Title

Abstract

1 Introduction

Generating genomic variants that carry genes responsible for essential cellular processes can open up new research directions. Adding genes or metabolic pathways to cells provides new variants with specialised phenotypes. Modified cells have different potential applications in biotechnology [Juhas et al., 2014], fuel production [Seo et al., 2013], healthcare, and food production [Juhas et al., 2013]. Another important application for studying essential genes is in drug discovery [Juhas et al., 2012]. Infectious diseases are among the top major causes of mortality worldwide. Even though antibiotic resistance is growing among bacteria, the antibiotic discovery and development rate is diminishing [Fischbach and Walsh, 2009, Nathan, 2004]. Therefore, it is urgent to find new drugs for infectious diseases. New antibiotics target genes that are essential for the survival of pathogenic bacteria in order to control disease. Some well-known antibiotics that target essential functions are tetracyclines that bind to small ribosomal subunit and interfere with protein translation [Brodersen et al., 2000], penicillins that target peptidoglycan and inhibit the cell wall synthesis [Chung et al., 2009], and quinolones that target DNA Gyrase [Marcusson et al., 2009].

In the earliest attempt for the identification of essential genes, Mushegian and Koonin compared the genomes of *Haemophilus influenzae* and *Mycoplasma genitalium*, assuming that the genes that are shared in these two phylogenetically distant bacteria are indispensable and reported 256 genes fulfilling this requirement [Mushegian and Koonin, 1996]. With the advent of sequencing technologies and availability of more genome sequences, the number of core genes in different prokaryotic genomes declined to less than 50 which is not enough for performing all essential functions in a cell [Charlebois and Doolittle, 2004]. Therefore, the use of experimental methods for the identification of essential genes is vital. Researchers have now studied the essential genes in organisms from all three domains of life [Luo et al., 2014] using a number of different methods. Baba et al. [Baba et al., 2006] made a library of single gene deletions for *Escherichia coli* K-12. The 303 genes where viable *Escherichia coli* colonies failed to grow are the candidate essential genes. Another group used an antisense RNA knockdown approach to study gene essentiality in *Staphylococcus* aureus [Forsyth et al., 2002]. In this method, if the expression of an antisense RNA hinders the growth of the cell, its cognate gene is known as essential. Both of these methods are labour intensive and are dependent on the accuracy of genome annotations. Another widely used procedure is transposon mutagenesis combined with high-throughput sequencing [Chao et al., 2016, van Opijnen and Camilli, 2013, Barquist et al., 2013a] which includes different approaches namely, Tn-Seq [van Opijnen et al., 2009], INSeq [Goodman et al., 2009], HITS [Gawronski et al., 2009] and TraDIS [Langridge et al., 2009]. These procedures differ in the type of transposon, sample preparation methods, and data analysis [van Opijnen and Camilli, 2013]. Nonetheless, all share the same workflow: pools of single insertion mutants are constructed using transposon mutagenesis. After a growth phase, mutants that are fitter outnumber the less fit ones. Using high-throughput sequencing and tallying transposon junctions gives an indication of whether a genomic region is essential or not. A high number of transposon insertions in a gene indicates that the gene is not essential in its growth medium and conversely, a low number of transposon insertions indicates that a gene is essential in the medium.

There are two groups of essential genes: core essential genes are indispensable for all cells, and accessory essential genes that are required for some organisms. Core essential genes can shed light on the genome structure of the last universal common ancestor and the evolution of living cells [Koonin, 2003] and have been used for the synthesis of minimal cells [Hutchison et al., 2016]. On the other hand, accessory essential genes are helpful in the study of specific lineages. Accessory essential genes may be useful in species-specific antibiotic discovery. Antibiotics that target core essential genes in pathogens may not be ideal, as they may attack homologous genes in their hosts or commensal bacteria.

Freed et al. [Freed et al., 2016] have investigated the difference between essential genes in *Shigella flexneri* 2a 2457T and *Escherichia coli* K12 BW25113 and shown that there are no genes that are essential in *Escherichi coli* and not essential in *Shigella flexneri*, while some genes are only essential in *Shigella flexneri.* These include a group of genes involved in cysteine, proline and sugar nucleotide biosynthesis, acetate utilisation, translation elongation, aminoacyl tRNA synthetase, murein DD-endopeptidase, and soxR-reducing complex, many of which are essential in *Shigella flexneri* due to the absence of paralogs or other alternative systems that exist in *Escherichia coli.*

Canals et al. [Canals et al., 2012] have compared the essentiality of genes in *Salmonella* typhimurium ATCC 14028, two isolates of *Salmonella typhi* Ty2 (varying in *htrA*, *aroC* and *aroD* genes), and *Escherichia coli* K12 BW25113 and found that these cells share 268 essential genes. Nine genes are essential in *Escherichia coli* and not essential in *Salmonella*s, three of which can tolerate insertions only in some parts and one has distant paralogs in *Salmonella*s that is missing in *Escherichia coli.* For other genes, there are a few insertions in *Salmonella*s, but the number is not small enough to call those genes essential. Moreover, 159 genes were almost essential in all *Salmonella*s but not in *Escherichia coli.* These include genes involved in replication and genes related to ribosome and its accessory proteins. The authors also found 26 genes that are under greater selection in *Salmonella* Typhimurium compared to *S. typhi* and 10 genes vice versa. Barquist et al. [Barquist et al., 2013b] have used transposon-directed insertion-site sequencing to compare the essentiality of genes in *Salmonella* Typhi Ty2, *Salmonella* Typhimurium SL1344, and *Escherichia coli* K12 BW25113. These genomes share 228 essential genes which are mostly involved in cell division, transcription, translation, and fatty acid and peptidoglycan biosynthesis. Additionally, many of the serovar-specific essential genes in *Salmonella*s are phage repressors. Another key difference between these two serovars is that a set of genes that are putatively involved in cell wall biosynthesis are essential in *Salmonella* Typhimurium and not essential in *salmonella* Typhi which gives an indication of the adaptation of these two *Salmonella* serovars to their niches.

Although there are some studies on differentiation of essentiality in different organisms, these studies usually include a few genomes. We are going to extend these studies to a larger scale in Enterobacteriaceae family. Our aim is to study the essentiality of genes in in 16 different organisms. These include *Enterobacter cloacae* NCTC 9394, *Klebsiella pneumoniae* Ecl8, *Klebsiella pneumoniae* RH201207, *Citrobacter rodentium* ICC168, *Salmonella* Typhimurium SL1344, *Salmonella* Typhimurium SL3261, *Salmonella* Typhimurium D23580, *Salmonella* Typhimurium A130, *Salmonella* Enteritidis P125109, *Salmonella* Typhi Ty2, *Escherichia coli* ST131 EC958, *Escherichia coli* UPEC ST131, *Escherichia coli* ETEC CS17, *Escherichia coli* ETEC H10407, *Escherichia coli* BW25113, and *Escherichia coli* K-12 MG1655.

2 Results and Discussion

2.1 Incorporating multiple measures of gene essentiality improves prediction

In this section we will compare a number of methods of quantifying the essentiality of genes based on transposon insertion data and introduce our own method.

2.1.1 Comparison of old methods

A number of methods have been used for evaluating the essentiality of genes using transposon insertion data. Freed et al. [Freed et al., 2016] have compared eleven features that can quantify the essentiality of genes. These include: the number of insertion sites within genes; the mean distance between insertion sites, the median distance between insertion sites, the number of base pairs before the first insertion in the $5'$ end, the ratio between the number of base pairs before the first insertion in the $5'$ end and gene length, the number of base pairs before the second insertion in the $5'$ end, the ratio between the number of base pairs before the second insertion in the $5'$ end and gene length, the length of the longest uninterrupted region in the gene, the ratio between the length of the longest uninterrupted region in the gene and gene length, the length of the longest region in the gene containing at most one insertion divided by gene length, and the number of insertion sites divided by gene length. Among these, the average distance between insertion sites and the length of the longest uninterrupted region were the best predictors.

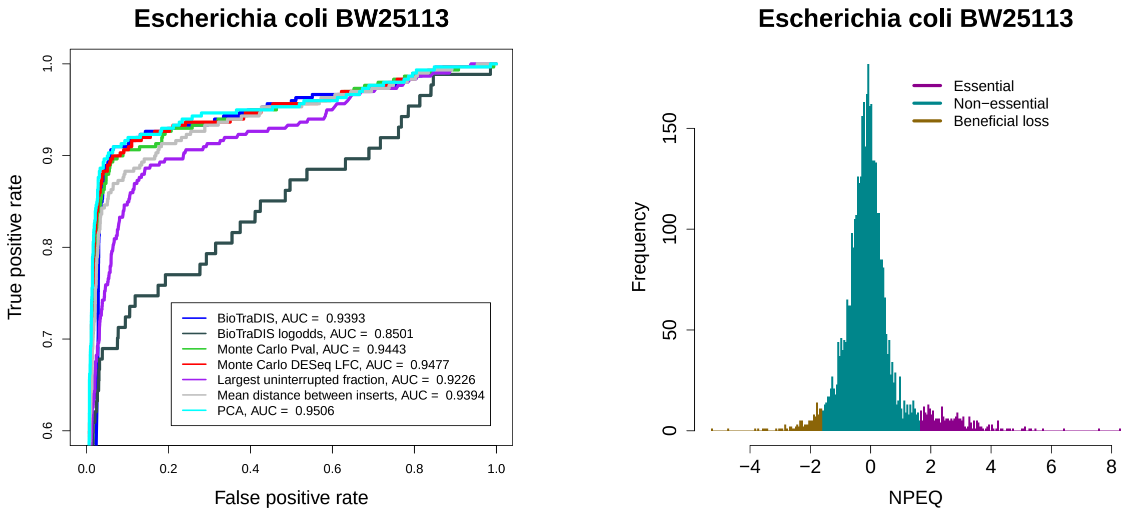
Barquist et al. [Barquist et al., 2016] have used insertion indices which are calculated by dividing the number of insertion sites by gene length for defining the essentiality of genes. Plotting all insertion indices for all genes in a genome gives a bimodal plot, each mode representing a group of genes (essential or non-essential). They have fitted two gamma distributions to these modes and calculated the log odds value for each gene to test which mode it belongs to.

Turner et al. [Turner et al., 2015] have randomly sampled insertions in a genome and using DESeq package [Anders and Huber, 2010] calculated log fold changes and P-values for the actual number of insertions compared to the expected number of insertions obtained from the sampling method.

We have compared the predictive power of the average distance between insertion sites in a gene, the length of the longest uninterrupted region, insertion index, log odds value obtained from insertion index value, and P-value and log fold change calculated using DESeq package after sampling. To evaluate the accuracy of these methods, we have compared the predicted essential genes to the essential genes in *Escherichia coli* K-12 MG1655 in EcoGene database [Zhou and Rudd, 2013]. The results are depicted in Figure 1a. Among all methods, those that try to fit a distribution to data and predict essentiality based on that, namely log odds value from insertion indices and P-value from DESeq after sampling, are the least accurate methods. Sampling and then calculating log fold changes using DESeq package is the most accurate predictor and the other three methods are also good predictors despite their simplicity.

2.1.2 PCA based method

We selected four of the predictors in previous section: insertion index, log fold change using sampling and DESeq, the average distance between insertion sites in a gene, and the length of the longest uninterrupted region. These predictors were selected as they were more accurate and less dependent on each other. Afterwards, we ran principal component analysis on these predictors using *prcomp* function in R and selected the first principal component.



(a) **(b)**

Figure 1. (a) The accuracy of 7 different prediction methods for quantifying the essentiality of genes. The higher the area under the curve, the more accurate the method is. (b) The distribution for our proposed method (NPEQ)

As Figure 1a shows, PCA results gave us the most powerful tool for quantifying the essentiality of genes compared to other methods. We have plotted the results for one sample genome and the results for other genomes are shown in SUPPLEMENTARY. Figure 1b shows the distribution of PCA results after zero mean unit variance normalisation. We call this value NPEQ (Normalised Pca based Essentiality Quantifier) in this paper. The genes whose NPEQ values are less than -1.644854 after normalisation are beneficial losses and the genes with NPEQ greater than 1.644854 are essential. The rest are non-essential genes. The cutoff 1.644854 has been selected because if we assume that NPEQ distribution is normal, this cutoff shows the P-value 0.95. Moreover, by trying to define a cutoff that maximises Matthews correlation coefficient for each genome using NPEQ, we get 1.650449 as the average cutoff which is very close to this value.

2.2 TraDIS data is biased

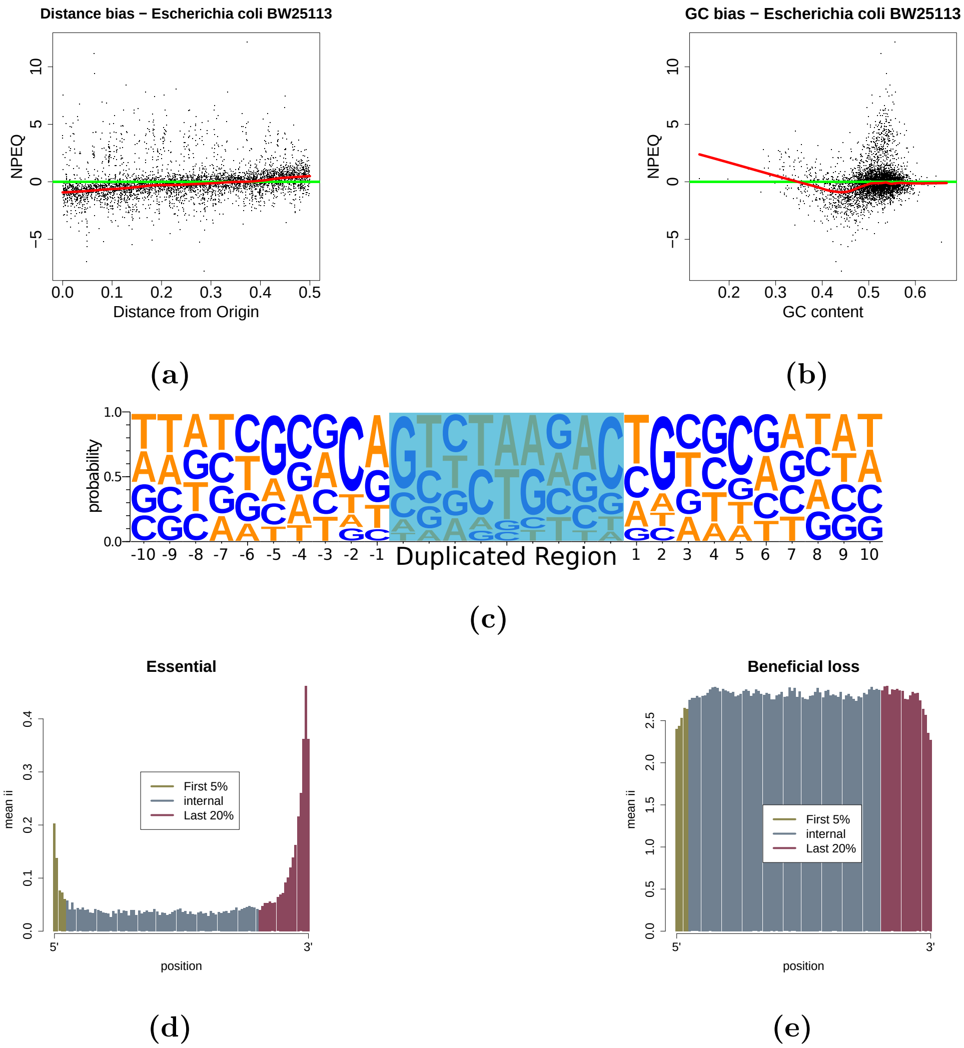
In Section 2.1.2 we proposed NPEQ method that quantifies the level of essentiality of a gene. However, if the transposon insertion is biased to specific regions in the genome, it can increase/decrease the values predicted by our method for genes and put them in a different level of essentiality. Different articles have reported biases in transposon insertion using Tn5 [Barquist et al., 2013b, Green et al., 2012, Rubin et al., 2015, Canals et al., 2012, Langridge et al., 2009]. We performed a detailed study of these biases. The biases that we studied include: origin of replication bias, preferred insertion motif bias, and positional bias within genes.

2.2.1 Distance from the origin of replication bias

While a study has reported Tn5 insertion bias near the origin of replication [Barquist et al., 2013b], another study has reported no bias [Rubin et al., 2015]. To study the bias towards the position of gene within genome, we plotted NPEQ for each gene versus the distance of the gene from the origin of replication normalised by the length of the genome in Fig. 2a. The figure indicates that NPEQ increases when the genes are located further from the origin of replication. When the bacteria are under replication during the transposn insertion process, there are more copies of the genes close to the origin of replication than the genes further away due to the initiation of different replication forks. This results in more insertions in the genes near the origin of replication which can influence the accuracy of our predictions.

2.2.2 Preferred nucleotide composition bias

Another concern while inferring essentiality from transposon insertion data is that transposon insertion is biased to certain compositions of nucleotides and high number of insertions in genes reflects the enrichment of the motifs that transposon insertion is biased towards, instead of their essentiality level. During Tn5 transposition, a sequence of 9 nucleotides is duplicated. Lodge et al. [Lodge et al., 1988] showed that these duplicated regions have G-C pairs at two ends and replacing these G-C pairs with A-T pairs reduces the number of transposon insertions by more than fivefold. Goryshin et al. [Goryshin et al., 1998] reported the palindromic sequence A-GNTYWRANC-T as the consensus target site for Tn5 transposition where the 9 letters in middle show the consensus sequence in the duplicated region. In some other research a similar consensus motif has been found for Tn5 [Canals et al., 2012] which is CGCGCA-GTTYWRAAC-TGCGCG. Others [Green et al., 2012, Rubin



(a)

(b)

(c)

(d)

(e)

Figure 2. (a)The distance of the genes from DnaA gene normalised by the lengths of the genomes versus NPEQ. The distance from DnaA gene has been calculated in both directions and then the minimum value has been used for distance. The red curves show the loess curve when the smoothness parameter is 0.2 and the green line shows no bias. (b) G-C content of genes against their NPEQ values. The red curves show the loess curve when the smoothness parameter is 0.2 and the green line shows no bias. (c) Sequence logo plots generated using sequences from 10 nucleotides flanking the 100 top most frequent insertion sites from each genome. The height of each character shows the relative frequency of that character. (d) and (e) The plots show the average insertion index in percentiles of all essential genes (d) and beneficial losses (e). The genes are divided into 3 segments: 5% of the genes on the 5′ end, 20% of the genes on the 3′ end, and the rest in the middle. These are shown by khaki, slate gray, and violet red respectively.

et al., 2015] have not found such a sequence but shown that the duplicated regions are G-C rich. We used Weblogo [Crooks et al., 2004] to generate a logo from duplicated regions and 10 nucleotides flanking the 100 top most frequent insertion sites in each genome. The results in Figure 2c show the consensus motif that we have found is very similar to [Green et al., 2012] and the only difference is in positions 3 and 7 within the duplicated region.

The other possible source of bias is if transpositions are more inclined to G-C or A-T rich regions. Rubin et al. [Rubin et al., 2015] have reported that the number of Tn5 insertions rises with the increase of G-C content and Green et al. [Green et al., 2012] have shown that the highest number of insertions occur in high G-C content regions. On the other hand, Langridge et al. [Langridge et al., 2009] have seen an increase in the number of Tn5 insertions in 40% G-C content. In Figure 2b, we plotted the G-C content of genes versus their NPEQ values. The red curves are loess curves with smoothness parameter 0.2 and the green line shows no bias. NPEQ decreases gradually (which means an increase in the number of insertions) as G-C content decreases and then somewhere between 40% and 50% G-C content, NPEQ starts to rise again in almost all genomes. In the region where most of the genes are packed, the loess curve is almost flat. On the left side of this flat region, there are genes with different G-C content which are enriched in mobile genetic elements. So, we expect to see more insertions (smaller NPEQ values) in this region. We have low NPEQ values between 40% and 50% G-C content which is expected. However, in most cases when we have less than 40% G-C content, NPEQ value is high. A possible reason for this phenomena is the association of A-T rich sequences and histone-like nucleotide structuring (H-NS) proteins, which reduces the insertions in A-T rich regions. This has been shown for Tn10 transposon [Kimura et al., 2016], but not for Tn5 transposon, yet. Overall, the results are consistent with Langridge et al. [Langridge et al., 2009]. The G-C content of the most of the genes with large NPEQ values is between 50% and 60% which is inconsistent with Green et al. [Green et al., 2012] as this region contains most of the genes and is not considered as a high G-C content region but rather an average G-C content region.

2.2.3 Positional bias within genes

Some research has indicated that himar1 transposons are more probable to get inserted into the two ends of a gene compared to the middle [Griffin et al., 2011]. We have tested this hypothesis using our TraDIS data. We divided every gene into 100 fragments with equal lengths (percentiles) and calculated the mean insertion index for each percentile. Insertion index is calculated using $\frac{\frac{n_p}{l_p}}{\frac{n_g}{l_g}}$, where n_p is the number of insertion sites in a specific percentile, l_p is the length of that percentile, n_g is the number of insertion sites in the whole genome and l_g is genome length. Mean insertion index for each percentile is calculated by averaging over all insertion indices for that specific percentile of genes. We saw almost no bias towards any location when considering all genes together (Fig. ?? SUPPLEMENTARY). We studied the bias in three different groups of genes: essential genes which have no or just a few insertions, non-essential genes which have an intermediate number of insertions, and beneficial losses which have a high number of insertions. The results imply that the number of insertions in the internal region of the essential genes is outnumbered by the number of insertions in the $5'$ and $3'$ ends (Fig. 2d) while it is the opposite in beneficial losses (Fig. 2e). High number of insertions at the $3'$ end of essential genes implies that the functional domains are located before the insertions and the insertions are not interfering with them. On the other hand, high number of insertions at the $5'$ end of the essential genes indicates there might be alternative start codons in the $5'$ end or it might be because of annotation errors that have predicted the start codon in an incorrect place before the actual start codon.

2.3 Most essential genes are ubiquitously essential in Enterobacteriaceae

Previous studies of gene essentiality in Enterobacteriaceae family have compared essential genes in different genomes in this family and studied the sets of core essential genes and accessory essential genes [Freed et al., 2016, Canals et al., 2012, Barquist et al., 2013b]. Core essential genes are responsible for essential processes such as cell division, DNA replication, transcription and translation and some important pathways like peptidoglycan and fatty acid biosynthesis [Barquist et al., 2013b]. Accessory essential genes differ in genomes due to different reasons such as niche adaptation, gene duplication and the existence of alternative pathways [Freed et al., 2016, Canals et al., 2012, Barquist et al., 2013b]. Another group of accessory essential genes are phage repressors [Barquist et al., 2013b]. Even though these genes are not essential for the growth of a cell, once phages are introduced to a cell, they become essential as long as the phage remains in the cell.

In this study, we have compared the genes in 16 bacteria from Enterobacteriaceae family. For this we needed to study core and accessory sets of genes. Moreover, in the presence of two redundant variations of one gene, if we knock out one copy using TraDIS, the other copy compensates and the organism can still survive [Dean et al., 2008]. This leads to a different essentiality inference using TraDIS. Therefore, in addition to core and accessory genes, we have studied duplicate genes. To study whether each gene in the 16 organisms is core, we used Jackhmmer from HMMER package [Eddy, 2011] to iteratively search for homologous proteins in our dataset and cluster them. We divided the clusters of homologous genes into three groups based on their conservation. Genus specific class contains genes that are present only in one genus, the genes in single copy class are present in more than one genus and more than 70% of them are not duplicated (core genes), and the genes in multi-copy class are present in more than one genus and more than 30% of them are duplicated (duplicate genes).

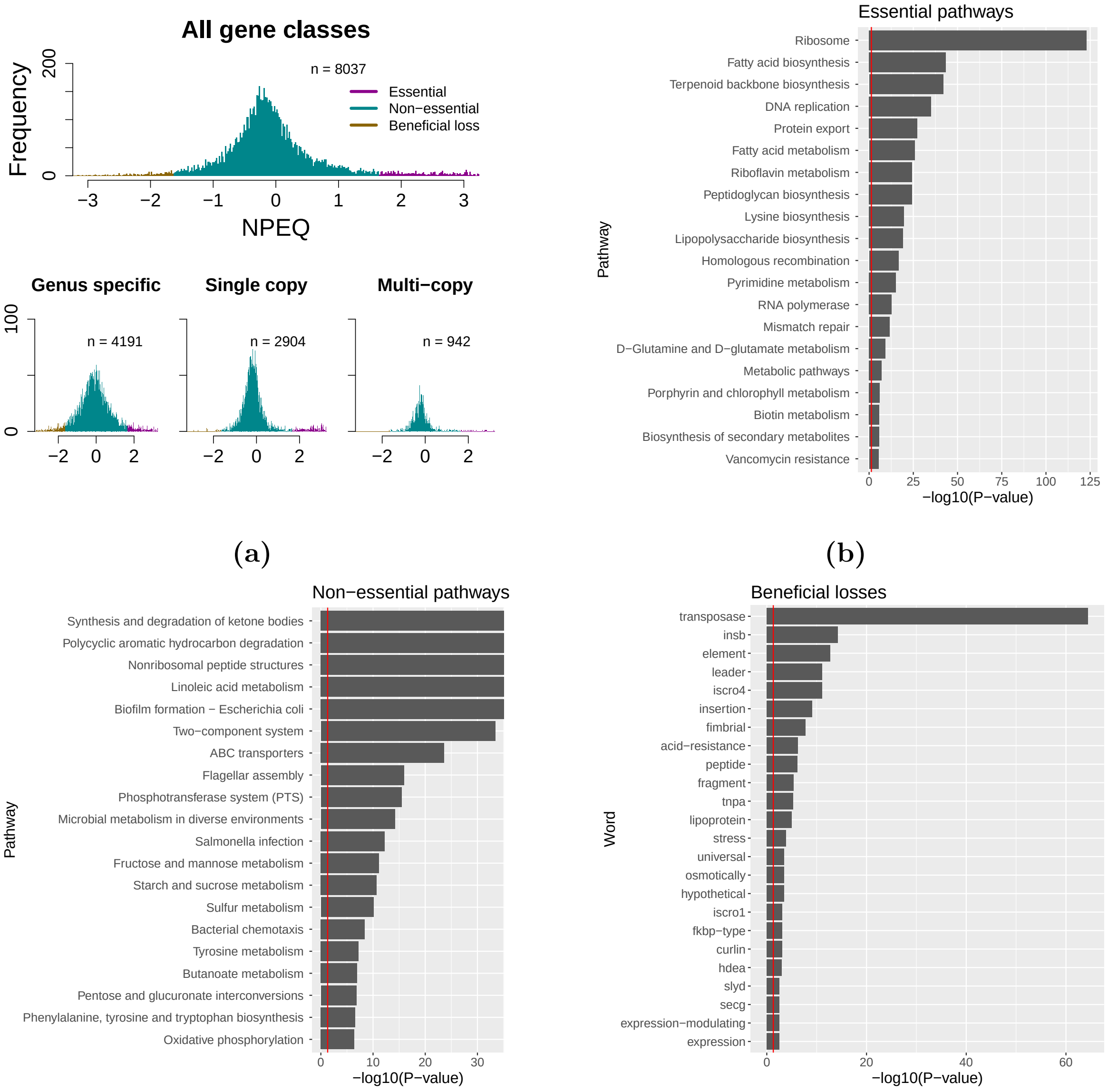
We also evaluated the essentiality of genes and divided the genes into different levels of essentiality using NPEQ values. If NPEQ is less than -1.644854, it means that the gene has tolerated many insertions, so it is beneficial for the organism to lose this gene in the rich medium that we used. However, if NPEQ is greater than 1.644854, the gene can tolerate very few or no insertions indicating that the gene is essential for cell viability in our test medium. Any other NPEQ value shows an intermediate number of insertions in genes meaning that it is not beneficial for the organism to lose these genes, but they are not essential, too. We have called this group of genes non-essential.

The results for comparing three levels of essentiality and three classes of conservation are depicted in Fig. 3a. The high number of single copy clusters in essential level, indicates that there is a set of essential genes in Enterobacteriaceae that are conserved and inclined to keep their essentiality. However, there are also many essential genus specific genes. The figure also shows that beneficial losses are over represented in genus specific class. Therefore, beneficial losses are mostly recent genes that the organism tends to lose in the long run. Besides, most of the multi-copy clusters are non-essential and there are only a few multi-copy clusters that are essential. This can be explained by the redundancy that duplicate genes can keep even after $\sim$ 100 million years [Dean et al., 2008].

To study which functions are enriched in each class of essentiality, we used KEGG pathway enrichment analysis [Kanehisa and Goto, 2000] and compared the genes in each class with the databases that were available for the genomes in this study. The results show that essential genes are enriched in pathways related to genetic information processing such as replication (DNA replication, homologous recombination and mismatch repair), transcription (RNA polymerase), translation (ribosome), and protein export. These are the essential functions that every cell needs for its viability. Other enriched pathways are mostly metabolism related. These include fatty acid biosynthesis that produces cell membrane, peptidoglycan and lipopolysaccharide biosynthesis that are essential components of cell wall, terpenoid backbone biosynthesis which feeds peptidoglycan biosynthesis, nucleotide and amino acid metabolism, and the metabolism of important cofactors and vitamins like riboflavin, biotin,porphyrin and chlorophyll (Figure 3b). CAN WE HAVE SOMETHING LIKE THIS? http://www.nature.com/articles/srep00125

Non-essential genes are mostly involved in carbohydrate, energy, and lipid metabolism, membrane transport, cell motility and cellular community. Even though these functions are important, they are not essential in a rich medium in lab.

As most of beneficial losses do not have homologs in other genomes, most of them are not studied and therefore there is no KEGG pathway for them. Because of this reason, we studied the description of these genes in their embl files and found the words that were enriched. These are shown in Figure 3d. The results show that most of beneficial losses are mobile genetic elements like transposases and insertion elements which are not essential in their host genomes, genes that are not essential in a rich medium like stress protein and acid-resistance protein, and hypothetical proteins.



(a)

(b)

(c)

(d)

Figure 3. (a) The genes have been clustered into homologous groups using Jackhmmer and divided into 3 groups: genus specific, single copy, and multi-copy genes. Then, the essentiality of the clusters has been defined using the insertion indices of the genes in the clusters. The figure shows that most of the essential genes are in single copy group, while most of the beneficial losses are genus-specific. (b) and (c) KEGG Pathways enriched in essential genes (b) and non-essential genes (c). (d) The words enriched in the description of beneficial losses in their embl files compared to other genes. The red line shows P-value = 0.05. The P-values are calculated using hypergeometric test and then corrected using Benjamini-Hochberg-Yekutieli procedure.

2.4 Phylogenetic analysis of gene essentiality identifies that essential genes are more likely to be conserved

Here, we have tested this hypothesis by comparing the number of core essential genes and core genes in the phylogenetic tree as we go up to the root. We have used Fitch's algorithm [Fitch, 1971] with a binary alphabet on both essentiality (0 for non-essential and 1 for essential) and conservation (1 for the presence and 0 for the absence of genes) to define if a gene is core essential or core at each level in the phylogenetic tree. If we get 1 at the top of the tree while studying essentiality of a gene using Fitch algorithm, it means that the gene is core essential and otherwise, the gene is non-essential. The same applies to the study of core genes. The phylogenetic tree has been annotated with the number of core essential genes (red) and the number of core genes (blue) at each level in Figure 4a. We then plotted the ratios at each level in the phylogenetic tree in Figure 4b and connected the medians in each level. The connecting line shows that the ratio between core essential genes and core genes rises as we go higher in the phylogenetic tree which means essential genes are more likely to be conserved in genomes compared to non-essential genes.

We have studied the essentiality status of genes involved in important biological processes in Table 1. These processes include cell division, DNA replication, transcription, translation, and important metabolic pathways such as peptidoglycan and fatty acid biosynthesis. If a gene was not core essential using Fitch algorithm, we looked at its essentiality in every genome: if it was essential in some of the genomes we classified it as accessory essential and if it was not essential in any genome we classified it as non-essential. FtsA and zipA genes are essential for cell division. FtsA is essential in all the genomes in our study, however zipA is essential everywhere except for Salmonella typhimurium A130 in which this gene is near essential. We investigated if this gene has a homolog in Salmonella typhimurium A130 and found no homologs. It has been shown that 21 mutations in ftsA gene can make it independent of zipA gene so that the zipA gene is no longer essential [Geissler et al., 2003, Pichoff et al., 2012]. However, ftsA gene does not have any of these mutations in Salmonella typhimurium A130, so there might be another factor that is causing zipA not to be essential. ADD MORE EXAMPLES HERE

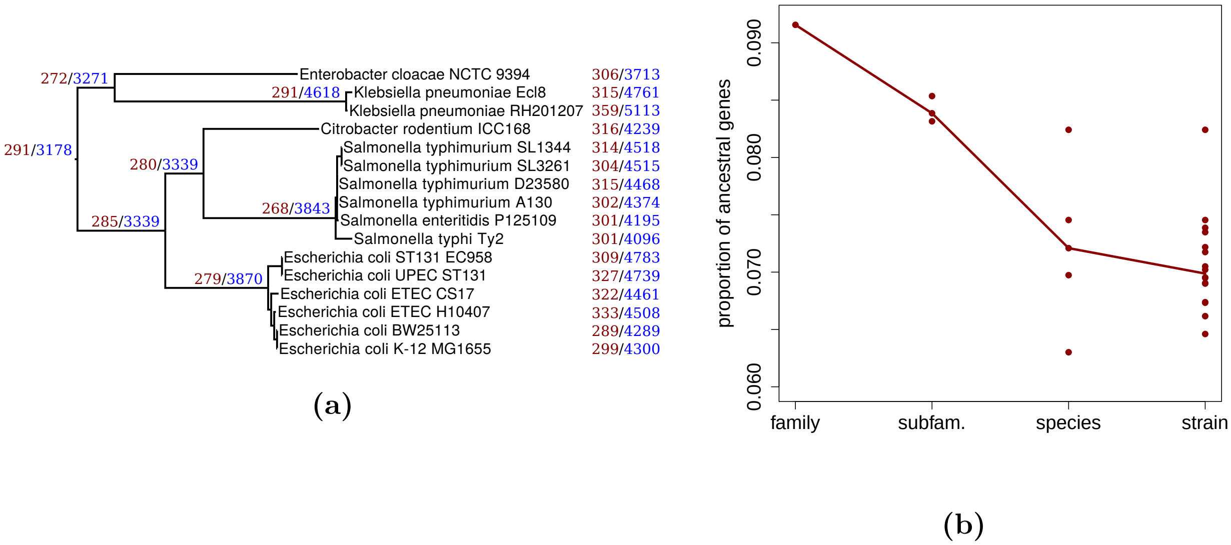


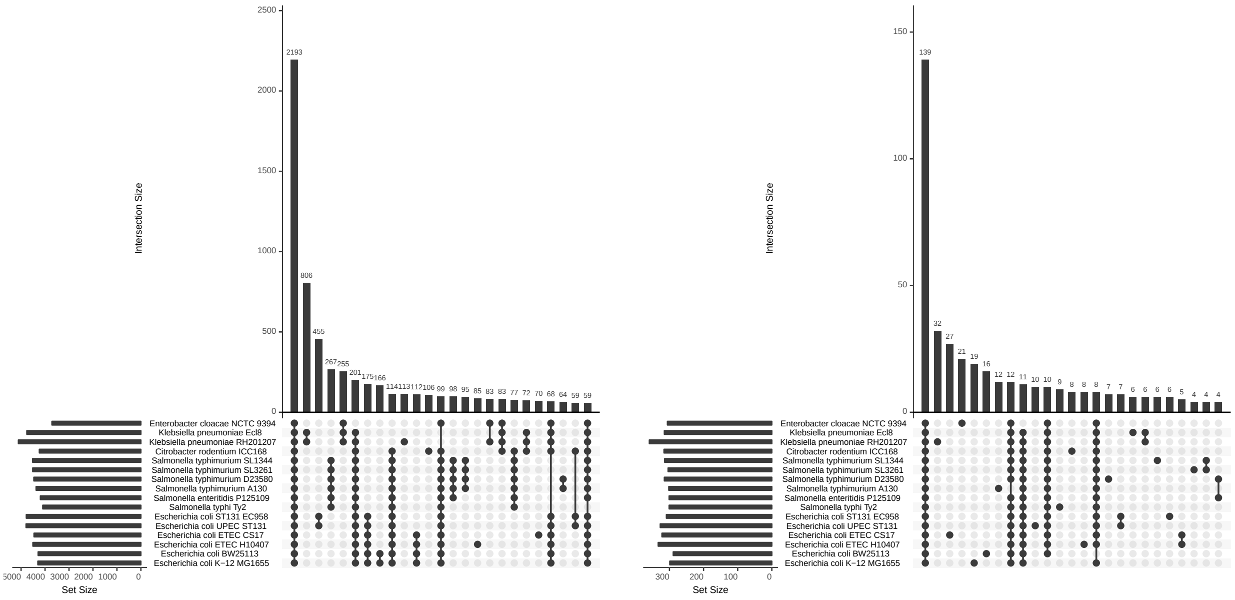
Figure 4. (a) The species tree for all genomes in this study. Numbers in red show the number of core essential genes at each level and numbers in blue show the number of core genes at each level. (b) The ratio between core essential genes and core genes at each level in the species tree. The dots in the strain level show the ratios for all 16 bacteria; the dots in the species level show the ratios for Enterobacter, Klebsiella, Citrobacter, Salmonella, and Escherichia; the subfam. level shows the ratios for the common ancestor of Enterobacter and klebsiella, the common ancestor of Citrobacter and Salmonella, and the common ancestor of Citrobacter, Salmonella, and Escherichia; and finally the dot in the family level shows the ratio for the root. The line connects the medians in each level.

So far we have shown that essential genes are less likely to be lost during evolution. The other question is if essential genes tend to keep their essentiality in different genomes. Bergmiller et al. [Bergmiller et al., 2012] have shown that only essential genes whose functions cannot be taken over by other genes keep their essentiality and others are possible to be lost or non-essential in other genomes. The genes that can replace other genes are not necessarily homologous to them, but they have similar functions.

We have compared the number of genes that are shared between our bacterial genomes (5a) and the number of genes essential in them (5b) and used UpSetR package [Conway and Gehlenborg, 2016] to visualise the results in Fig. 5. As shown in the figures, among 2162 genes that are shared between all the bacteria under study, only 135 are essential everywhere. We looked at subsets of the genes that were shared in every combination of our bacterial strains to see whether they show a trend similar to the phylogenetic tree. The results propose that although conservation of genes follows a tree-like trend, the essentiality does not show such a trend. The reason is that many of the essential genes that are not shared in all genomes are probably compensable by other genes and so the essentiality does not look tree-like.

Table 1. The essentiality status of genes involved in important biological processes

Biological process	Subprocess	Core essential	Accessory essential	Non-essential
Cell Division		ftsAHLQWYZ, minE, mukB, zipA	ftsKNX, minD	CedA, ftsJ, minC, sdiA, sulA
DNA replication	Polymerases I, II, III	dnaENQX, holABD, polA	holC	holE, polB
	Supercoiling	gyrAB, parCE		
	Primosome-associated	dnaBCGT, priA, ssb	priB, rep	priC
Transcription	RNA polymerase	rpoABC		
	Sigma, elongation, anti- and termination factors	nusABG, rho, rpoDH	rpoEN	rpoS
Translation	tRNA-synthetases	alaS, argS, asnS, aspS, cysS, glnS, gltX, glyQS, hisS, ileS, leuS, lysS, metG, pheST, proS, serS, thrS, tyrS, valS	trpS	
	Ribosome components	rplBCDEFJKLMNOPQRSTUV, rplWXY, rpmABCD, rpsABCDEFGHJKLMNPQRSTU	rplA, rpmEGHI, rpsIO	rplI, rpmF
	Initiation, elongation and peptide chain release factors	fusA, infABC, prfAB, tsf	efp	prfC, selB, tufAB
Biosynthetic pathways				
Peptidoglycan		MraY, murABCDEFGI		ddlAB
Fatty acids		accABCD, fabABDGHIZ		



(a) **(b)**

Figure 5. (a) The number of genes and (b) essential genes shared between different groups of bacteria. The bars show the number of genes and the filled circles show which bacteria are sharing those genes.

3 Materials and Methods

References

Anders and Huber, 2010. Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, 11:R106.

Baba et al., 2006. Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K. A., Tomita, M., Wanner, B. L., and Mori, H. (2006). Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Molecular Systems Biology*, 2:2006.0008.

Barquist et al., 2013a. Barquist, L., Boinett, C. J., and Cain, A. K. (2013a). Approaches to querying bacterial genomes with transposon-insertion sequencing. *RNA Biology*, 10(7):1161–1169.

Barquist et al., 2013b. Barquist, L., Langridge, G. C., Turner, D. J., Phan, M.-D., Turner, A. K., Bateman, A., Parkhill, J., Wain, J., and Gardner, P. P. (2013b). A comparison of dense transposon insertion libraries in the Salmonella serovars Typhi and Typhimurium. *Nucleic Acids Research*, page gkt148.

Barquist et al., 2016. Barquist, L., Mayho, M., Cummins, C., Cain, A. K., Boinett, C. J., Page, A. J., Langridge, G. C., Quail, M. A., Keane, J. A., and Parkhill, J. (2016). The TraDIS toolkit: sequencing and analysis for dense transposon mutant libraries. *Bioinformatics*, 32(7):1109–1111.

Bergmiller et al., 2012. Bergmiller, T., Ackermann, M., and Silander, O. K. (2012). Patterns of Evolutionary Conservation of Essential Genes Correlate with Their Compensability. *PLOS Genetics*, 8(6):e1002803.

Brodersen et al., 2000. Brodersen, D. E., Clemons Jr., W. M., Carter, A. P., Morgan-Warren, R. J., Wimberly, B. T., and Ramakrishnan, V. (2000). The Structural Basis for the Action of the Antibiotics Tetracycline, Pactamycin, and Hygromycin B on the 30s Ribosomal Subunit. *Cell*, 103(7):1143–1154.

Canals et al., 2012. Canals, R., Xia, X.-Q., Fronick, C., Clifton, S. W., Ahmer, B. M., Andrews-Polymenis, H. L., Porwollik, S., and McClelland, M. (2012). High-throughput comparison of gene fitness among related bacteria. *BMC Genomics*, 13:212.

Chao et al., 2016. Chao, M. C., Abel, S., Davis, B. M., and Waldor, M. K. (2016). The design and analysis of transposon insertion sequencing experiments. *Nature Reviews Microbiology*, 14(2):119–128.

Charlebois and Doolittle, 2004. Charlebois, R. L. and Doolittle, W. F. (2004). Computing prokaryotic gene ubiquity: Rescuing the core from extinction. *Genome Research*, 14(12):2469–2477.

Chung et al., 2009. Chung, H. S., Yao, Z., Goehring, N. W., Kishony, R., Beckwith, J., and Kahne, D. (2009). Rapid beta-lactam-induced lysis requires successful assembly of the cell division machinery. *Proceedings of the National Academy of Sciences*, 106(51):21872–21877.

Conway and Gehlenborg, 2016. Conway, J. and Gehlenborg, N. (2016). UpSetR: A More Scalable Alternative to Venn and Euler Diagrams for Visualizing Intersecting Sets.

Crooks et al., 2004. Crooks, G. E., Hon, G., Chandonia, J.-M., and Brenner, S. E. (2004). WebLogo: a sequence logo generator. *Genome Research*, 14(6):1188–1190.

Dean et al., 2008. Dean, E. J., Davis, J. C., Davis, R. W., and Petrov, D. A. (2008). Pervasive and Persistent Redundancy among Duplicated Genes in Yeast. *PLOS Genet*, 4(7):e1000113.

Eddy, 2011. Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLOS Comput Biol*, 7(10):e1002195.

Fischbach and Walsh, 2009. Fischbach, M. A. and Walsh, C. T. (2009). Antibiotics for Emerging Pathogens. *Science*, 325(5944):1089–1093.

Fitch, 1971. Fitch, W. M. (1971). Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology. *Systematic Biology*, 20(4):406–416.

Forsyth et al., 2002. Forsyth, R. A., Haselbeck, R. J., Ohlsen, K. L., Yamamoto, R. T., Xu, H., Trawick, J. D., Wall, D., Wang, L., Brown-Driver, V., Froelich, J. M., C, K. G., King, P., McCarthy, M., Malone, C., Misiner, B., Robbins, D., Tan, Z., Zhu Zy, Z.-y., Carr, G., Mosca, D. A., Zamudio, C., Foulkes, J. G., and Zyskind, J. W. (2002). A genome-wide strategy for the identification of essential genes in Staphylococcus aureus. *Molecular Microbiology*, 43(6):1387–1400.

Freed et al., 2016. Freed, N. E., Bumann, D., and Silander, O. K. (2016). Combining Shigella Tn-seq data with gold-standard E. coli gene deletion data suggests rare transitions between essential and non-essential gene functionality. *BMC microbiology*, 16(1):203.

Gawronski et al., 2009. Gawronski, J. D., Wong, S. M. S., Giannoukos, G., Ward, D. V., and Akerley, B. J. (2009). Tracking insertion mutants within libraries by deep sequencing and a genome-wide screen for Haemophilus genes required in the lung. *Proceedings of the National Academy of Sciences*, 106(38):16422–16427.

Geissler et al., 2003. Geissler, B., Elraheb, D., and Margolin, W. (2003). A gain-of-function mutation in ftsA bypasses the requirement for the essential cell division gene zipA in Escherichia coli. *Proceedings of the National Academy of Sciences*, 100(7):4197–4202.

Gerdes et al., 2003. Gerdes, S. Y., Scholle, M. D., Campbell, J. W., Balázsi, G., Ravasz, E., Daugherty, M. D., Somera, A. L., Kyrpides, N. C., Anderson, I., Gelfand, M. S., Bhattacharya, A., Kapatral, V., D'Souza, M., Baev, M. V., Grechkin, Y., Mseeh, F., Fonstein, M. Y., Overbeek, R., Barabási, A.-L., Oltvai, Z. N., and Osterman, A. L. (2003). Experimental Determination and System Level Analysis of Essential Genes in Escherichia coli MG1655. *Journal of Bacteriology*, 185(19):5673–5684.

Goodman et al., 2009. Goodman, A. L., McNulty, N. P., Zhao, Y., Leip, D., Mitra, R. D., Lozupone, C. A., Knight, R., and Gordon, J. I. (2009). Identifying genetic determinants needed to establish a human gut symbiont in its habitat. *Cell Host & Microbe*, 6(3):279–289.

Goryshin et al., 1998. Goryshin, I. Y., Miller, J. A., Kil, Y. V., Lanzov, V. A., and Reznikoff, W. S. (1998). Tn5/IS50 target recognition. *Proceedings of the National Academy of Sciences of the United States of America*, 95(18):10716–10721.

Green et al., 2012. Green, B., Bouchier, C., Fairhead, C., Craig, N. L., and Cormack, B. P. (2012). Insertion site preference of Mu, Tn5, and Tn7 transposons. *Mobile DNA*, 3:3.

Griffin et al., 2011. Griffin, J. E., Gawronski, J. D., DeJesus, M. A., Ioerger, T. R., Akerley, B. J., and Sassetti, C. M. (2011). High-Resolution Phenotypic Profiling Defines Genes Essential for Mycobacterial Growth and Cholesterol Catabolism. *PLoS Pathogens*, 7(9).

Hutchison et al., 2016. 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., and Venter, J. C. (2016). Design and synthesis of a minimal bacterial genome. *Science*, 351(6280):aad6253.

Jordan et al., 2002. Jordan, I. K., Rogozin, I. B., Wolf, Y. I., and Koonin, E. V. (2002). Essential Genes Are More Evolutionarily Conserved Than Are Nonessential Genes in Bacteria. *Genome Research*, 12(6):962–968.

Juhas et al., 2013. Juhas, M., Davenport, P. W., Brown, J. R., Yarkoni, O., and Ajioka, J. W. (2013). Meeting report: The Cambridge BioDesign TechEvent – Synthetic Biology, a new "Age of Wonder"? *Biotechnology Journal*, 8(7):761–763.

Juhas et al., 2012. Juhas, M., Eberl, L., and Church, G. M. (2012). Essential genes as antimicrobial targets and cornerstones of synthetic biology. *Trends in Biotechnology*, 30(11):601–607.

Juhas et al., 2014. Juhas, M., Reuß, D. R., Zhu, B., and Commichau, F. M. (2014). Bacillus subtilis and Escherichia coli essential genes and minimal cell factories after one decade of genome engineering. *Microbiology*, 160(11):2341–2351.

Kanehisa and Goto, 2000. Kanehisa, M. and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28(1):27–30.

Kimura et al., 2016. Kimura, S., Hubbard, T. P., Davis, B. M., and Waldor, M. K. (2016). The Nucleoid Binding Protein H-NS Biases Genome-Wide Transposon Insertion Landscapes. *mBio*, 7(4):e01351–16.

Koonin, 2003. Koonin, E. V. (2003). Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nature Reviews Microbiology*, 1(2):127–136.

Krylov et al., 2003. Krylov, D. M., Wolf, Y. I., Rogozin, I. B., and Koonin, E. V. (2003). Gene Loss, Protein Sequence Divergence, Gene Dispensability, Expression Level, and

Interactivity Are Correlated in Eukaryotic Evolution. *Genome Research*, 13(10):2229–2235.

Langridge et al., 2009. Langridge, G. C., Phan, M.-D., Turner, D. J., Perkins, T. T., Parts, L., Haase, J., Charles, I., Maskell, D. J., Peters, S. E., Dougan, G., Wain, J., Parkhill, J., and Turner, A. K. (2009). Simultaneous assay of every Salmonella Typhi gene using one million transposon mutants. *Genome Research*, 19(12):2308–2316.

Lodge et al., 1988. Lodge, J. K., Weston-Hafer, K., and Berg, D. E. (1988). Transposon Tn5 target specificity: preference for insertion at G/C pairs. *Genetics*, 120(3):645–650.

Luo et al., 2014. Luo, H., Lin, Y., Gao, F., Zhang, C.-T., and Zhang, R. (2014). DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements. *Nucleic Acids Research*, 42(Database issue):D574–D580.

Marcusson et al., 2009. Marcusson, L. L., Frimodt-Møller, N., and Hughes, D. (2009). Interplay in the Selection of Fluoroquinolone Resistance and Bacterial Fitness. *PLOS Pathogens*, 5(8):e1000541.

Mushegian and Koonin, 1996. Mushegian, A. R. and Koonin, E. V. (1996). A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 93(19):10268–10273.

Nathan, 2004. Nathan, C. (2004). Antibiotics at the crossroads. *Nature*, 431(7011):899–902.

Pichoff et al., 2012. Pichoff, S., Shen, B., Sullivan, B., and Lutkenhaus, J. (2012). FtsA mutants impaired for self-interaction bypass ZipA suggesting a model in which FtsA's self-interaction competes with its ability to recruit downstream division proteins. *Molecular Microbiology*, 83(1):151–167.

Rocha and Danchin, 2004. Rocha, E. P. C. and Danchin, A. (2004). An Analysis of Determinants of Amino Acids Substitution Rates in Bacterial Proteins. *Molecular Biology and Evolution*, 21(1):108–116.

Rubin et al., 2015. Rubin, B. E., Wetmore, K. M., Price, M. N., Diamond, S., Shultzaberger, R. K., Lowe, L. C., Curtin, G., Arkin, A. P., Deutschbauer, A., and Golden, S. S. (2015). The essential gene set of a photosynthetic organism. *Proceedings of the National Academy of Sciences*, 112(48):E6634–E6643.

Seo et al., 2013. Seo, S. W., Yang, J., Min, B. E., Jang, S., Lim, J. H., Lim, H. G., Kim, S. C., Kim, S. Y., Jeong, J. H., and Jung, G. Y. (2013). Synthetic biology: Tools to design microbes for the production of chemicals and fuels. *Biotechnology Advances*, 31(6):811–817.

Silander and Ackermann, 2009. Silander, O. K. and Ackermann, M. (2009). The constancy of gene conservation across divergent bacterial orders. *BMC Research Notes*, 2:2.

Turner et al., 2015. Turner, K. H., Wessel, A. K., Palmer, G. C., Murray, J. L., and Whiteley, M. (2015). Essential genome of Pseudomonas aeruginosa in cystic fibrosis sputum. *Proceedings of the National Academy of Sciences*, 112(13):4110–4115.

van Opijnen et al., 2009. van Opijnen, T., Bodi, K. L., and Camilli, A. (2009). Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. *Nature Methods*, 6(10):767–772.

van Opijnen and Camilli, 2013. van Opijnen, T. and Camilli, A. (2013). Transposon insertion sequencing: a new tool for systems-level analysis of microorganisms. *Nature Reviews Microbiology*, 11(7):435–442.

Wilson et al., 1977. Wilson, A. C., Carlson, S. S., and White, T. J. (1977). Biochemical evolution. *Annual Review of Biochemistry*, 46:573–639.

Zhou and Rudd, 2013. Zhou, J. and Rudd, K. E. (2013). EcoGene 3.0. *Nucleic Acids Research*, 41(D1):D613–D624.