Identification of Essential Protein Domains From High-density Transposon Insertion Sequencing

A.S.M. Zisanur Rahman  
University of Manitoba

Lukas Timmerman  
University of Manitoba

Flyn Gallardo  
University of Manitoba

Silvia T. Cardona (silvia.cardona@umanitoba.ca)  
University of Manitoba

Research Article

Keywords: DUFs, Essential genes, Tn-seq, Protein Domain

DOI: https://doi.org/10.21203/rs.3.rs-589027/v1

License: ©  This work is licensed under a Creative Commons Attribution 4.0 International License.  
Read Full License
Identification of Essential Protein Domains from High-density Transposon Insertion Sequencing

A. S. M. Zisanur Rahman\textsuperscript{1}, Lukas Timmerman\textsuperscript{2}, Flyn Gallardo\textsuperscript{1}, Silvia T. Cardona\textsuperscript{1, 3*}

\textsuperscript{1}Department of Microbiology, University of Manitoba, Winnipeg, MB, Canada.
\textsuperscript{2}Department of Computer Science, University of Manitoba, Winnipeg, MB, Canada.
\textsuperscript{3}Department of Medical Microbiology & Infectious Diseases, University of Manitoba, Winnipeg, Canada.

*To whom correspondence should be addressed: silvia.cardona@umanitoba.ca
Abstract

A first clue to gene function can be obtained by examining whether a gene is required for life in certain standard conditions, that is, whether a gene is essential. In bacteria, essential genes are usually identified by high-density transposon mutagenesis followed by sequencing of insertion sites (Tn-seq). These studies assign the term “essential” to whole genes rather than the protein domain sequences that confer the essential functions. However, genes can code for multiple protein domains that evolve their functions independently. Therefore, when essential genes code for more than one protein domain, only one of them could be essential. In this study, we defined this subset of genes as “essential domain-containing” (EDC) genes. Using a Tn-seq data set built-in *Burkholderia cenocepacia* K56-2, we developed an *in silico* pipeline to identify EDC genes and the essential protein domains they encode. We found forty candidate EDC genes and demonstrated growth defect phenotypes using CRISPR interference (CRISPRi). This analysis included two knockdowns of genes encoding the protein domains of unknown function DUF2213 and DUF4148. These essential domains are conserved in more than two hundred bacterial species, including human and plant pathogens. Together, our study suggests that essentiality should be assigned to individual protein domains rather than genes, contributing to a first functional characterization of protein domains of unknown function.
Keywords
DUFs, Essential genes, Tn-seq, Protein Domain

Author contributions
ASMZR- performed the majority of the experiments and wrote the manuscript
LT- created the python script and contributed to manuscript editing
FG- created CRISPRi mutants and contributed to manuscript editing
STC- conceived the idea, supervised the work, provided financial support, and edited the final version of the manuscript.
Introduction

A first step when characterizing gene function should be asking whether a given gene encodes an essential cellular function, whether the gene is necessary for the survival of the organism. A widely accepted method to identify essential genes in bacteria is high-density transposon mutagenesis, followed by Illumina-sequencing of the transposon insertion junctions (Tn-seq)\(^1\). During Tn-seq, transposon mutant cells are pooled and grown in optimal conditions, allowing cells with a transposon insertion located in a non-essential element to survive. Cells with a transposon insertion in an essential element should be lost or depleted from the population. When transposon insertions are identified by Illumina sequencing, read counts per gene in the central 70-90% of the open reading frame (disruptive insertions) are normalized by gene length and used to predict essentiality. 5-15% sequences from the 3’ and 5’ ends are usually removed from the analysis, as insertions within the terminal regions are likely non-disruptive\(^2\)-\(^5\). While disrupted genes are regarded as “non-essential,” the method yields a list of putative essential genes as those with zero or very few mapped reads (Figure 1a and b)\(^3\).

Another step towards identifying gene function is the annotation of the protein domains encoded by genes. Protein domains are functional or structural units that can fold, evolve, and function independently. Homology-based protein domain prediction and function assignment are effective starting points for understanding protein function, even when diverse protein architectures add complexity to functional annotations\(^6,7\). While domain databases such as Pfam\(^8\) and InterPro\(^9\) aim to provide maximum sequence coverage to predict protein domain identity, approximately 30% of all domains listed in these databases (Pfam 33.1 and InterPro 81.0) are ‘domains of unknown function (DUFs).’ Single DUFs are usually predicted to span through functionally uncharacterized proteins. However, studies suggest that at least some of these proteins may contain more than one domain\(^10,11\).
While robust and comprehensive, Tn-seq studies do not consider that genes may encode for more than one protein domain. Tn-seq analysis may classify a gene as “non-essential” due to the presence of transposon insertions in a non-essential coding region, despite the gene coding for a second domain not spanning through the whole gene length that might be essential. We operationally defined this subclass of essential genes as “essential domain-containing” (EDC) genes (Figure 1c and d) and set out to identify them in a Tn-seq dataset built in *Burkholderia cenocepacia* K56-2. By analyzing biases in transposon density in genes previously identified as “non-essential”, we found 40 genes where the encoded proteins contained putative essential and non-essential domains. Using a CRISPR Interference (CRISPRi) platform we developed for *Burkholderia* 15, we experimentally confirmed growth defects, representing the loss of an essential function, in 27 EDC gene knockdowns. The identified EDC genes include ten encoding known multidomain proteins and two entirely uncharacterized genes encoding different N-terminal DUFs, demonstrating the utility of the approach. This study highlights that gene essentiality depends on the function of individual protein domains rather than entire proteins.

**Results**

**Identification of EDC Genes from Tn-seq Data**

To identify EDC genes in *B. cenocepacia* K56-2, we built a custom script that used our previous Tn-seq data 14 to select genes that i) were not previously found to be essential in *B. cenocepacia* K56-2, and ii) had an asymmetric distribution of transposon insertions (Figure 2). The script split each gene into two equal parts and selected genes with reads in only one region to identify genes with transposon insertion biases. We worked under the assumption that (i) each half could represent one functional domain and (ii) one of the domains may be essential while the other may not. We arbitrarily set the parameters “min ratio” and “min reads” to 0 and 0.14, respectively (see Material and Methods and Supplementary Figure 1). These settings
looked for genes that had zero reads at one end, while the number of reads in the non-empty end was at least 14% of that region's length. For example, if a section of a gene was 100 bp in length, it would require at least 14 reads mapped to that section to be considered non-essential. With these settings, the script produced an extensive list of 178 candidate EDC genes (Supplementary Table 1).

**Bioinformatic Analysis of the Candidate EDC Genes**

We reasoned that if EDC genes contained essential protein domains, then the essential protein domains may be encoded by essential genes in at least some other bacteria. We then searched for essential ortholog genes of the 178 candidate EDC genes by BLASTx searches against the ‘Database of Essential Genes (DEG)’ using 50% sequence alignment and 30% sequence identity as the cut-off. We found that 40 of the 178 genes had orthologs annotated as ‘essential’ in other bacterial species. We wished to interrogate the domains encoded by these 40 genes using UniProt based on InterPro domains. InterPro predicts the domain information by matching the protein or nucleic acid sequences against the member databases (collectively known as InterPro consortium) to identify ‘signatures’ associated with known domains. Thus, the InterPro prediction relies on the availability of sequence characterization and annotation. This analysis showed that from the 40 candidate EDC genes predicted to be essential by homology with other essential genes, 10 genes encoded multidomain proteins, and 7 of them were well-characterized, such as the N-terminal domain of DnaK and NusA (Figure 3a). The remaining genes were predicted to have one single annotated domain (19 genes) that did not span the whole gene-length or encoded uncharacterized proteins (11 genes) (Supplementary Table 2). All 40 genes had transposon insertions located in one half of the gene, showing that the script was able to identify genes with biased transposon insertions (Supplementary Figure 2). Taken together, these results suggest that the identified genes
Table 1: Essential genes and domains identified based on biased transposon insertions.

| K56-2 Locus Tag | Homolog J2315 Locus Tag | Product Name | Function | Reads at 5' Half | Reads at 3' Half | Identified Essential Domain |
|-----------------|-------------------------|--------------|----------|------------------|-----------------|----------------------------|
| WQ49_RS0 0050   | BCAL3469                | cell division protein FtsL | Essential cell division protein | 0 | 23 | Domain (FtsL) |
| WQ49_RS0 0770   | BCAL3328                | NUDIX hydrolase | Nucleoside-diphosphatase | 0 | 49 | Domain (Nudix hydrolase) |
| WQ49_RS0 0885   | BCAL3305                | preprotein transllocase subunit YajC | Secretase/insertase | 21 | 0 | new |
| WQ49_RS0 1035   | BCAL3270                | DnaK | Chaperone | 0 | 227 | N-terminal Domain |
| WQ49_RS0 2920   | BCAM1451                | hypothetical protein | Unknown | 43 | 0 | new |
| WQ49_RS0 3160   | BCAM1502                | hypothetical protein | Unknown | 59 | 0 | new |
| WQ49_RS0 3550   | QU43_RS6 2245           | hypothetical protein | Unknown | 33 | 0 | new |
| WQ49_RS0 3805   | BCAM1624                | MaoC family dehydratase | MaoC-like dehydratase | 46 | 0 | new |
| WQ49_RS0 4450   | BCAM1749                | hypothetical protein | Unknown | 17 | 0 | new |
| WQ49_RS0 7360   | BCAM2338                | glycosyl transferase family 1 | UDP-glycosyltransferase | 0 | 152 | Domain (Glyco_transf_28) |
| WQ49_RS0 7395   | QU43_RS6 6100           | hypothetical protein | Unknown | 0 | 58 | new |
| WQ49_RS0 9185   | BCAS0417                | cytochrome biogenesis protein CcdA | electron transfer | 0 | 38 | new |
| WQ49_RS1 0495   | BCAS0158                | hypothetical protein | Unknown | 0 | 34 | Domain (DUF4148) |
| WQ49_RS1 1915   | BCAL0324                | TatB | protein transmembrane transporter | 0 | 57 | Domain (TatA_B_E) |
| WQ49_RS1 2045   | BCAL0298                | thiamine biosynthesis protein ThiS | thiamine biosynthesis protein ThiS | 0 | 50 | Domain (ThiS) |
| Accession   | Gene Name     | Description                                      | Domain/Region         |
|-------------|---------------|--------------------------------------------------|-----------------------|
| WQ49_RS1 2280 | BCAL0250      | 50S ribosomal protein L18                        | structural constituent of ribosome 0 65 Domain (Ribosomal_L18p) |
| WQ49_RS1 2305 | BCAL0245      | RpIX                                            | structural constituent of ribosome 20 0 Domain (L24-Pfam) |
| WQ49_RS1 2315 | BCAL0243      | 30S ribosomal protein S17                       | structural constituent of ribosome 0 64 new |
| WQ49_RS1 2365 | BCAL0233      | RpsJ                                            | structural constituent of ribosome 0 25 new |
| WQ49_RS1 6145 | BCAM1066      | hypothetical protein                            | Unknown 0 425 Domain (DUF2213) |
| WQ49_RS1 8705 | BCAM0549      | molecular chaperone GroES                       | Chaperone 0 21 Domain (Cpn10) |
| WQ49_RS2 2170 | BCAM2699      | alpha/beta hydrolase                            | Putative hydrolase 120 0 Domain (Abhydrolase_3) |
| WQ49_RS2 3945 | BCAL0558      | Cca                                             | 3'-cytidine-cytidine-tRNA adenylyltransferase 0 79 Domain (PolyA Polymerase)/Domain (Binding) |
| WQ49_RS2 4070 | BCAL0585      | hypothetical protein                            | Unknown 0 23 new |
| WQ49_RS2 5525 | BCAL0878      | FmdB family transcriptional regulator           | Regulatory activity 0 30 Domain (CxxC_CXXC_SSSS) |
| WQ49_RS2 5680 | BCAL0909      | 16S rRNA maturation RNase YbeY                  | Endoribonuclease activity 68 0 Domain (UPF0054) |
| WQ49_RS2 6625 | BCAL2715      | RpmG                                            | structural constituent of ribosome 0 31 Domain (Ribosomal_L33) |
| WQ49_RS2 7920 | BCAL2334      | NADH-quinone oxidoreductase subunit K           | NADH dehydrogenase 0 21 Domain (Oxidored_q2) |
| WQ49_RS2 8635 | BCAL2199      | Fe-S cluster assembly transcriptional regulator IscR | DNA-binding transcription factor 39 0 Domain (Rrf2) |
| WQ49_RS2 9230 | BCAL2091      | 30S ribosomal protein S2                        | structural constituent of ribosome 0 86 Domain (Ribosomal_S2) |
CRISPRi Knockdowns of EDC Genes Show Growth Defects

To phenotypically characterize the effect of knocking down EDC genes, we used CRISPR interference or CRISPRi \(^{16}\) to create knockdown mutants of the genes of interest. CRISPRi comprises a chromosomally integrated dCas9 under the control of a rhamnose-inducible promoter and plasmid-borne sgRNA driven by a constitutively active synthetic promoter, \(P_{J23119}\) \(^{16}\). Simultaneous expression of dCas9 and a target-specific sgRNA allows the dCas9 to bind the target DNA region and, thus, sterically interfere with transcription by RNA polymerase \(^{15,16}\). To inhibit the expression of the candidate genes, we designed two sgRNAs against each of the candidate genes targeting the start codon and adjacent region on the non-template strand (Supplementary Figure 3a and c). For phenotypic characterization, we grew
the cells in LB with and without rhamnose. Upon induction of dCas9 with rhamnose, 27 out of the 40 candidate genes showed at least 25% growth inhibition relative to the uninduced condition (Supplementary Figure 3b-d).

**DUF2213 and DUF4148 Appear to be Essential Domains**

The presence of DUFs is a common feature of hypothetical or uncharacterized proteins. To initiate functional characterization of DUFs, we focused on two genes containing DUF-coding sequences, which their respective CRISPRi mutants demonstrated a conditional growth defect (Figure 3b). WQ49_RS16145 (BCAM1066) and WQ49_RS10495 (BCAS0158) contain DUF2213 (Pfam accession PF09979) and DUF4148 (Pfam accession PF13663), respectively at the N-terminal end of the proteins (Figure 3b). BLAST searches of BCAM1066 and BCAS0158 genes as a query against the DEG showed that BCAM1066 (WQ49_RS16145) had 30% sequence similarity with lysK (B8GXH3) from Caulobacter crescentus, and BCAS0158 (WQ49_RS10495) had a 52% sequence identity with a predicted amino acid permease (BPSS1112) from Burkholderia pseudomallei K96243 (data not shown). Mining of the Pfam database (https://pfam.xfam.org/) showed that these DUFs are well conserved across the bacterial species: DUF2213 is present in 209 bacterial species, including bacterial pathogens (Acinetobacter baumannii, Enterobacter cloacae, Haemophilus influenzae, Burkholderia cepacia, Shigella flexneri), plant pathogens (Agrobacterium tumefaciens), and biotechnologically relevant species (Pseudomonas putida) (Figure 4a and Supplementary Table 5). DUF4148 is found in 204 bacterial species, primarily in Burkholderia species (i.e. Burkholderia cepacia, Burkholderia mallei, Burkholderia vietnamiensis) and plant pathogens such as Ralstonia solanacearum (Figure 4b and Supplementary Table 5). DUF2213 is also present in many phage-related proteins (Figure 4a). Eight unique domain architectures were observed for proteins containing DUF2213 and five for DUF4148 (Figure 4c-d). DUF2213 is associated with another essential domain PF00293, a NUDIX hydrolase (Figure 4c). In other
proteins, DUF2213 is associated with the LPD3 domain (PF18798) and DUF1073 (PF06381) which is also conserved across bacterial species (Figure 4c). On the other hand, Pfam analysis of DUF4148 shows that DUF4148 differs in domain length among species and is associated with the Pfam domain PF00144, known to confer resistance against β-lactams (Figure 4d). Nonetheless, the N-terminus was highly conserved, suggesting it is functionally significant. The Pfam-based analysis of species distribution also revealed that DUF2213 is present in six eukaryotic species (five metazoans and one fungal species), whereas DUF4148 is present in five eukaryotic species (three viridiplantae species and two metazoan species). The widespread distribution of these DUFs indicates the functional importance of these essential domains, creating an impetus for further characterization.

Discussion

A first step in the functional characterization of proteins is performed through protein depletion and growth phenotype characterization. As multidomain proteins can perform multiple functions driven by the activity of their individual domains, the function assigned to a gene product could indeed correspond to one of its domains and not to the whole protein. That is the case of essential genes identified by Tn-seq. In standard Tn-seq analysis the condition of essentiality is assigned to genes and not to domains, resulting in incorrect classification of many essential genes as non-essential. Rather, essentiality assignment pipeline should be revised to analyze the essentiality of individual protein domains. Indeed, essentiality can be assigned to individual domains of a multidomain protein rather than the entire protein. In this work, we defined as essential-domain-containing (EDC) genes those genes that encode more than one protein domain, with one of the domains coding for an essential function. By analyzing a Tn-seq dataset for transposon insertion biases, we show that standard Tn-seq analysis pipelines may miss EDC genes, whose detection often requires either manual curation or additional considerations.
We validated our approach by identifying previously characterized multidomain essential proteins in which the essential function is assigned to one single domain. For instance, our analysis of biases in the Tn-seq dataset showed that the N-terminal domain of NusA is sufficient to mediate the essential function, in agreement with previous work. Similarly, the \textit{B. cenocepacia} K56-2 \textit{dnaK} gene was previously defined as non-essential; however, we found that the Tn-seq reads mapped onto \textit{dnaK} were biased toward the C-terminal domain (CTD), suggesting that only the NTD is necessary for its essential function. (Figure 3b, \textbf{Supplementary Figure 2}). DnaK is a multidomain protein and a master regulator of the chaperone network. DnaK comprises an N-terminal ATPase domain (NTD) and a C-terminal substrate-binding domain (CTD). Perturbations either within the NTD that leads to the abrogation of the ATPase activity or within the conserved linker peptide that impairs the interdomain mechanistic interaction abrogate the \textit{in vivo} activity of DnaK.

While 14 EDC genes that demonstrated a growth defect when knocked down code for proteins annotated to have a single domain, none of these domains span the entire gene, and transposon insertions are only mapped to the annotated domain (\textbf{Supplementary Figure 2}). Thus, it is possible that the remaining regions code for novel domains that perform the essential biological functions independently of the adjacent sequences. Indeed, multidomain proteins that are involved in direct protein-protein interactions are more often detected as essential than proteins with a single domain, hinting towards the functional contribution of individual domains within a protein complex.

We demonstrated a conditional growth defect in 27 out of 40 CRISPRi mutants of EDC genes. It remains a possibility that the sgRNAs designed for CRISPRi-mediated gene silencing of the remaining 13 genes were not efficient in target binding, thus yielding no growth defect. CRISPRi is more effective in blocking transcription initiation than elongation, and is the most efficient in silencing gene expression when promoter regions are targeted with gRNAs.
However, as promoter regions for *B. cenocepacia* genomes remained largely unannotated we targeted translation start sites. It remains to be investigated whether targeting the promoter region to block the transcription initiation rather than elongation might yield conditional a growth phenotype in the remaining 13 genes.

A large portion of the protein domains that lack functional assignment can be grouped within the DUF category. DUFs are members of ever-increasing uncharacterized protein families; they are the object of experimental and computational efforts towards their functional characterization. Determining if a DUF is essential is among the first steps in functional characterization. In this study, we focused on two EDC genes that encode putative essential DUFs: DUF2213 and DUF4148. Both domains have a high degree of conservation across diverse phyla, which highlights their biological relevance. DUF2213, a phage-associated domain (PF09979), is well distributed across bacteria and phages. Interestingly, we found that DUF4148 (PF13663) is putatively essential and associated with β-lactamase (PF00144) (Figure 4).

In summary, our study identified 27 EDC genes whose knockdown produced a growth defect, highlighting the essential nature of one of their protein domains. By leveraging a Tn-Seq dataset in *B. cenocepacia* K56-2, we demonstrate that the essential nature of protein-coding genes is a function of the individual protein domains they encode. We propose that determining essentiality of a domain of unknown function should be the first step in the process to define their function.

**Methods**

**Bacterial Strains and Growth Conditions**

The list of bacterial strains and plasmids used in this study is provided in Supplementary Table 3. Bacterial strains were grown in LB-Lennox medium (Difco) at 37°C. *E. coli* strain MM290 carrying the helper plasmid pRK2013 was selected in kanamycin
40µg/mL (Fisher Scientific). Donor strains of *E. coli* DH5α and *B. cenocepacia* K56-2 carrying the sgRNA plasmids were selected in trimethoprim 50µg/mL and 100µg/mL (Sigma), respectively.

**Identification of EDC Genes from Tn-Seq Dataset**

Candidate EDC genes were identified with a custom python script using the Tn-seq dataset. The script analyzed every gene previously classified as “non-essential” by splitting it into two equal halves and counting the number of reads mapped to each half-gene. The script then used the “min ratio” and “min reads” as filtering criteria to call EDC genes. “Min ratio” was defined as the desired ratio of reads between the halves of the gene. “Min reads” was defined as the minimum number of reads in the non-empty end that is equal to a 14% of that half's length. Min reads was set to 0.14, while min ratio was set as 0. For each gene, 10% from each end of the gene was discarded from the analysis. The parameters can be changed to yield either more stringent or more general results. The script is available at https://github.com/cardonalab/EssentialDomains

**Bioinformatic Analysis**

Orthologous essential genes were identified using BLASTx against DEG 15. Multidomain information was fetched from the UniProt database based on Pfam and InterPro domain features. DUF containing genes were characterized using the Pfam tool available on the Pfam website (https://pfam.xfam.org/). Domain sequences were retrieved in FASTA format from the Pfam database and aligned by Clustal Ω. Maximum-likelihood phylogenetic trees were generated with MEGA-X using a Jones-Taylor-Thornton (JTT)-based model applying 100 bootstrap values. Phylogenetic trees were visualized, edited and taxonomic labels were assigned using Interactive Tree Of Life (i-TOL). Bootstrap values are represented on a scale of 0 to 1. Taxonomic annotations were labelled based on the NCBI taxonomy database using UniProt identifiers.
CRISPRi mutants were of the EDC genes were created as previously described \(^{16}\). Briefly, pSCB2-sgRNAv2, a modified plasmid from pSCB2-sgRNA \(^{16}\), was used as the template for inverse PCR to insert 20bp target-specific sgRNA sequence. Inverse PCR was performed using Q5 high-fidelity polymerase (NEB), forward primers with individual sgRNAs as 5’ tail, and 1092 as the reverse primer. The resultant fragments were ligated to create circular plasmids by incubating 0.5\(\mu\)L of the respective PCR products with quick ligation buffer (NEB), 0.25 \(\mu\)L DpnI, 0.25 \(\mu\)L T4 polynucleotide kinase (NEB), and 0.25 \(\mu\)L T4 ligase (NEB) for 30 minutes at 37\(^\circ\)C. Resultant plasmids were transformed into *E. coli* DH5\(\alpha\), recovered for 2h and selected in LB supplemented with trimethoprim 50\(\mu\)g/mL (Sigma). The transformants were further confirmed by colony PCR using primers 1409 and 848. *E. coli* strains carrying the sgRNA plasmids were used as donors, and *E. coli* MM290/pRK2013 as the helper for triparental mating to introduce the sgRNA plasmids into *B. cenocepacia* K56-2 containing the chromosomally integrated dCas9 under the control of a rhamnose inducible promoter, as described previously \(^{38}\). Trimethoprim resistant colonies (100\(\mu\)g/mL) were selected and screened by colony PCR using the primers 1409 and 848. The list of all the primers used in this study is provided in Supplementary Table 4.

**Conditional Growth Phenotype Analysis of the CRISPRi Mutants**

To determine the conditional growth phenotype of the candidate genes, overnight cultures of the CRISPRi mutants were back diluted to \(\text{OD}_{600\text{nm}}\) 0.01. The cultures were grown at 37\(^\circ\)C for 20-24 hours with continuous shaking in a 384-well plate containing LB broth supplemented with trimethoprim 100\(\mu\)g/mL and with/without 1\% rhamnose. \(\text{OD}_{600\text{nm}}\) readings were taken at 1 hour intervals using BioTek Synergy 2 microplate reader.

**Acknowledgments**
This work was supported by grants from the Canadian Institutes of Health Research (CIHR), Cystic Fibrosis Foundation, Cystic Fibrosis Canada to STC; ASMZR was supported by a University of Manitoba Graduate Fellowship (UMGF). The authors thank Dr. Georg Hausner, Andrew Hogan, Dustin Maydaniuk and rest of the Cardona lab members for critically reading the manuscript.

**Conflict of Interest Statement**

The authors declare no conflict of interest.

**References**

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

2. Akerley, B. J. *et al.* Systematic identification of essential genes by in vitro mariner mutagenesis. *Proc Natl Acad Sci U A* **95**, 8927–8932 (1998).

3. Chao, M. C., Abel, S., Davis, B. M. & Waldor, M. K. The design and analysis of transposon insertion sequencing experiments. *Nat. Rev. Microbiol.* **14**, 119–128 (2016).

4. Langridge, G. C. *et al.* Simultaneous assay of every *Salmonella Typhi* gene using one million transposon mutants. *Genome Res.* **19**, 2308–2316 (2009).

5. Shields, R. C., Zeng, L., Culp, D. J. & Burne, R. A. Genomewide Identification of Essential Genes and Fitness Determinants of Streptococcus mutans UA159. *mSphere* **3**, e00031-18 (2018).

6. Forslund, S. K., Kaduk, M. & Sonnhammer, E. L. L. Evolution of Protein Domain Architectures. in *Evolutionary Genomics* (ed. Anisimova, M.) vol. 1910 469–504 (Springer New York, 2019).
7. Schnoes, A. M., Brown, S. D., Dodevski, I. & Babbitt, P. C. Annotation Error in Public Databases: Misannotation of Molecular Function in Enzyme Superfamilies. *PLoS Comput. Biol.* **5**, e1000605 (2009).

8. El-Gebali, S. *et al.* The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**, D427–D432 (2019).

9. Mitchell, A. L. *et al.* InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res.* **47**, D351–D360 (2019).

10. Bateman, A., Coggill, P. & Finn, R. D. DUFs: families in search of function. *Acta Crystallograph. Sect. F Struct. Biol. Cryst. Commun.* **66**, 1148–1152 (2010).

11. Goodacre, N. F., Gerloff, D. L. & Uetz, P. Protein Domains of Unknown Function Are Essential in Bacteria. *mBio* **5**, e00744-13 (2014).

12. Lluch-Senar, M. *et al.* Defining a minimal cell: essentiality of small ORFs and ncRNAs in a genome-reduced bacterium. *Mol. Syst. Biol.* **11**, 780 (2015).

13. Lu, Y. *et al.* A novel essential domain perspective for exploring gene essentiality. *Bioinformatics* **31**, 2921–2929 (2015).

14. Gislason, A. S., Turner, K., Domaratzki, M. & Cardona, S. T. Comparative analysis of the *Burkholderia cenocepacia* K56-2 essential genome reveals cell envelope functions that are uniquely required for survival in species of the genus *Burkholderia*. *Microb. Genomics* **3**, e000140 (2017).

15. Qi, L. S. *et al.* Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* **152**, 1173–1183 (2013).

16. Hogan, A. M., Rahman, A. S. M. Z., Lightly, T. J. & Cardona, S. T. A Broad-Host-Range CRISPRi Toolkit for Silencing Gene Expression in Burkholderia. *ACS Synth. Biol.* **8**, 2372–2384 (2019).
17. Luo, H. et al. DEG 15, an update of the Database of Essential Genes that includes built-in analysis tools. Nucleic Acids Res. 49, D677–D686 (2021).

18. The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 47, D506–D515 (2019).

19. Gao, M., Glenn, A. E., Blacutt, A. A. & Gold, S. E. Fungal Lactamases: Their Occurrence and Function. Front. Microbiol. 8, 1775 (2017).

20. Kanaan, S. P., Huang, C., Wuchty, S., Chen, D. Z. & Izaguirre, J. A. Inferring Protein–Protein Interactions from Multiple Protein Domain Combinations. in Computational Systems Biology (eds. Ireton, R., Montgomery, K., Bumgarner, R., Samudrala, R. & McDermott, J.) vol. 541 43–59 (Humana Press, 2009).

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

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

23. Qayyum, M. Z., Dey, D. & Sen, R. Transcription Elongation Factor NusA Is a General Antagonist of Rho-dependent Termination in Escherichia coli. J. Biol. Chem. 291, 8090–8108 (2016).

24. Ha, K. S., Toulokhonov, I., Vassylyev, D. G. & Landick, R. The NusA N-Terminal Domain Is Necessary and Sufficient for Enhancement of Transcriptional Pausing via Interaction with the RNA Exit Channel of RNA Polymerase. J. Mol. Biol. 401, 708–725 (2010).

25. Wu, C.-C., Naveen, V., Chien, C.-H., Chang, Y.-W. & Hsiao, C.-D. Crystal Structure of DnaK Protein Complexed with Nucleotide Exchange Factor GrpE in DnaK Chaperone
26. Barthel, T. K., Zhang, J. & Walker, G. C. ATPase-Defective Derivatives of Escherichia
coliDnaK That Behave Differently with Respect to ATP-Induced Conformational Change
and Peptide Release. *J. Bacteriol.* **183**, 5482–5490 (2001).

27. Vogel, M., Mayer, M. P. & Bukau, B. Allosteric Regulation of Hsp70 Chaperones
Involves a Conserved Interdomain Linker. *J. Biol. Chem.* **281**, 38705–38711 (2006).

28. Bikard, D. *et al.* Programmable repression and activation of bacterial gene expression
using an engineered CRISPR-Cas system. *Nucleic Acids Res.* **41**, 7429–7437 (2013).

29. Hawkins, J. S., Wong, S., Peters, J. M., Almeida, R. & Qi, L. S. Targeted Transcriptional
Repression in Bacteria Using CRISPR Interference (CRISPRi). *Methods Mol. Biol.
Clifton NJ* **1311**, 349–362 (2015).

30. Vigouroux, A., Oldewurtel, E., Cui, L., Bikard, D. & Teeffelen, S. van. Tuning dCas9’s
ability to block transcription enables robust, noiseless knockdown of bacterial genes.
*Mol. Syst. Biol.* **14**, e7899 (2018).

31. Bastard, K. *et al.* Revealing the hidden functional diversity of an enzyme family. *Nat.
Chem. Biol.* **10**, 42–49 (2014).

32. Dessailly, B. H. *et al.* PSI-2: Structural Genomics to Cover Protein Domain Family
Space. *Structure* **17**, 869–881 (2009).

33. Zhang, X. *et al.* Assignment of function to a domain of unknown function: DUF1537 is a
new kinase family in catabolic pathways for acid sugars. *Proc. Natl. Acad. Sci.* **113**,
E4161–E4169 (2016).

34. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence
alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).
35. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol. Biol. Evol.* **35**, 1547–1549 (2018).

36. Jones, D. T., Taylor, W. R. & Thornton, J. M. The rapid generation of mutation data matrices from protein sequences. *Bioinformatics* **8**, 275–282 (1992).

37. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* **47**, W256–W259 (2019).

38. Hogan, A. M. *et al*. Competitive Fitness of Essential Gene Knockdowns Reveals a Broad-Spectrum Antibacterial Inhibitor of the Cell Division Protein FtsZ. *Antimicrob. Agents Chemother.* **62**, e01231-18 (2018).
While disrupted genes are regarded as “non-essential,” the method yields a list of putative essential genes as those with zero or very few mapped reads (1a and 1b). We operationally defined this subclass of essential genes as “essential domain-containing” (EDC) genes (1c and 1d).
Figure 2

Asymmetric distribution of transposon insertions
Figure 3

DUF2213 and DUF4148 Appear to be Essential Domains
DUF2213 is also present in many phage-related proteins (4a). DUF4148 is found in 204 bacterial species, primarily in Burkholderia species (4b). Eight unique domain architectures were observed for proteins containing DUF2213 and five for DUF4148 (4c-d).

Figure 4

DUF2213 is also present in many phage-related proteins (4a). DUF4148 is found in 204 bacterial species, primarily in Burkholderia species (4b). Eight unique domain architectures were observed for proteins containing DUF2213 and five for DUF4148 (4c-d).

Supplementary Files
This is a list of supplementary files associated with this preprint. Click to download.

- SupplementalFiles.pdf
- SupplementaryTable1.csv
- SupplementaryTable2.xlsx
- SupplementaryTable5.xlsx