corresponded with a mutant in a neighbouring well, 4 of the mutants corresponded with a mutant elsewhere within the same plate, while the kanamycin resistance

cassette insertion site of 4 mutants (*ypdI*, *yiaD*, *thyA* and *oxyR*) correspond with mutants elsewhere in the library (Supplementary Table 3). Initially, we were concerned that the wrong isolates may have been PCR-checked in error, but our 96-well plates are sealed with foil and when we conducted the validation checks the seals were not removed, instead the foil was pierced to access the stock below and patched over. We checked the tested wells and confirmed these matched the intended positions. It is therefore likely that these wells were either incorrectly stocked or have been contaminated during replication of the library.

At least one of the mutants we identified as incorrect (*oxyR*) has been reported previously as an impure mutant (Tamae et al. 2008; Liu et al. 2010). To test whether the incorrect mutants in our library are representative of the Keio mutants in other labs, we reviewed the phenotypic profiles of the incorrect mutants using a publicly available phenotypic database for the Keio library (Nichols et al. 2011). The work by Nichols et al. (2011) screened the Keio library under a range of different stresses and recorded the resulting colony sizes of each mutant. Using this data, they were able to assign quantitative fitness scores for each gene under each condition. This data can be used to assign a function to genes of unknown function as genes involved in the same process or pathway have similar phenotypes and therefore have a high correlation score. Therefore, we compared the phenotypic profile of each incorrect mutant with its positive counterpart, for example, the insertion site for the '*hns*' mutant was identified within *hupA* (Supplementary Table 3), as such, you might expect the '*hns*' mutant to behave as the *hupA* mutant. Although we only validated one copy of each mutant clone by PCR, the data presented by Nichols et al. (2011) is a combined dataset that merges data from both mutant clones; entries with a low correlation coefficient between clones were discarded, therefore we assume that the clones are comparable for all remaining mutants. However, phenotypic comparison revealed that none of the expected mutants correlated highly (>0.6) with the identified mutants (Supplementary Fig. 4), and only 2 comparisons (*pgaD* with *ycfL*, 0.56; and *kch* with *oppD*, 0.46), had a correlation coefficient greater than 0.4 suggesting the identified contaminations are specific to our library.

Discussion

The construction of the Keio library was a laborious undertaking and an impressive feat, made all the more remarkable by the fact that despite the discrepancies we identified here and previously, these are very few in number relative to the whole library. Nevertheless, in specific

examples highlighted here, some issues still remain. Altogether, our method only missed 10 correctly constructed mutants out of a total of 3,728 confirmed isolates. This was a consequence of nucleotide polymorphisms, insertions or deletions within the kanamycin cassette or the genomic DNA of the mutant resulting in the reads for these data being discarded during processing. TraDIS-validate therefore had a success rate of 99.7% for identifying the correct kanamycin cassette position. This dataset provides a benchmark for the purity of the Keio mutants and a screening method for mapping the position of an antibiotic resistance cassette in any ordered library. A limitation of this method is that it only informs whether a mutant is present within the total pool of mutants, the nature of the method cannot distinguish whether the mutant of a given well is correct; further validation checks are required to test this. One caveat to consider is that the presence of a correctly positioned kanamycin resistance cassette is not sufficient evidence for the correct construction of a mutant, as is seen with *yciM*, *yhcB*, *hda* and *holD* mutants (Mahalakshmi et al. 2014; Goodall et al. 2021). In each of these examples, the kanamycin cassette is in the correct location, but secondary mutations have been reported elsewhere in the genome. Similarly, Yamamoto *et al.* (2009) have reported gene duplication events in a handful of Keio mutants. As such, when working with individual isolates from any ordered library it is recommended to check the mutant by PCR and Sanger sequencing, and when found to be correct it should be transduced into a clean background before use (Held et al. 2012).

A noteworthy point during examination of the data was the identification of 73 genes with annotation differences between the *E. coli* BW25113 reference genome and the coordinates used for the initial construction of the Keio library. This highlights the importance of the annotation quality of the reference genome for high-throughput whole-genome screens, but also raises the question: how best to annotate a genome? Within the *E. coli* BW25113 reference genome annotation (CP009273.1), pseudogenes are annotated with the CDS extending beyond premature stop codons. Arguably, this is an informative representation of the gene as this displays the gene alterations that led to the classification of ‘pseudogene’, however, in some cases the truncated CDSs are still expressed and functional, and therefore perhaps not pseudogenes but observed evolution. Additionally, for simplicity the complexity of the genome is not always reflected in standard genome annotation. For example, the gene *dnaX* encodes 2 subunits of DNA polymerase III: τ and γ expressed from the same locus in equal proportions; subunit τ is expressed by standard translation while γ is expressed when a -1 nt frameshift

occurs at the A_AAA_AAG consensus sequence (Blinkowa and Walker 1990; Flower and McHenry 1990; Tsuchihashi and Kornberg 1990; Gurvich et al. 2003). The capacity to encode 2 different protein subunits depending upon ribosomal processivity is not evident from standard gene annotation. This is pertinent for the Keio collection as some coding sequences in the *E. coli* BW25113 genome have been annotated across multiple coding sequences, such as *arpB*, which contains the same frameshift consensus sequence as *dnaX* but, unlike *dnaX*, the *arpB* gene was deleted in two parts. Primer design is therefore of considerable importance when constructing gene-replacement libraries. We highlighted some specific examples within the Keio library for consideration and demonstrate the utility of our method for identifying these errors.

Finally, we have been able to resolve some of the discrepancies raised in earlier analysis of a TraDIS library in the *E. coli* strain BW25113 (Goodall et al. 2018). We observed mutations in two genes, *yhhQ* and *tnaB*, previously reported as essential, and conclude that both are in fact non-essential, which is consistent with the TraDIS data. Secondly, we identified that the *ybeY*, *thyA* and *dnaT* mutants, reported as essential in the TraDIS screen, are not stocked within the Keio library and therefore likely essential. Conversely, we validated the presence of several mutants in the Keio collection that the TraDIS data predicted to be essential (Supplementary Table 1). There are a number of explanations for the false-positive prediction of essential genes in transposon mutagenesis screens, as has been noted by us and others (Goodall et al. 2018; Cain et al. 2020). As a metric, insertion frequency can be influenced by other biological factors besides gene essentiality. Notwithstanding, transposon insertion sequencing screen provides data rich information that can be enhanced with the development of more refined analysis methods, in particular the use of machine learning methods. Our data highlight that care needs to be taken with any high-throughput screen to understand the limitations of the library construction method, the validity of the library and the limitations of the analyses used; especially with ever increasing high-throughput screens used to assign biological function to genes. We conclude TraDIS-validate is a quick and robust method for curating ordered gene libraries to identify errors ahead of often costly and time-intensive screens.

Materials and Methods

Strains and plasmids

E. coli K-12 strain BW25113, the parent strain of the Keio library, was used for construction of a transposon library. The strain has the following genotype: *rrnB3* Δ *lacZ4787* *hsdR514* Δ (*araBAD*)567 Δ (*rhaBAD*)568 *rph-1* (Datsenko and Wanner 2000).

Culture of bacterial isolates and determination of growth kinetics

Keio collection isolates were cultured on Lysogeny broth (LB; 10 g tryptone, 5 g yeast extract, 10 g NaCl) supplemented with kanamycin sulfate, 50 μ g/mL (Fluka Biochemika, Switzerland) and Agar Bacteriological No. 1 (Oxoid, Basingstoke, UK), at 37°C.

Amplicon sequencing

Keio library isolates were cultured using a multiple point inoculator to inoculate large square plates of solid LB medium and incubated overnight for 18 h at 37°C. Colonies were scraped from LB plates and pooled. Genomic DNA was extracted using a Qiagen QIAamp DNA Blood Mini Kit, according to the manufacturer's specifications. DNA was fragmented using an ultrasonicator (Diagenode, Bioruptor® Plus) with the following profile: 13 cycles of 30 s ON, 90 s OFF at low intensity, to an average fragment size of ~300 bp. The kanamycin resistance cassette-gDNA junctions were prepared for sequencing using NEB Ultra I kit (New England Biolabs). Following adapter ligation and treatment with the NEB USER Enzyme, the sample was purified using AMPure XP beads (Beckman Coulter) to remove adapter dimer, and select fragments around ~250 bp, the final sample was eluted in 17 μ l EB buffer, 15 μ l of which was transferred to a microcentrifuge tube. Next, cassette-gDNA junctions were enriched by PCR using a forward primer with homology to the kanamycin resistance cassette (K1F 5'-TCGCCTTCTTGACGAGTTCTTCTAATAAGG-3'), and a reverse primer with homology to the ligated adapter (A1R 5'-GACTGGAGTTCAGACGTGTGCTCTTCCGATC-3'). The PCR reaction contained 25 μ l 2x Q5 DNA Polymerase (NEB), 2.5 μ l K1F (10 μ M), 2.5 μ l A1R (10 μ M), 15 μ l sample and 5 μ l nuclease-free water and was amplified using the following cycle conditions: 98°C, 48 s; (98°C, 15 s; 65°C, 30 s; 72°C, 30 s) x10; 72°C, 1 min; 4°C hold. Samples were purified using AMPure XP beads at a ratio of 1:0.9 respectively. Finally, samples were prepared for sequencing using a custom forward primer K2F (5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGA TCTNNNNNNAAAGTATAGGAAGCTTCGAAGCAGCT-3'), which contains the Illumina P5 flow cell and adapter sequences and a custom barcode denoted by 'N'. The barcodes varied in length from 6-9 bp and were unique to each primer. The purpose of the barcode is to add

nucleotide diversity during the early cycles of sequencing and stagger the introduction of the kanamycin resistance cassette. Without the barcode, sequencing the cassette-gDNA junction on an Illumina platform would result in 100% nucleotide identity during the early sequencing cycles (corresponding with the uniform cassette nucleotide sequence), and would result in the data being classified as ‘overclustered’ and subsequently binned. The reverse primers for this reaction were the NEBNext® Multiplex Oligos for Illumina (NEB; Supplementary information; Supplementary Fig. 1). The PCR reaction contained 25 µl 2x Q5 DNA Polymerase (NEB), 2.5 µl K2.F (10 µM), 2.5 µl NEBNext®_oligo.R (10 µM), 15 µl sample and 5 µl nuclease-free water and was amplified using the following cycle conditions: 98°C, 48 s; (98°C, 15 s; 65°C, 30 s; 72°C, 30 s) x20; 72°C, 1 min; 4°C hold. Samples were quantified by qPCR using a Library Quantification kit for Illumina platforms (Kapa) and sequenced using an Illumina MiSeq. Data are available from the European Nucleotide Archive accession PRJEB37909.

Sequencing analysis

The Fastx barcode splitter and trimmer tools, of the Fastx toolkit, were used to assess and trim the sequences (Pearson et al. 1997). Sequence reads were first filtered by their sample barcodes, allowing zero mismatches. Similarity matching of the cassette was done by identifying the first 35 bp of the kanamycin resistance cassette in two parts: 25 bases (5' to 3', corresponding with the PCR2 primer binding site, termed ‘Ktag1’) were matched allowing for 3 mismatches, trimmed, and then the remaining 10 bases (corresponding with sequenced cassette, termed ‘Ktag2’) matched allowing for 1 mismatch, and trimmed (Supplementary Fig. 1). Sequences less than 20 bases long were removed using Trimmomatic (Bolger et al. 2014). Trimmed, filtered sequences were then aligned to the reference genome *E. coli* BW25113, obtained from the NCBI genome repository (CP009273.1). The aligner bwa was used, with the mem algorithm (Li and Durbin 2009). Samtools was used for the subsequent conversion from sam to bam files, and the sorting and indexing of data (Li et al. 2009). The bedtools suite was used to create bed files, which were intersected against the BW25113 reference genome. Data were inspected manually using the Artemis genome browser, after processing for visualisation using custom python scripts (Rutherford et al. 2000). The data can also be visualised in our online browser, available at <https://tradis-vault.qfab.org/apollo/jbrowse/>.

PCR validation of Keio mutants

A single colony of each mutant was resuspended in 100 µl water and boiled at 98°C for 10 min. Cell debris was pelleted by centrifugation and 5 µl of supernatant was used as a template for PCR. The cassette-genomic DNA junction was amplified using a primer specific for the kanamycin resistance cassette (Kan-down: 5'-CTATCGCCTTCTTGACGAGTTC-3') and a reverse degenerate primer (HIB17: 5'-CGGAATTCCGGATNGAYKSNGGNTC-3') that non-specifically binds to the genome (Levano-Garcia et al. 2005). A 'touchdown' thermal profile was used to maximise primer annealing, using the following thermal profile: 95°C for 5 min; then 25 cycles: 45 s at 95°C; 45 s at 60°C, decreasing by 1°C each cycle; 2 min at 72°C; followed by 25 cycles: 45 s at 95°C; 45 s at 50°C; 2 minutes at 72°C; and then held at 4°C. Following purification, the amplified product was sequenced using Sanger sequencing and a nested primer specific for the kanamycin resistance cassette (Kan-down-seq: 5'-CTAATAAGGGGATCTTGAAGTTC-3').

Data and Software Availability

The datasets produced in this study are available at the following databases:

- Sequence data: ENA accession: PRJEB37909 (<https://www.ebi.ac.uk/ena>)

The data can also be visualised in our online browser available at <https://tradis-vault.qfab.org/apollo/jbrowse/>

Acknowledgements

We thank Professor Mori for his helpful tables and insight into the Keio library.

Author contributions

The project was devised by FCM and IRH. SAM extracted the Keio isolates and prepared the samples for sequencing; ES sequenced the samples using the TraDIS-like protocol. ECAG processed and analysed the sequencing data. JAB, RS; IAW; CI and GB validated mutants by PCR. Phenotypic analysis of mutants was done by MB. The manuscript was written and edited by ECAG, FCM, AC, JAC, MB, JAB, and IRH.

Conflict of interest

The authors declare that they have no conflict of interest.

References

- Baba, T., T. Ara, M. Hasegawa, Y. Takai, Y. Okumura, M. Baba, K. a Datsenko, M. Tomita, et al. 2006. Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Molecular systems biology* 2: 2006.0008. doi:10.1038/msb4100050.
- Bachmann, B. J. 1972. Pedigrees of some mutant strains of *Escherichia coli* K-12. *Bacteriological reviews* 36. American Society for Microbiology (ASM): 525–57.
- Baym, M., L. Shaket, I. A. Anzai, O. Adesina, and B. Barstow. 2016. Rapid construction of a whole-genome transposon insertion collection for *Shewanella oneidensis* by Knockout Sudoku. *Nature Communications* 7. Nature Publishing Group. doi:10.1038/ncomms13270.
- De Berardinis, V., D. Vallenet, V. Castelli, M. Besnard, A. S. Pinet, C. Cruaud, S. Samair, C. Lechaplais, et al. 2008. A complete collection of single-gene deletion mutants of *Acinetobacter baylyi* ADP1. *Molecular Systems Biology*. doi:10.1038/msb.2008.10.
- Blattner, F. R., G. Plunkett, C. A. Bloch, N. T. Perna, V. Burland, M. Riley, J. Collado-Vides, J. D. Glasner, et al. 1997. The complete genome sequence of *Escherichia coli* K-12. *Science (New York, N.Y.)* 277: 1453–62.
- Blinkowa, A. L., and J. R. Walker. 1990. Programmed ribosomal frameshifting generates the *Escherichia coli* DNA polymerase III γ subunit from within the τ subunit reading frame. *Nucleic Acids Research* 18: 1725–1729. doi:10.1093/nar/18.7.1725.
- Bolger, A. M., M. Lohse, and B. Usadel. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30: 2114–2120. doi:10.1093/bioinformatics/btu170.
- Bosl, M., and H. Kersten. 1994. Organization and functions of genes in the upstream region of *tyrT* of *Escherichia coli*: Phenotypes of mutants with partial deletion of a new gene (*tgs*). *Journal of Bacteriology* 176. American Society for Microbiology: 221–231. doi:10.1128/jb.176.1.221-231.1994.
- Cain, A. K., L. Barquist, A. L. Goodman, I. T. Paulsen, J. Parkhill, and T. van Opijnen. 2020. A decade of advances in transposon-insertion sequencing. *Nature Reviews Genetics*. Nature Research. doi:10.1038/s41576-020-0244-x.
- Cameron, D. E., J. M. Urbach, and J. J. Mekalanos. 2008. A defined transposon mutant library and its use in identifying motility genes in *Vibrio cholerae*. *Proceedings of the National Academy of Sciences of the United States of America* 105: 8736–8741. doi:10.1073/pnas.0803281105.
- Chaudhuri, R. R., E. Morgan, S. E. Peters, S. J. Pleasance, D. L. Hudson, H. M. Davies, J. Wang, P. M. van Diemen, et al. 2013. Comprehensive Assignment of Roles for

- Salmonella Typhimurium Genes in Intestinal Colonization of Food-Producing Animals.
Edited by Diarmaid Hughes. *PLoS Genetics* 9: e1003456.
doi:10.1371/journal.pgen.1003456.
- Datsenko, K. A., and B. L. Wanner. 2000. One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proceedings of the National Academy of Sciences of the United States of America* 97: 6640–5. doi:10.1073/pnas.120163297.
- Dijkstra, A. J., and W. Keck. 1996. Identification of New Members of the Lytic Transglycosylase Family in *Haemophilus influenzae* and *Escherichia coli*. *Microbial Drug Resistance* 2: 141–145. doi:10.1089/mdr.1996.2.141.
- Enstrom, M., K. Held, B. Ramage, M. Brittnacher, L. Gallagher, and C. Manoil. 2012. Genotype-phenotype associations in a nonmodel prokaryote. *mBio* 3. doi:10.1128/mBio.00001-12.
- Flower, A. M., and C. S. McHenry. 1990. The γ subunit of DNA polymerase III holoenzyme of *Escherichia coli* is produced by ribosomal frameshifting. *Proceedings of the National Academy of Sciences of the United States of America* 87: 3713–3717. doi:10.1073/pnas.87.10.3713.
- Gallagher, L. A., E. Ramage, M. A. Jacobs, R. Kaul, M. Brittnacher, and C. Manoil. 2007. A comprehensive transposon mutant library of *Francisella novicida*, a bioweapon surrogate. *Proceedings of the National Academy of Sciences of the United States of America* 104: 1009–1014. doi:10.1073/pnas.0606713104.
- Giaever, G., A. M. Chu, L. Ni, C. Connelly, L. Riles, S. Véronneau, S. Dow, A. Lucau-Danila, et al. 2002. Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* 418: 387–391. doi:10.1038/nature00935.
- Goodall, E. C. A., A. Robinson, I. G. Johnston, S. Jabbari, K. A. Turner, A. F. Cunningham, P. A. Lund, J. A. Cole, et al. 2018. The Essential Genome of *Escherichia coli* K-12. Edited by Swaine L. Chen. *mBio* 9: e02096-17. doi:10.1128/mBio.02096-17.
- Goodall, E. C. A., G. L. Isom, J. L. Rooke, K. Pullela, C. Icke, Z. Yang, G. Boelter, A. Jones, et al. 2021. Loss of YhcB results in dysregulation of coordinated peptidoglycan, LPS and phospholipid synthesis during *Escherichia coli* cell growth. *PLoS Genetics* 17. Public Library of Science: e1009586. doi:10.1371/JOURNAL.PGEN.1009586.
- Gray, A. N., B. M. Koo, A. L. Shiver, J. M. Peters, H. Osadnik, and C. A. Gross. 2015. High-throughput bacterial functional genomics in the sequencing era. *Current Opinion in Microbiology*. Elsevier Ltd. doi:10.1016/j.mib.2015.07.012.
- Grenier, F., D. Matteau, V. Baby, and S. Rodrigue. 2014. Complete Genome Sequence of

Escherichia coli BW25113. *Genome announcements* 2. American Society for Microbiology (ASM). doi:10.1128/genomeA.01038-14.

Gurvich, O. L., P. V Baranov, J. Zhou, A. W. Hammer, R. F. Gesteland, and J. F. Atkins. 2003. Sequences that direct significant levels of frameshifting are frequent in coding regions of Escherichia coli. *The EMBO journal* 22. European Molecular Biology Organization: 5941–50. doi:10.1093/emboj/cdg561.

Hayashi, K., N. Morooka, Y. Yamamoto, K. Fujita, K. Isono, S. Choi, E. Ohtsubo, T. Baba, et al. 2006. Highly accurate genome sequences of Escherichia coli K-12 strains MG1655 and W3110. *Molecular systems biology* 2. NATURE PUBLISHING GROUP, 75 VARICK ST, 9TH FLR, NEW YORK, NY 10013-1917 USA: 2006.0007. doi:10.1038/msb4100049.

Held, K., E. Ramage, M. Jacobs, L. Gallagher, and C. Manoil. 2012. Sequence-verified two-allele transposon mutant library for Pseudomonas aeruginosa PAO1. *Journal of Bacteriology* 194: 6387–6389. doi:10.1128/JB.01479-12.

Jacobs, M. A., A. Alwood, I. Thaipisuttikul, D. Spencer, E. Haugen, S. Ernst, O. Will, R. Kaul, et al. 2003. Comprehensive transposon mutant library of Pseudomonas aeruginosa. *Proceedings of the National Academy of Sciences of the United States of America* 100: 14339–14344. doi:10.1073/pnas.2036282100.

Kobayashi, K., S. D. Ehrlich, A. Albertini, G. Amati, K. K. Andersen, M. Arnaud, K. Asai, S. Ashikaga, et al. 2003. Essential Bacillus subtilis genes. *Proceedings of the National Academy of Sciences of the United States of America* 100: 4678–4683. doi:10.1073/pnas.0730515100.

Lane, J. M., and E. J. Rubin. 2006. Scaling down: A PCR-based method to efficiently screen for desired knockouts in a high density Mycobacterium tuberculosis picked mutant library. *Tuberculosis* 86: 310–313. doi:10.1016/j.tube.2006.01.015.

Levano-Garcia, J., S. Verjovski-Almeida, and A. C. R. da Silva. 2005. Mapping transposon insertion sites by touchdown PCR and hybrid degenerate primers. *BioTechniques* 38. Future Science Ltd London, UK : 225–229. doi:10.2144/05382ST03.

Li, H., and R. Durbin. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25: 1754–1760. doi:10.1093/bioinformatics/btp324.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, et al. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25: 2078–2079. doi:10.1093/bioinformatics/btp352.

Liberati, N. T., J. M. Urbach, S. Miyata, D. G. Lee, E. Drenkard, G. Wu, J. Villanueva, T.

- Wei, et al. 2006. An ordered, nonredundant library of *Pseudomonas aeruginosa* strain PA14 transposon insertion mutants. *Proceedings of the National Academy of Sciences of the United States of America* 103. National Academy of Sciences: 2833–2838. doi:10.1073/pnas.0511100103.
- Liu, A., L. Tran, E. Becket, K. Lee, L. Chinn, E. Park, K. Tran, and J. H. Miller. 2010. Antibiotic Sensitivity Profiles Determined with an *Escherichia coli* Gene Knockout Collection: Generating an Antibiotic Bar Code. *Antimicrobial Agents and Chemotherapy* 54. American Society for Microbiology: 1393–1403. doi:10.1128/AAC.00906-09.
- Ma, C., H. Zhou, J. Li, J. Dai, W. He, H. Wang, L. Wu, and Y. Wang. 2011. Construction of 4"-isovalerylspiramycin-I-producing strain by in-frame partial deletion of 3-O-acyltransferase gene in streptomyces spiramyceticus WSJ-1, the bitespiramycin producer. *Current Microbiology* 62: 16–20. doi:10.1007/s00284-010-9664-8.
- Mahalakshmi, S, M. R. Sunayana, L. SaiSree, and M. Reddy. 2014. yciM is an essential gene required for regulation of lipopolysaccharide synthesis in *Escherichia coli*. *Molecular microbiology* 91: 145–57. doi:10.1111/mmi.12452.
- Mahalakshmi, S., M. R. Sunayana, L. SaiSree, and M. Reddy. 2014. yciM is an essential gene required for regulation of lipopolysaccharide synthesis in *Escherichia coli*. *Molecular Microbiology* 91: 145–157. doi:10.1111/mmi.12452.
- Mesarich, C. H., J. Rees-George, P. P. Gardner, F. A. Ghomi, M. L. Gerth, M. T. Andersen, E. H. A. Rikkerink, P. C. Fineran, et al. 2017. Transposon insertion libraries for the characterization of mutants from the kiwifruit pathogen *Pseudomonas syringae* pv. actinidiae. Edited by Mikael Skurnik. *PLOS ONE* 12: e0172790. doi:10.1371/journal.pone.0172790.
- Nichols, R. J., S. Sen, Y. J. Choo, P. Beltrao, M. Zietek, R. Chaba, S. Lee, K. M. Kazmierczak, et al. 2011. Phenotypic landscape of a bacterial cell. *Cell* 144: 143–56. doi:10.1016/j.cell.2010.11.052.
- Pearson, W. R., T. Wood, Z. Zhang, and W. Miller. 1997. Comparison of DNA Sequences with Protein Sequences. *Genomics* 46: 24–36. doi:10.1006/geno.1997.4995.
- Porwollik, S., C. A. Santiviago, P. Cheng, F. Long, P. Desai, J. Fredlund, S. Srikumar, C. A. Silva, et al. 2014. Defined Single-Gene and Multi-Gene Deletion Mutant Collections in *Salmonella enterica* sv Typhimurium. Edited by Michael Hensel. *PLoS ONE* 9: e99820. doi:10.1371/journal.pone.0099820.
- Rutherford, K., J. Parkhill, J. Crook, T. Horsnell, P. Rice, M. A. Rajandream, and B. Barrell.

2000. Artemis: sequence visualization and annotation. *Bioinformatics (Oxford, England)* 16: 944–5. doi:10.1093/BIOINFORMATICS/16.10.944.

Tamae, C., A. Liu, K. Kim, D. Sitz, J. Hong, E. Becket, A. Bui, P. Solaimani, et al. 2008. Determination of antibiotic hypersensitivity among 4,000 single-gene-knockout mutants of *Escherichia coli*. *Journal of bacteriology* 190. American Society for Microbiology Journals: 5981–8. doi:10.1128/JB.01982-07.

Tsuchihashi, Z., and A. Kornberg. 1990. Translational frameshifting generates the γ subunit of DNA polymerase III holoenzyme. *Proceedings of the National Academy of Sciences of the United States of America* 87: 2516–2520. doi:10.1073/pnas.87.7.2516.

Yamamoto, N., K. Nakahigashi, T. Nakamichi, M. Yoshino, Y. Takai, Y. Touda, A. Furubayashi, S. Kinjo, et al. 2009. Update on the Keio collection of *Escherichia coli* single-gene deletion mutants. *Molecular systems biology* 5: 335. doi:10.1038/msb.2009.92.

Tables

Table 1. Number of reads surviving each stage of processing

Number of barcoded reads	K tag 1 match	K tag 2 match	Reads > 20 bp and passing QC filter	Total mapped reads
1,670,483	1,629,609 (97.5%)	1,076,508 (66.06%)	1,057,274 (98.21%)	884,545 (83.67%)

Figures

Figure 1. Sequencing the Keio library

(A) Mutants of the Keio library are stocked in 96-well plates. Each 96-well plate was inoculated onto an LB agar plate and cultures incubated overnight. Colonies were pooled and genomic DNA was extracted from the pool of mutants. (B) The Keio library of mutants was constructed by individually targeting each gene for replacement by the kanamycin resistance cassette amplified from pKD13. The resulting mutant retained a start codon and its terminal 7 codons (inclusive of a stop codon) with the central coding sequence replaced by the resistance cassette. To sequence the Keio library, genomic DNA was extracted from the pooled mutants and fragmented. An adapter was ligated to the fragmented ends, and the cassette-gDNA junction was amplified using primers specific for the 3' end of the kanamycin phosphotransferase gene and the adapter. (C) Output FASTQ data was mapped to a reference genome, and the first coordinate of each mapped read was counted as the kanamycin resistance cassette insertion site. Abbreviations: FRT, Flp recognition target.

Figure 2. Identification of the kanamycin resistance cassette insertion sites

(A) Schematic of the correctly located insertion site of a Keio mutant. (B) Examples of genes encoded across >1 reading frame, which were excluded from our analysis. (C) Decision chart for classifying mutants. Caveats: ¹Genes encoded across >1 reading frame were discounted from our analysis (73). ²The gene could be essential (296), the gene could be a recent annotation and was not targeted for deletion (151), an insertion was detected but had ≤ 3 reads and was filtered from our analysis (7), or no insertion was detected for an expected mutant (54). ³Reference genome annotation is incomplete, but the mutant appears correct (23). (D) The presence and status of the corresponding Keio mutant for every annotated gene in BW25113. The kanamycin resistance cassette was correctly positioned within the majority of genes (3,718 genes; grey). For a subset of genes, there were discrepancies with the identified location of the cassette and the gene annotation (148 genes; magenta). There are 296 annotated genes in the BW25113 reference genome reported to be essential during construction of the Keio library (teal) that had no insertion, and 151 genes annotated in the BW25113 reference genome that were not targeted for deletion during construction of the Keio library (purple).

Figure 3. Comparison of expected and constructed mutants

(A) Primers for construction of a Keio mutant; mapping of the reverse primer to the reference genome reveals the expected insertion site (red, above). (B) Comparison of expected insertion sites (above; red) with identified insertion sites (below; blue). (i) A correctly disrupted gene; (ii) A gene reported to be deleted but with no identified insertion; (iii) a gene reported as essential as the mutant could not be constructed; (iv) a gene reported as essential but with a correctly positioned resistance cassette.

Figure 4. Identification of the mapped positions of primers used to construct specific Keio mutants

(A-D) Construction of the *flu*, *mltC*, *csiE*, and *rzoR* & *rzpR* gene deletion mutants using the primers reported in the original Keio library paper. Incorrectly positioned primers are shown as arrows in magenta, mapped primer sites are shown as red spikes above the gene track, identified insertion sites are shown in blue below. (A) Recombination of the FRT sites within the *flu* mutant will result in an artefact peptide composed of the *flu* start codon, the 34-residue scar left by the cassette, and residues of the in-frame reading frame (magenta). (B) (i) Primers to construct the *mltC* mutant map to the incorrect location (magenta), construction of this mutant would maintain the C-terminal portion of the *mltC* CDS. (ii) Two cassette-gDNA junctions were identified within the *mltC* gene (blue). (C) (i) Primers to construct the *csiE* mutant map to the incorrect location (magenta), construction of this mutant would delete *csiE* and the 3' end of the neighbouring gene *hcaT*. (ii) A cassette-gDNA junctions corresponding with the mis-localised primer position (*) was identified within the *hcaT* gene (blue), in addition to a correctly position cassette within the *csiE* gene (blue). (D) Primers to construct the *rzoR* and *rzpR* mutants map to the same, incorrect location (magenta). (ii) Sequence data for these loci reveal 2 mapped primer positions (red), and 3 identified cassette insertion sites (blue).

734 **Supplementary Tables**

735 **Supplementary Table 1. Gene insertion mutant status**

736 **Supplementary Table 2. Mutants that failed to grow**

737 **Supplementary Table 3. Identification of the kanamycin resistant cassette insertion site by**

738 **PCR**

739