cture(0,0)(-30,0)10 (-30,-5)(0,1)10 (-35,0)(1,0)30 (0,30)10 (-5,30)(1,0)10 (0,35)(0,-1)30 picturepicture(0,0)(30,0)10 (30,-5)(0,1)10 (35,0)(-1,0)30 (0,30

Published online DD MM YYYY

Journal, YYYY, Vol. xx, No. xx **1–14**
doi:foo/bar/gkn000

StORF-Reporter: Finding Genes between Genes

Nicholas J. Dimonaco [1,*], Wayne Aubrey [2,*], Kim Kenobi [3,*], Amanda Clare [2,*], and Christopher J. Creevey [4,]**

[1]Institute of Biological, Environmental and Rural Sciences, Aberystwyth University, Aberystwyth, SY23 3PD, Wales, UK
[2]Department of Computer Science, Aberystwyth University, Aberystwyth, SY23 3DB, Wales, UK
[3]Department of Mathematics, Aberystwyth University, Aberystwyth, SY23 3BZ, Wales, UK
[4]School of Biological Sciences, Queen's University Belfast, Belfast, BT7 1NN, Northern Ireland, UK

ABSTRACT

Large regions of prokaryotic genomes are currently without any annotation, in part due to well-established limitations of annotation tools. We present StORF-Reporter, a tool that takes as input an annotated genome and returns missed CDS genes from the unannotated regions. StORF-Reporter consists of two parts. The first part begins with the extraction of unannotated regions from an annotated genome. Next, Stop-ORFs (StORFs) are identified in these unannotated regions. StORFs are Open Reading Frames that are delimited by stop codons.

We show that this methodology recovers complete coding sequences (with/without similarity to known genes) which were missing from both canonical and novel genome annotations. StORF-Reporter recovered sequences that exhibited high levels of sequence identity to proteins in the SwissProt database and the proteomes of the genome they were identified from (gene-duplicates). We inspected in detail the results from the genomes of six model organisms, the pangenome of Escherichia coli, and a further 6,223 annotated prokaryotic genomes of 179 genera from the Ensembl Bacteria database. StORF-Reporter was able to extend the core, soft-core and accessory gene-collections, and identify novel gene families and families which were extended into additional genera, not previously identified in the canonical annotations. Many of the gene families these sequences form are routinely misreported or completely omitted by state-of-the-art annotation methods.

INTRODUCTION

Prokaryotic genomes are often observed with high levels of genome density, with the compact nature of their genes, both coding and non-coding, resulting in little 'extraneous' DNA (1). Much of this extraneous DNA is often referred to as intergenic, islands of DNA between functional elements such as genes, promoters or other functionally and structurally important elements. In eukaryota studies, the relative proportion of a genome reported to be intergenic has been linked to the biological complexity of the organism (2). However, due to the much reduced proportion of their genomes reported as intergenic, prokaryotic intergenic regions are often assumed not to be functionally important, even though studies have reported finding putative coding and non-coding genes (both complete and pseudogenised) in them (3, 4, 5). Therefore, in this work, we will be using the term *unannotated region* (UR) to describe the regions of a genome which are yet to be characterised. This work specifically investigates their potential to harbor yet-to-be discovered CoDing Sequence (CDS) genes.

There is a distinct paucity in the analysis of prokaryote URs and what they may contain (4, 6). Those studies which have investigated the conservation and selection of putative UR sites have shown that there is "consistent evidence of strong purifying selection on IGRs [URs]" (7). These regions should now be investigated to determine whether they may contain genes missed by contemporary prediction methods.

Unfortunately, the methods employed to study these regions often rely heavily on sequence alignments of the entire UR sequence to known proteins deposited in databases. This practice is inefficient and more importantly inaccurate, as the part of UR which may exhibit similarity to a known protein may be overshadowed by the surrounding DNA 'noise' and requires the homologous sequence to have been previously identified. Furthermore, similarity searching to previously identified genes is unlikely to work if there is a systematic reason behind why such genes are routinely missed by contemporary genome annotation, and therefore they are unlikely to have representatives in such databases.

To elucidate these regions and overcome the weaknesses and biases in genome annotation, many of which continue to be outlined in the literature (8), we have developed a novel methodology to supplement genome annotation. This approach, named 'StORF-Reporter', is performed in two stages. First we extract URs from an annotated genome (see Figure 1. Then we identify and filter StORFs (Open Reading Frames that are delimited by two stop codons) from these extracted URs.

*To whom correspondence should be addressed. Email: nicholas@dimonaco.co.uk

© YYYY The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

cture(0,0)(-30,0)10 (-30,-5)(0,1)10 (-35,0)(1,0)30 (0,-30)10 (-5,-30)(1,0)10 (0,-35)(0,1)30 picturepicture(0,0)(30,0)10 (30,-5)(0,1)10 (35,0)(-1,0)30 (0,

"output" — 2022/3/31 — 20:58 — page 2 — #2
cture(0,0)(-30,0)10 (-30,-5)(0,1)10 (-35,0)(1,0)30 (0,30)10 (-5,30)(1,0)10 (0,35)(0,-1)30 picturepicture(0,0)(30,0)10 (30,-5)(0,1)10 (35,0)(-1,0)30 (0,30

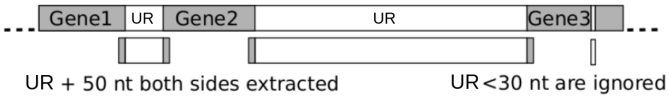

Figure 1. Visual representation of how Unannotated Regions (URs) are selected for extraction. URs that are less than 30 nt are not extracted. URs are extracted with additional 50 nt on their 5' and 3' ends to allow for overlapping genes.

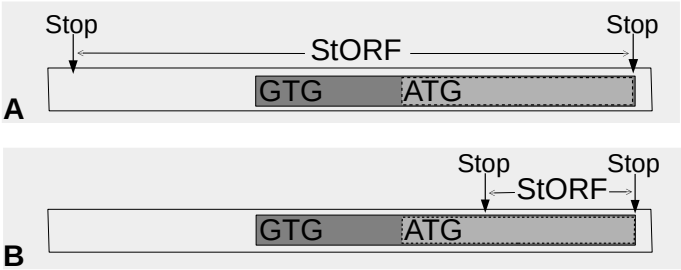

Figure 2. Visual representation of a StORF and how it can capture multiple potential start codons for a single gene in an unannotated region. Image A depicts a StORF capturing the two possible start positions/codons for a CDS gene and image B shows how a StORF can comprise of only a partial segment of a gene if that gene either recodes a canonical stop codon or has had an in-frame stop codon mutation.

An example of a StORF, extracted from a previously annotated genome, can be seen in Figure 2, potentially capturing multiple possible start codons. The reporting of stop codons, instead of predicting a specific start codon, removes a number of complexities, such as the Shine–Dalgarno motif (9), promoter regions (10), ribosomal binding sites, and operons (11). Additionally, the existence of a StORF with a gene-like length is unlikely in the out of frame region of another, unreported gene. This is due to out of frame stop codons having been observed as a widespread phenomenon among prokaryotes and theorised to carry functional significance of being selected in the course of genome evolution to act against unintended frameshift (12). Thus, the length of a StORF is likely to be a useful factor in their discovery and removal of false positives.

Current prokaryotic annotation tools still struggle to determine the correct start codon, resulting in incomplete or missed gene detection (8). The reporting of StORFs, therefore, can also remove the potential for an incorrect or alternative start codon selection to truncate the start of a protein product, as the entire coding potential of the CDS is encapsulated by the StORF. This is important as while the canonical 'ATG' codon is most-often used to initiate approximately 80% of prokaryotic CDS genes, many species and cross-species gene families have been shown to use very different start codon profiles (13). Furthermore, there is growing evidence of the existence of "organisms that adjust their genetic code in response to changing environments", thus making traditional start codon identification ever more complex (14). Although stop codon usage in prokaryote CDS genes has also been linked to a number of genomic features such as GC content, there is still much we do not yet know and many theories concerning their selection are still being posited by researchers (15, 16). However, unlike start codons, the three canonical (TGA,TAG,TAA) stop codons are the terminus for almost all CDS genes across almost all species, meaning that identification can be based solely on their presence.

The existence of multiple, and often contrary, definitions of an ORF has been recently raised (17) with the positives and negatives of the different definitions discussed. Interestingly, while most prokaryotic annotation tools will annotate from start codon to stop codon, the Sequence Ontology (18) defines an ORF as "The in-frame interval between the stop codons of a reading frame which when read as sequential triplets, has the potential of encoding a sequential string of amino acids.". StORFs therefore use the same interpretation as the Sequence Ontology but for clarity and due to their specific use case, will continued to be named StORFs hereafter.

This work proposes the StORF-Reporter methodology specifically as an additional step to be performed on the output of contemporary prokaryotic genome annotation. The combination of these StORFs with the existing annotations (for example, from Ensembl, or from a tool such as Prodigal (19)) can be referred to as 'GFF+' and represent a refining of current annotations as opposed to the traditional mantra of reannotation. We investigate StORFs identified from three sets of genomes. Firstly, we look at six bacterial model organisms which exhibit many URs of lengths long enough to contain full length CDS genes that can expand the narrative of these well studies genome gene collections. While a number of StORF sequences recovered were found to be duplicates or at least shared a high level of sequence identity to genes already annotated elsewhere in the genome in the canonical genomic annotations from Ensembl (20), others exhibited high levels of sequence identity to proteins in the Swiss-Prot database (21).

We next investigate the presence of conserved StORFs across the Escherichia coli (*E. coli*) pangenome. As pangenomes consist of continuously expanding and contracting species-wide gene sets which are dynamically exchanged in response to exogenous pressures, missing annotations can have detrimental impacts on the utility of their study. Additionally, as shown in the literature, the types of genes most often missing from genome annotations are those which are transient or non-core accessory genes, which are less common or missing in genomic databases (8, 22). Therefore, StORFs were extracted from a large collection of *E. coli* genomes and analysis of these StORFs showed that they formed and redefined a number of core, soft-core and accessory gene families. The resulting novel gene families and known families supplemented with additional strains only identified through StORF sequences, has essentially rewritten the pangenome gene-collection of *E. coli*.

Lastly, we apply the same methodology to 6,223 prokaryotic genomes from Ensembl Bacteria. We find highly conserved StORFs across different genera in both 'StORF-Only' gene families and as additional sequences, expanding the Ensembl annotated gene families. The high levels of sequence conservation observed between genera suggests that these StORF sequences are likely to be functional genes that must now be added to the canonical annotations.

cture(0,0)(-30,0)10 (-30,-5)(0,1)10 (-35,0)(1,0)30 (0,-30)10 (-5,-30)(1,0)10 (0,-35)(0,1)30 picturepicture(0,0)(30,0)10 (30,-5)(0,1)10 (35,0)(-1,0)30 (0,

cture(0,0)(-30,0)10 (-30,-5)(0,1)10 (-35,0)(1,0)30 (0,30)10 (-5,30)(1,0)10 (0,35)(0,-1)30 picturepicture(0,0)(30,0)10 (30,-5)(0,1)10 (35,0)(-1,0)30 (0,30

MATERIALS AND METHODS

Datasets

The canonical annotation and sequence data for 44,048 genomes of 8,244 prokaryotic species, which included 43,347 bacterial genomes and 493 archaeal genomes, were acquired from the 46th release of Ensembl Bacteria (20). For each genome, two data files were downloaded; the complete DNA sequence (*∗_dna.toplevel.fa*) and the GFF (Generic Feature Format) file (*∗.gff3*) containing the position information for each gene (coding and non-coding). The genomic elements (including both coding and non-coding genes) presented in the Ensembl Bacteria GFF annotations were taken as reference annotations for this study. Fragmented genomes were removed by filtering out those with more than 5 contigs (not including plasmids). Next, genera which had less than 5 genomes were removed. This resulted in 6,223 genomes (see Supplementary Table 1). Out of this 6,223, the six model organisms; *Bacillus subtilis (B. subtilis)* BEST7003 strain, *Caulobacter crescentus (C. crescentus)* CB15 strain, *Escherichia coli (E. coli)* K-12 ER3413 strain, *Mycoplasma genitalium (M. genitalium)* G37 strain, *Pseudomonas fluorescens (P. fluorescens)* UK4 strain, *Staphylococcus aureus (S. aureus)* 502A strain, shown in Supplementary Table 2, were selected for this study. Additionally, *Escherichia_coli* was selected to undergo a pangenomic analysis of its Ensembl annotated genes and extracted StORFs. As many of the Ensembl Bacteria genomes were unlabeled at the species level, to make sure the correct genomes were extracted, only those labeled specifically as *'Escherichia_coli'* were extracted. This resulted in a group of 219 *E. coli* genomes from the original filtered group of 6,223 (see https://github.com/NickJD/StORF-Reporter for further detail).

StORF-Reporter

StORF-Reporter was developed as an additional supplementary step to be performed after a 'traditional' genome annotation had been undertaken. It is comprised of two distinct parts which can be used independently or together as they were in this study: UR_Extractor(.py) and StORF_Finder(.py). The processes of both are described in the following sections and the code is available at https://github.com/NickJD/StORF-Reporter.

Unannotated Region – Extractor (UR-Extractor) To facilitate the extraction of URs from different prokaryotic genomes (and the various interpretations of the GFF format), UR_Extractor was developed. It allows user-defined genomic features (coding and non-coding genomic elements) to be used to identify the boundaries of genomic elements. Written in Python3 (23) and using the feature coordinates presented in GFF3 format, regions of DNA without annotation were extracted from the provided DNA sequence file. GFF3 and FASTA files with the details of the sequence and loci for each UR were reported for further analysis (see Supplementary Section 3.1 for command line menu of UR_Extractor.py).

When recovering URs from a genome it is important to consider that unannotated CDSs (to be captured by StORF-Finder) may overlap with annotated genes (coding and non-coding) in either the same or an alternative frame (24). Therefore, it was necessary to determine the extent to which the URs should be extended into the annotated regions. The 6,223 genomes from Ensembl Bacteria, including the six model organisms (MOs) in Supplementary Table 2 were studied to identify representative parameters for the extraction of URs across this large set of Ensembl Bacteria genomes. The median gene overlap observed for each of the six genomes was 3 nt and the 75th percentile was below 50 nt, as can be seen in the left half (A) of Figure 3. Furthermore, the median distance from a gene's canonical start codon to the nearest in-frame upstream stop codon (representing the untranslated region of a StORF) was 39 nt and in 77.4% of the CDS genes it was less than or equal to 100 nt (see the right half (B) of Figure 3). The lengths of overlap between genes (both coding and non-coding) across the 6,223 prokaryotic genomes were investigated and as can be seen in Supplementary Figure 1, the gene overlap lengths observed for the 6 genomes are similar to the majority of the Ensembl annotated genomes. These findings further illustrate that these model bacterial genomes are relatively compact and demonstrate that StORFs most often contain very little extraneous upstream DNA. To account for the gene and StORF features observed, URs were only extracted if they were at least 30 nt long but were extended by 50 nt at both ends (into annotated genes) by the UR Extractor tool (see Figure 1), totaling a minimum length of 130 nt. This extension of 50 nt at both the 5-prime and 3-prime ends, was to allow for the capture of the parts of StORFs which may overlap with existing gene annotations. This process took into account the untranslated region between the start of the StORF and putative start codon of the protein sequence it captured and was necessary as the direction of a potential StORF is not known in advance (see Figure 2 for an illustration of how a StORF captures upstream DNA).

All genomic repositories, including Ensembl Bacteria, are known to contain varying levels of assembly and annotation error due to a number of constraints. Therefore, UR Extractor has a maximum UR length cutoff of 100kb. A genome with a 100 kb UR either warrants further investigation for biological interest or is most likely to contain annotation error.

Stop - Open Reading Frame – Finder (StORF-Finder) To identify StORFs, StORF-Finder starts by scanning the previously extracted URs (or alternatively any DNA sequence) for loci encapsulated by in-frame, user-defined stop codons (see Figure 2). By the nature of nucleotide triplet abundance, influenced by GC content and other constraints, the three canonical stop codons, TAG, TGA and TAA, are frequently present throughout both out-of-frame genic and intergenic DNA (25). To investigate whether StORF-Finder should itself select which stop codons to use, a comparison was performed of the presence of the triplets with their usage as stop codons across the Ensembl Bacteria genomes. The results of this analysis (see Supplementary Figure 2) show that while there are differences between the three codons, they are not universal and it is likely not beneficial to make generalisations.

Additionally, as with start codons, stop codons are differentially preferred by groups of genes and gene-families both from within the same and across different genomes (26). For example, we can see in Supplementary Figure 2 that the codon TAA is often used more frequently than its distribution across the genomes suggests. Similarly, TGA is often used

cture(0,0)(-30,0)10 (-30,-5)(0,1)10 (-35,0)(1,0)30 (0,-30)10 (-5,-30)(1,0)10 (0,-35)(0,1)30 picturepicture(0,0)(30,0)10 (30,-5)(0,1)10 (35,0)(-1,0)30 (0,

cture(0,0)(-30,0)10 (-30,-5)(0,1)10 (-35,0)(1,0)30 (0,30)10 (-5,30)(1,0)10 (0,35)(0,-1)30 picturepicture(0,0)(30,0)10 (30,-5)(0,1)10 (35,0)(-1,0)30 (0,30

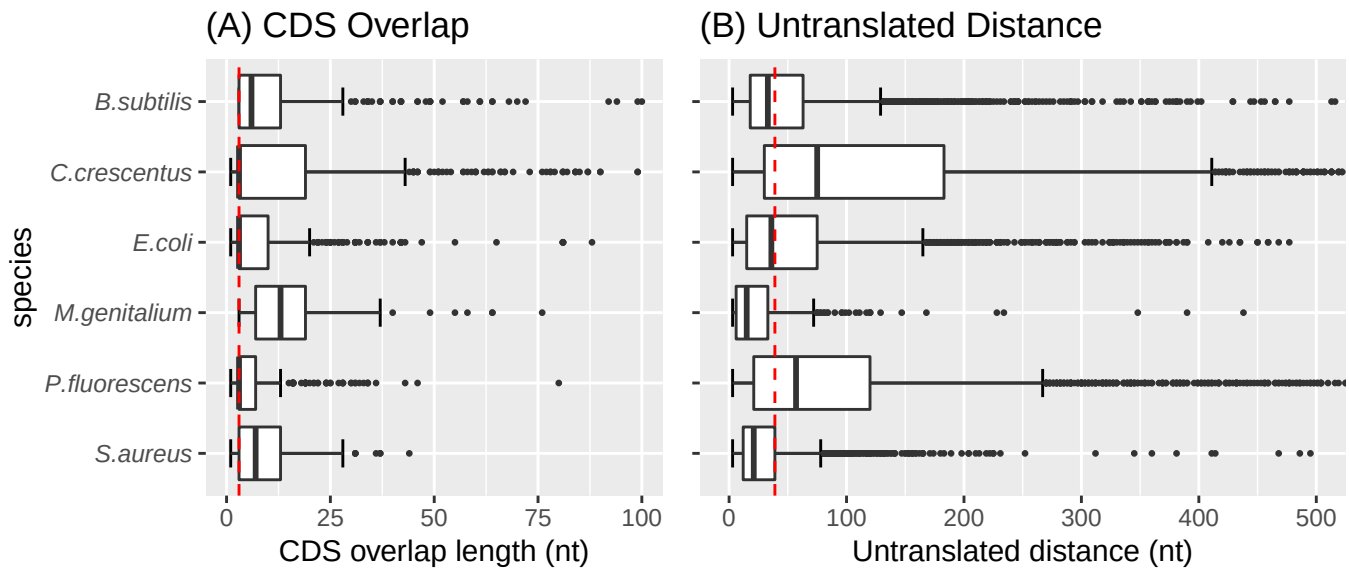

Figure 3. This double plot reports the analysis of the 6 model organism which was used during the parameterisation of StORF-Reporter. Figure A reports the distributions of the Ensembl annotation gene overlap lengths for each model organism with a dotted red line representing the overall median (3 nt) with the x-axis truncated at 100 nt. Figure B reports the distance between an Ensembl gene's start codon and the first in-frame upstream stop codon for the model organisms with the x-axis truncated at 500 nt. The dotted red line represents the overall median (39 nt). These plots indicate that the extension of 50 nt from each end of the extracted intergenic regions is often enough to capture both the true overlap between an annotated gene and the putative gene identified by a StORF and the small amount of upstream non-coding DNA which the StORF will contain.

less frequently than its distribution across the genome would suggest in all 6 model organisms (with TGA never used as a stop codon in *M. genitalium*. For TAG, the usage is mostly in line with the overall distribution across the genomes. While there is evidence of stop codon preference and even re-coding for amino acids in species such as *M. genitalium*, it can be assumed that for the majority of prokaryotic genomes, the three canonical stop codons can be treated equally.

While being able to effectively circumvent start codon identification, StORFs do have their own drawbacks. As we are not able to identify which codon is the start of transcription, this being ATG or otherwise, when translating an identified StORF to amino acid, we use universal genetic code codon table 11 for all sequences. This is a problem as "The biosynthesis of all proteins from all living cells begins with methionine" (27). Therefore, although an alternative start codon may be used, the codon is still translated as Methionine in the final amino acid sequence (28). As a countermeasure against this, StORF-Finder reports both the DNA and amino acid sequences in its output.

As with traditional CDS prediction methods, without any form of filtering, StORF-Finder can report a very high number of overlapping and nested StORFs, even for short regions of DNA. Therefore, a filtering step was developed to reduce the number of StORFs reported. This is done by first ordering all StORFs by their length in descending order and then iteratively removing the nested or shorter StORFs in cases where two StORFs overlap beyond the defined length of 50 nt. This is done individually for each set of StORFs for each UR. This filtration method is applied to all StORFs equally and is intended to prioritise capturing as much of an UR as is possible, while reducing the number of nested and overlapping StORFs. The remaining StORFs are reported in FASTA (DNA and/or amino acid form) and GFF format with both locus coordinates for their original position within the original genome sequence, and their position relative to the UR in which they were reported.

Extracting StORFs from Prodigal and Ensembl annotations of the six Model Organisms

To investigate the ability of StORF-Reporter to expand contemporary annotations, URs and StORFs were extracted from both the annotations and predictions provided by Ensembl Bacteria and the prokaryote gene prediction tool Prodigal (19), for the 6 model organism genomes. The default parameters of UR_Extractor (-gene_ident "ID=gene") were used to record the complete set of genomic elements (coding and non-coding) which was then used to extract URs from the Ensembl provided annotations. Prodigal was applied to the complete DNA sequences of the six model organisms and a novel CDS prediction was performed using the tool's default parameters. As Prodigal only predicted CDSs, UR_Extractor was only able to use the CDS coordinates to extract URs (-gene_ident "CDS").

Escherichia coli Pangenome Analysis

To investigate the impact that StORFs may have on a well-studied prokaryotic model organism pangenome, the Ensembl Bacteria protein sequences and StORFs extracted from *E. coli* strains were studied.

Although gene loci information can be used to build a pangenome, many of the URs which harbor the CDS genes that StORF-Reporter identifies are unlikely to be in the same location across all genomes. The fragmented state of the majority of genomes held in Ensembl Bacteria (or most genome repositories in general) further adds to the necessity

cture(0,0)(-30,0)10 (-30,-5)(0,1)10 (-35,0)(1,0)30 (0,-30)10 (-5,-30)(1,0)10 (0,-35)(0,1)30 picturepicture(0,0)(30,0)10 (30,-5)(0,1)10 (35,0)(-1,0)30 (0,

to use alternative techniques to gather the detail of the pangenome. Amino acid sequence similarity has long been used to group genes of similar function together in gene family analysis, and has been used in this study. CD-Hit (29) is a sequence clustering tool which does not require additional information such as loci or genome data and can create gene families based solely on sequence similarity. The *E. coli* annotated amino acid sequences from Ensembl *(.pep.all.fa)* were combined into one FASTA file and gene-clustering was then undertaken with CD-Hit (version 4.8.1) (29). The gene clustering was performed with the following parameters: aligned sequence identity threshold of 0.9 (90%), length difference cutoff of 0.6 (60%), and the '-g' option was set to the 'more accurate' option (see Supplementary Subsection 3 for more details). The consistent use of the 0.6 length difference cutoff between clustered sequences allowed for the instances where the StORF sequences contained additional upstream non-coding DNA which would have hindered the clustering of the matching coding regions. The strict sequence identity threshold of 0.9 determined that the resulting clusters were very similar across the regions where they did align.

The output of CD-Hit consists of two datafiles, a '.clstr' (cluster) file containing the full sequence cluster metadata and a FASTA file containing the representative ID and sequence for each cluster. The CD-Hit cluster file, which consisted of the identified Ensembl gene clusters, was used as a baseline against which to compare the additional StORF predictions. The amino acid sequences of the *E. coli* StORFs recovered from the URs of the same set of 219 genomes were then combined with the previous CD-Hit output (FASTA) from the Ensembl proteins, which consisted of one representative sequence for each cluster (representatives are often the longer of the sequences in any one cluster). This combined FASTA file then underwent another round of CD-Hit clustering with the same parameters and produced the final *E. coli* pangenome datafiles (see Supplementary Subsection 4.2 for more details on the files created and used in this process).

The results of these clusters are complex and have been classified differently in respect to the sequences they have clustered. As such they will be used throughout the rest of this paper and are referred to as follows:

- **Ensembl-Only**, which refers to the clusters with only Ensembl annotated protein sequences.
- **Ensembl-StORF**, which refers to the clusters which contain at least one sequence from both Ensembl annotated protein sequences and one StORF amino acid sequence.
- **StORF-Only**, which refers to the clusters which solely contain StORF amino acid sequences.

To interpret the output from CD-Hit, a number of Python3 scripts were developed and are available on the StORF-Reporter Github repository https://github.com/NickJD/StORF-Reporter. **single_Genera_CD_Hit_StORF_Reporter_Core_Extension.py** was built to first identify the core-gene family groups of the 219 *E. coli* strains and then identify whether any StORFs clustered with these representative gene cluster groups, but also whether there were any StORF-Only gene clusters spanning large numbers of *E. coli* strains. This allows for the recording of unique strains/genomes that make up each cluster and therefore the ability to calculate each cluster's pangenomic status (core, soft-core, accessory). Next, the .clstr file produced from the combined Ensembl representative sequences and StORF sequences was used to identify whether the Ensembl gene clusters had been added to by StORFs or whether there were any StORF-Only gene clusters. The StORF-Finder output sequence ID tag 'Stop-ORF' was used to distinguish between the Ensembl and StORF sequences.

It is difficult to predict what the result of clustering many thousands of protein sequences together could be. Through introducing hard-coded computation parameters into the natural world, it is inevitable that certain cases are unaccounted for. For example, multiple representatives from separate Ensembl gene clusters can be combined into a new single cluster with the addition of StORF sequences. To account for this, the Ensembl clusters which were combined when StORF sequences were added were first recorded separately in the Ensembl gene cluster results, and then additionally recorded in their new combined clusters with the additional StORFs (see results rows 1 and 2 respectively in Supplementary Table 3).

Extracting StORFs from 6,223 Ensembl genomes

An important aspect of gene family research concerns their diversity and spread across different genera. The StORF-Reporter methodology was applied to the multi-genera collection of 6,223 filtered Ensembl Bacteria genomes using the Ensembl annotations for the UR and StORF extraction. The diversity of assembly and annotation quality of these genomes presented a number of obstacles. As with the*E. coli* genomes, even after filtering, the genomes in this study were often in a fragmented state with a number of low quality regions containing large sections of ambiguous nucleotides. With this in mind, URs and StORFs were extracted individually for each of the 6,223 genomes using default parameters. The resulting StORF-Finder output FASTA files were combined into a single FASTA file with their original genus and species name appended to the start of each sequence header. For example, the full genome name of 'Enterobacter_cloacae_ecwsu1.asm23997v1' was appended to the start of the protein name 'AEW71445' to make 'Enterobacter_cloacae_ecwsu1.asm23997v1|AEW71445' next, the same CD-Hit parameters and workflow were undertaken as for the *E. coli* pangenomic study. A modification of the previously developed single species script (see Subsection) was built to handle multiple genera. This script records multiple different levels of taxa information for each of the sequences such as genera, species and strain. However, as the study on *E. coli* had already covered StORF-Reporter's ability to capture gene sequences at the species level, only the genera specific to each sequence in this study were recorded.

All scripts developed and used are available at https://github.com/NickJD/StORF-Reporter.

Validation of Recovered StORFs

Different validation methods were required to determine what was being detected by the StORF sequences. For the

cture(0,0)(-30,0)10 (-30,-5)(0,1)10 (-35,0)(1,0)30 (0,30)10 (-5,30)(1,0)10 (0,35)(0,-1)30 picturepicture(0,0)(30,0)10 (30,-5)(0,1)10 (35,0)(-1,0)30 (0,30

StORFs identified from within the URs of the Prodigal annotations, an extension to the 'ORForise' platform (8) was developed, which originally was designed to compare annotations from different CDS prediction tools to a reference annotation. This extension, 'StORForise', was used to identify the Ensembl CDS genes missed by Prodigal that StORF-Reporter was able to recover. The default parameters of what classified a gene as 'detected' were taken from the ORForise package (a minimum of 75% coverage of an in-frame CDS prediction for a reference CDS gene). StORF-Reporter was only presented with the extracted URs according to the Prodigal CDS predictions. Mispredictions by Prodigal result in either elongation of an Ensembl CDS gene prediction, or prediction of a CDS where no Ensembl annotation existed. These would constrain the available regions that could be searched for StORFs. Therefore, as part of the ORForise platform, two additional scripts were developed to identify the Ensembl genes which Prodigal missed and then identify the number of missed Ensembl genes which were non-vitiated (not corrupted or contaminated) by the mispredictions of Prodigal (false positives, overlapping Ensembl genes by more than 50 nt).

The protein sequence aligner Diamond (30) (version v0.9.30.131) was used to align the recovered StORFs from the Prodigal and Ensembl annotations of the six model organisms to known proteins from both the same genomes and to the SwissProt protein database (31) (release 2021_01, downloaded 05/03/2021). The sequence similarity alignments were performed with the Diamond blastp option with two separate parameter runs, one with a minimum bitscore of 60 and the second with the same bitscore but also a subject coverage cut-off of >=80%. This was done to report whether some StORFs were as a result of gene duplication or whether they had previously been identified in other studies. This was performed at the amino acid level as although the coding start sites are not identified by StORF-Finder, the coding frame is, and so the upstream region will include the start codon in the same frame (unless mutations are present). This indicates that StORFs can be directly translated into amino acid sequences, and then undergo similarity analysis to find homologous proteins already reported across other genomes and protein databases.

In addition to identifying whether the reported StORFs have been previously reported, the functional profiles of the identified StORFs and how they may differ to the CDS genes from Ensembl, Clusters of Orthologous Groups (COG) functional categories from EggNOG were reported. This functional analysis of the representative sequences from each cluster was performed with EggNOG-Mapper version 2.15 (32) with the latest databases available as downloaded on the 01/08/2021 and with default parameters.

The full ORForise package along with StORForise and all other additional scripts are available at https://github.com/NickJD/ORForise and https://github.com/NickJD/StORF-Reporter.

RESULTS

Unannotated Regions of Model Organisms

UR-Extractor was first applied to both the canonical Ensembl annotations and the Prodigal predictions of the six model organisms. The results of this can be seen in Supplementary Tables 4 and 5. These show that the numbers of URs reported for each model organism were similar between the Ensembl and Prodigal annotations, except for *M. genitalium* which were 157 and 636 respectively. This was likely due to the inability of Prodigal to account for reassignment of the 'UGA/TGA' stop codon and its resultant misreporting of a number of truncated ORFs. The Ensembl annotations also included non-coding genes which were treated the same as coding genes by the UR-Extractor tool. As Prodigal does not predict non-coding genes, it is likely to report additional or longer URs compared to Ensembl. This can explain some of the higher numbers and longer UR lengths in the Prodigal analysis. For example, while the longest UR length in *S. aureus* was 2,591 in the Ensembl annotation, this increased to 12,332 in the Prodigal annotation. However, this was not consistent across all MOs. Interestingly, the longest UR in *P. fluorescens* decreased from 20,088 to 4,264 for Ensembl and Prodigal annotations respectively.

By studying the URs extracted with either the annotations of Prodigal or Ensembl, there were clearly a number of URs with lengths long enough to contain complete CDS genes. While some of the URs of Prodigal are likely to contain a number of known (in Ensembl annotations) non-coding genes and other genomic elements, StORFs are specifically designed to identify coding genes and these are the focus of this paper.

StORF-Reporter Recovers Ensembl Genes Missed By Prodigal

An analysis of the StORFs extracted from the URs of the six MO genomes separately, was undertaken using the annotations from Prodigal. StORF-Reporter was able to recover Ensembl-annotated genes that were missed by Prodigal, and also found many other potential genes that had sequence similarity to the Swiss-Prot database or proteins already annotated in the MO genomes.

The ORForise platform reported that Prodigal was able to identify the vast majority of Ensembl genes from each of the MOs, except for *M. genitalium* (8). The genomic regions containing no CDS predictions were extracted with UR-Extractor and StORFs were reported from these URs using StORF-Finder. For each of the MOs, StORF-Reporter identified a high number of StORFs, as can be seen in Supplementary Table 6. However, StORF-Reporter is impeded by the mispredictions of Prodigal. For each model organism, Prodigal reported a number of CDSs which either had no representative in the Ensembl annotation, were elongated versions of Ensembl genes or were too inaccurate to be classified as 'detected', according to the ORForise platform. This meant that StORF-Reporter was only able to search for StORFs in the non-vitiated regions of the genome. For each MO, between 1-26 non-vitiated missed Ensembl genes were recovered by using StORF-Reporter with the Prodigal annotations.

cture(0,0)(-30,0)10 (-30,-5)(0,1)10 (-35,0)(1,0)30 (0,-30)10 (-5,-30)(1,0)10 (0,-35)(0,1)30 picturepicture(0,0)(30,0)10 (30,-5)(0,1)10 (35,0)(-1,0)30 (0,

Model Organism	# StORFs	Swiss-Prot Hits **[Subject Hit >=80%]**	Intra-Genome Hits **[Subject Hit >=80%]**
B. subtilis	2,322	26 **[20]**	8 **[2]**
C. crescentus	1,554	8 **[5]**	15 **[3]**
E. coli	2,798	148 **[107]**	57 **[13]**
M. genitalium	168	63 **[4]**	50 **[0]**
P. fluorescens	3,404	73 **[46]**	173**[42]**
S. aureus	2,503	23 **[4]**	17 **[3]**

Table 1. Table containing the number of StORFs found in the URs recovered from Ensembl annotations for six model organisms. The numbers of StORFs which had a high sequence similarity and >=80% subject coverage to a protein in Swiss-Prot and Ensembl proteome are listed.

The StORFs from each of the model organisms were DIAMOND blastp searched against databases created from both the Swiss-Prot protein database and the proteomes of the respective model organism (see supplementary Table 7). From these results, it appears that while the ability of StORF-Reporter to recover Ensembl genes which were missed by Prodigal is limited, the number of StORFs which were observed with a high level of sequence similarity to proteins in the curated Swiss-Prot database and from those already identified in the Ensembl annotations requires further investigation.

StORF-Reporter Finds Complete Genes Not Present In Ensembl Annotations

StORF-Reporter found a high number of StORFs representing potential genes not just in the URs reported by Prodigal, but also in the URs from the curated Ensembl annotation, for each of the six MOs. Each of these StORFs have the potential to contain not only undiscovered genes but also historic gene fragments or other functional units waiting to be characterised. A number of StORFs had high sequence similarity to known protein coding genes in the Swiss-Prot protein database. Each of these StORFs have the potential to contain not only undiscovered genes but also historic gene fragments or other functional units waiting to be characterised. There were also a number of StORFs which, while of considerable length, did not have a high level of sequence similarity. This could indicate that these sequences are not yet present in the databases or are fragmented in the genome.

The number of StORFs which attained a match to a protein sequence from the curated Swiss-Prot database are reported in Table 1. There were also a number of StORFs which, while of considerable length (>100 amino acids), did not have a high level of sequence similarity. This could indicate that these sequences are not yet present in the databases or are fragmented in the genome.

To study the possibility that StORFs may capture instances of gene-duplication, for each of the MOs, the reported StORFs were aligned to protein sequence databases built from each of the Ensembl annotations. Between 8-173 StORFs were observed to contain a high quality alignment to putative duplicate 'genic' sequences from within the same genome. These StORFs are likely to represent genes that have undergone duplication and either subsequent loss mutation or have not been detected in the original genome annotation process. While *M. genitalium* observed 50 intra-genome StORFs 'hits', it was the only MO not to have a StORF alignment >=80% of an Ensembl, likely due to the 'UGA/TGA' reassignment. The results are reported in Table 1.

Extending the *Escherichia coli* Pangenome

StORF-Reporter found StORFs that both extend the core-gene set and create novel core-gene clusters which has the potential to extend our knowledge of the *E. coli* pangenome.

The 6,223 Ensembl Bacteria collection was filtered to give a set of 219 *E. coli*. From the 1,042,068 Ensembl protein sequences annotated in these 219 genomes, 34,737 gene-clusters were formed, of which 20,676 were non-singletons. A median number of 3,038 URs and 2,958 StORFs were identified for each the 219 genomes for a total of 673,136 URs and 652,056 StORFs (see Supplementary Table 8 for more detail).

The representatives from each of the Ensembl-Only clusters (all 34,737) were combined with the 652,056 StORFs identified from the same 219 *E. coli* genomes (see Supplementary Table 8 for more detail) and the same CD-Hit sequence clustering was applied. This resulted in a total of 86,579 clusters, of which 31,676 were non-singletons. 51,929 clusters were formed of only StORFs (StORF-Only clusters) and 28,094 clusters were formed of only Ensembl genes (Ensembl-Only clusters). The remainder clustered StORFs together with Ensembl genes (Ensembl-StORF clusters). Some of these clusters containing StORFs and Ensembl genes contained multiple Ensembl cluster representatives, so bringing together multiple gene families (defined henceforth as clusters formed from Ensembl annotation protein sequences). Out of a total of 6,643 Ensembl gene families that clustered with StORF sequences and extended them into additional *E. coli* genomes, 87 were formed where the clustered StORF sequences did not extend the gene family into additional genomes. Supplementary Figure 4 shows that the distribution of the unique *E. coli* genomes present in the Ensembl-Only, Ensembl-StORF and StORF-Only clusters both followed a similar U-shaped curve. Interestingly, the size of the clusters which were extended by the StORFs was often increased by an substantial amount. For example, the proportion of clusters with 10 or fewer unique genomes was reduced from just below 60% to 40% with the addition of the StORF sequences.

The 2,612 core-gene (found in $>=99\%$ genomes) gene families consisting of only Ensembl genes is a similar figure to that of the number of core genes found in other studies of *E. coli* strains (33). Taking the varying levels of genome assembly and annotation quality into consideration, a large number of gene clusters (with and without StORF sequences) were found to contain genes from more than 95% of the *E. coli* genomes (see Supplementary Table 3). Interestingly, some clusters which fell below this 95% level were extended into the 95% and 99% thresholds by the addition of their clustered StORF sequences. There were also StORF-Only clusters with sequences from more than 95% and 99% of the *E. coli* genomes.

The inherently limiting process of using hard-coded sequence similarity and length cutoffs to distinguish gene families has inevitably brought forward a number of interesting cases. One example can be seen in the combining

cture(0,0)(-30,0)10 (-30,-5)(0,1)10 (-35,0)(1,0)30 (0,30)10 (-5,30)(1,0)10 (0,35)(0,-1)30 picturepicture(0,0)(30,0)10 (30,-5)(0,1)10 (35,0)(-1,0)30 (0,30

of multiple Ensembl protein representative sequences, and thus clusters, into one new cluster by the addition of StORFs. The 71 amino acid representative protein AKD71933 of Ensembl cluster 4,714 (consisting of 59 sequences and spanning 24 strains, some of which have duplicates or paralogs within the same genome), was clustered with the 66 amino acid protein AHM40952 of Ensembl cluster 6,112 (consisting of 26 sequences from 26 strains) when the StORFs sequences were included in the CD-Hit clustering. Previously these proteins were distinct enough to be clustered separately. The two Ensembl representatives were clustered together in combined Cluster 4 with 335 StORFs from 99 strains. The StORFs in this cluster are longer than the Ensembl genes (around 100 aa), and so CD-Hit chose a StORF to be its representative sequence for the combined cluster. As can be seen in the Clustal Omega (34) multiple sequence alignment of the representative sequences of the above-mentioned clusters in Supplementary Figure 6, their sequence identity was high, $\geq 90\%$ (the minimum CD-Hit percentage identity for clustering) along the regions in which they aligned.

There were also examples of Ensembl gene clusters combined together with StORF sequences, which could completely change the dynamics of the Ensembl gene families. As seen in Supplementary Figure 7, the three Ensembl representatives sequences, AHM40736, AIT36070 and AFS84250, are combined into a new cluster with a single StORF sequence (Cluster 17,575). As we can observe from their alignment, they are in fact very similar and are further examples of the limitations we impose on ourselves when attempting to categories natural elements into computationally-friendly representations. There were 77 of these types of combined clusters which together had clustered 164 Ensembl representative sequences together. These clusters are presented separately in Supplementary Table 3. Interestingly, a number of these combined clusters contained more unique *E. coli* strains via the StORF sequences than via the Ensembl genes. Of the clusters that combined more than one previous Ensembl cluster, there were no cases of core ($>=99\%$) gene families but 1 example of soft-core ($>=95\%$ to $<99\%$) and 11 accessory gene families ($>=15\%$ to $<95\%$).

As can be seen in Supplementary Table 3, each of the three categories of the overall pangenome of *E. coli* have been extended by the addition of the StORF sequences. Among the most interesting results are the 179 gene-families which were extended to the core of the pangenome by the sequences identified by StORF-Reporter. Additionally, there were cases where clusters which contained both Ensembl and StORF sequences, were classified as core-gene families, on the basis of their StORF sequences alone. While they were grouped with additional Ensembl genes, the Ensembl genes did not contribute to the distribution of the cluster (9 were core, 15 were soft-core and 648 were accessory). Most intriguingly, there were a large number of clusters which were made up entirely of StORF sequences (StORF-Only gene families) and spanned all three pangenome gene categories (239 were core, 216 were soft-core and 3,426 were accessory). Lastly, there were 162 Ensembl clusters which although did cluster with StORF sequences, none of the StORFs were from additional genomes and as such were not counted in the above data.

EggNOG COG functional categories were identified for the representative sequences from the initial set of 34,737 gene clusters in order to investigate whether the functional profiles are different between Ensembl and StORF clusters. The sequences which obtained a COG annotation were separated into three groups: those which did not cluster with a StORF sequence (Ensembl-Only), those which contained both Ensembl and StORF sequences (Ensembl-StORF) and those which only contained StORF sequences (StORF-Only). This resulted in 15,296, 23,419 and 6,481 clusters, respectively. While some singleton Ensembl-Only and StORF-Only clusters did have COG annotations, only those clusters which had sequences from at least 2 different genomes were reported. The 20 COG functional categories were grouped together into their 4 respective domains and are presented in Supplementary Table 9. As the COG group 'Information Storage & Processing' was observed with the largest difference and increase between the Ensembl-Only and StORF-Only clusters, the individual COG categories were extracted and reported in Supplementary Table 10. While the number of Ensembl-Only sequences which obtained a COG classification are much higher than for StORF-Only, the reported distribution of COG categories are similar in both. Both 'A' and 'B' are observed in very low numbers (9 'A', 1 'B' for Ensembl-only and 0 'A', 0 'B' for StORF-Only respectively). StORF-Only sequences have less 'K' but more 'L' than Ensembl-Only sequences, possibly hinting at a functional difference of missing gene function from canonical genome annotations.

StORFs Identified Within and Across Multiple Genera

The previous analyses of URs and StORFs have been conducted on the same species. It could be assumed that even in the case of the *E. coli* pangenome, many of the StORFs which have been identified are remnants from a multitude of different genomic factors such as assembly error, genome structural elements or additional processes we are yet to understand. In this study, all 21,503,164 protein sequences were used from the 6,223 genomes from Ensembl Bacteria (see Supplementary Table 1 for details on genomes used).As with the *E. coli* pangenome analysis, a substantial number of URs and StORFs were identified across the genomes studied. A median number of 2,305 and 1,981 URs and StORFs were identified for each of the 6,223 genomes for a total of 14,221,482 and 13,301,175 respectively (see Supplementary Table 11 for more detail). Additionally, Ensembl-Only, Ensembl-StORF and StORF-Only clusters, were identified spanning multiple genera across the 179 genera from Ensembl Bacteria. As would be expected, the distribution of genera in the Ensembl-Only clusters is wide and many clusters are reported with more than six different genera (see Supplementary Table 12). Furthermore, not only have Ensembl-Only clusters been added to and extended by StORFs (creating Ensembl-StORF clusters), but also substantial numbers of StORF-Only clusters have been found with StORFs from multiple genera.

Many clusters were identified with both Ensembl and StORF sequences from the same genera and genomes. While these are likely to represent important candidates for gene duplication studies, they were not the aim of this study.

cture(0,0)(-30,0)10 (-30,-5)(0,1)10 (-35,0)(1,0)30 (0,-30)10 (-5,-30)(1,0)10 (0,-35)(0,1)30 picturepicture(0,0)(30,0)10 (30,-5)(0,1)10 (35,0)(-1,0)30 (0,

cture(0,0)(-30,0)10 (-30,-5)(0,1)10 (-35,0)(1,0)30 (0,30)10 (-5,30)(1,0)10 (0,35)(0,-1)30 picturepicture(0,0)(30,0)10 (30,-5)(0,1)10 (35,0)(-1,0)30 (0,30

StORFs have changed the diversity and spread of gene families by extending the canonical Ensembl gene clusters into additional genera which were not annotated in the Ensembl annotations. One example of this, from the Ensembl-Only clusters, is a singleton cluster (Cluster 1,845,585) which did not cluster with any other Ensembl gene from any of the 6,223 genomes. It consisted of a 35 amino acid *E. coli* Ensembl gene (reported in GenBank as a hypothetical protein - *E. coli kte42* - gene ELF73632). This singleton sequence, when combined with the StORF sequences from all 6,223 Ensembl genomes, was clustered into the 4th largest combined cluster (Ensembl and StORF combined 'Cluster 3'), with StORFs spanning multiple different genera. This cluster consisted of 1,162 StORF sequences from a total of 149 genomes and 5 genera (*Shigella, Escherichia, Pantoea, Klebsiella, Enterobacter*). The fact that the only organism to have this sequence annotated was *E. coli* may not be a coincidence.

Another example is a large Ensembl gene family (Ensembl Cluster 31), reported in Genbank to be protein 'CQR83579' which is a '30S ribosomal protein S18' (35), which spanned 26 genera and 1,030 sequences. This gene family was extended to a Ensembl-StORF cluster by StORF sequences in the combined cluster 277,221. The Ensembl representative protein 'CQR83579' from an *E. coli* k12 strain was clustered with 4 StORFs from three strains of *Yersinia pestis* and one *E. coli*. These StORFs came from genomes which were not represented in the original Ensembl cluster 31. The four StORF sequences were 81 amino acids long as opposed to the 80 amino acids of the Ensembl representative and had between 96% and 100% sequence similarity. These results were interesting for the specific reason that without the addition of the StORF sequences, analysis on those four genomes could incorrectly conclude that this likely essential ribosomal gene is not present. This may lead to the assumed dynamics of the genomes and their subsequent pangenomes being incorrect or at the very least incomplete.

As a final example, the small Ensembl Cluster 83,472, which consisted of 32 sequences between 63-70 amino acids long and spanning 4 genera (*Escherichia, Streptococcus, Yersinia* and *Shigella*) was extended significantly by 726 StORF sequences from 7 additional genera (*Pantoea, Klebsiella, Citrobacter, Enterobacter, Serratia, Salmonella* and *Bacillus*) as part of the Ensembl-StORF combined Cluster 14. The function of this protein is unclear. A BLAST search against the NCBI non-redundant protein database reported a number of different annotations which indicate that it could be either part of a transposase family, a plasmid stability protein or something else entirely.

As with the *E. coli* pangenome study, the hard-coded cutoffs used for the CD-Hit gene family analysis produced a number of clusters which combined multiple Ensembl representative sequences which were previously clustered separately. Six Ensembl-StORF combined clusters were found to contain more than one Ensembl representative sequence and together combined 14 Ensembl-Only clusters. Additionally, many of the StORF sequences were found in genera which were not present in the original Ensembl-Only clusters. An example of one such cluster, Ensembl-StORF cluster 124,470, combined three Ensembl-Only representatives together with seven StORF sequences. Supplementary Figure 8 shows the phylogenetic tree built with ClustalO and FastTree (36) and plotted with iTOL (37). This tree clearly reports the diversity and at the same time the similarity of the StORFs and Ensembl annotated sequences. Arguably, as with the combined clusters in the *E. coli* pangenome study, these StORF sequences bridge the gap between the three separate Ensembl gene families and as such have changed the dynamics of these inter-genera gene families. The multiple sequence alignment for these sequences is shown in Supplementary Figure 9 and shows how the 3 Ensembl genes have been clustered together with the StORF sequences. While the percentage identities are lower than 90% between the 3 Ensembl genes (the reason they did not cluster together), they are more than 90% when compared to some of the StORF sequences. Additionally, this multiple sequence alignment depicts how for some amino acid positions, StORF sequences can be more similar to one or more Ensembl gene as they are compared to each other. These type of clusters were included in the results of Supplementary Table 12, separately in row 'Ensembl-StORF-Combined'.

As seen in the types of clusters reported in Supplementary Table 12, StORF-Reporter was also able to identify a number of clusters containing only StORF sequences that spanned many different genera. The biggest of these, combined cluster 15,543, spans 22 genera and consists of 81 genomes and 92 sequences. At 380 amino acids long, the representative of this cluster had a 90% percent identity and 55% query coverage to the 'fatty-acid-CoA ligase fadD18' protein in GenBank. This large, widespread protein was mostly observed as a single gene in each genome but in 5 species it was found duplicated with between 2 and 6 copies in non-contiguous URs.

The clusters that were found in at least two or more genomes were investigated for COG functional categories using the EggNOG-Mapper tool. The COG functions assigned to the cluster representatives are presented in Supplementary Table 13. Although the low proportion of StORF-Only clusters with a COG annotation (2.97%) does limit these results, the more than 10% COG category changes for 'METABOLISM' and 'INFORMATION STORAGE & PROCESSING' compared to Ensembl-Only, are still informative of the functional differences between the two groups. As the COG category 'INFORMATION STORAGE & PROCESSING' was reported with the largest proportional increase for and Chi squared statistic, this specific COG group was further investigated in Supplementary Table 14. While the low number of StORF-Only COGs does limit these results, there is still a clear upwards shift to 'Replication, recombination and repair' (L). This is interesting as genes assigned with this COG are often perceived to be essential.

DISCUSSION

The Unannotated Regions of Prokaryote Genomes

StORF-Reporter uses a number of known but often overlooked aspects of genomics, such as the presence of out of frame stop codons (12) and the potential for alternative use of start codons across prokaryotes, to direct the development of a homology-free method of gene discovery to further complement our current genomic knowledge. Additional genomic features such as the high level gene density observed in canonical prokaryotic annotations have been "perceived as evidence of

cture(0,0)(-30,0)10 (-30,-5)(0,1)10 (-35,0)(1,0)30 (0,-30)10 (-5,-30)(1,0)10 (0,-35)(0,1)30 picturepicture(0,0)(30,0)10 (30,-5)(0,1)10 (35,0)(-1,0)30 (0,

cture(0,0)(-30,0)10 (-30,-5)(0,1)10 (-35,0)(1,0)30 (0,30)10 (-5,30)(1,0)10 (0,35)(0,-1)30 picturepicture(0,0)(30,0)10 (30,-5)(0,1)10 (35,0)(-1,0)30 (0,30

adaptive genome streamlining" (1), but are at odds with the observations that many parokoryote genomes contain large numbers of long URs (see Supplementary Tables 4 and 5). As shown in this work, the six model organisms exhibited thousands of URs which were often overlooked and contain genes which are missed by state-of-the-art gene prediction methods and are also missing from the canonical Ensembl annotations. We can only speculate as to whether these genes identified by StORF-Reporter are in fact expressed. Considering experimental evidence of expression in the model organism *E. coli* is missing from large portions of their gene collections (38), it will likely require a large community effort.

This work has clearly identified a number of novel CDS genes not in the current genome annotations. It is difficult to know whether the proportion of the StORFs identified in this study are a result of the current ineffectiveness of genome annotation methods or in fact genes which are no longer expressed due to mutation, possibly of the upstream region (putative pseudogenes). However, evidence that we are finding real genes can be seen in the 'Intra-Genome' results in Table 1. StORF-Reporter identified the presence of duplicates of genes already present in the Ensembl annotations. While pseudogenes are often ignored, or worse, misnamed as 'junk DNA', even in human disease research (39), their presence is important and their absence from annotation is limiting to future studies. This is worth considering in future studies considering the vast numbers of clusters identified with high numbers of additional StORFs sequences which have been identified with high levels of sequence similarity to Ensembl annotation genes across many different genera, to SwissProt proteins and functional COG groups. However, as many of these genes are already annotated elsewhere and are most likely examples of incomplete genome annotation, it could be argued that the most important findings of this work were in fact the StORF-Only clusters which did not have any similarity to any known genes. Interestingly, the stop codon abundances and usages presented in Supplementary Tables 15 and 16, suggest that StORFs without homology to known CDS genes may be at least partially validated with stop codon usage analysis. As coding and non-coding regions of prokaryote genomes are believed to evolve separately, reacting to intra- and inter-genomic pressures independently (40), SNP and mutation rate studies could be used as evidence in determining whether these are still functional genes. Many novel StORFs formed clusters which spanned more than 10 genera and as such are prime examples of the current level of omissions in genome annotation. As such, irrespective of whether they are CDS sequences or other more cryptic genomic elements, they are nonetheless target sequences that future developments in genome annotation and experimental work should investigate.

The 'Stop - Open Reading Frame'

The concept of a StORF shows that despite all the advancements in computational biology, there are still clear problems in how genome annotation is undertaken. Previous findings (8) confirmed that genome annotation methods in use today such as Prodigal (19) accurately detect the majority of CDS genes (in genomes which use the 'universal' codon table). However, in general, while all these tools overpredicted the number of genes, there were still large (>=10-20%) portions of the genomes studied without any annotation.

The specific use of stop codons to identify CDS genes dates back to the early days of automated gene prediction where stop-to-stop regions were identified and then studied for 'signal' (41). While it could be thought that StORFs are exceedingly common in prokaryotic genomes, it has been long known that out-of-frame stop codons (OSCs) appear across prokaryote genomes (both intergenic and genic) more often than GC content and nucleotide abundance would suggest alone (12). These OSCs provide a mechanism of early termination of translation in incorrect reading frames so that the metabolic cost and potential toxic bi-products associated with frameshift events can be reduced. This function of OSCs may have been selected for during the course of genome evolution to act against unintended frameshift occurrences. As such, the higher levels of stop codons observed in out-of-frame genic, intergenic and UR regions, lends further credence to the gene-length uninterrupted StORFs found in these regions.

While some amount of intergenic DNA in any prokaryotic genome is to be expected, the levels observed in the canonical annotations were at odds with the evolutionary theory of genome streamlining (42). Therefore, not only was there a need for a tool to improve upon contemporary annotations, but it was also a fundamental requirement that the URs be the target for investigation. Subsequently, through the development and use of the StORF-Reporter methodology, the established doctrine of what intergenic DNA is, has been undermined in this work.

Identifying the True Intergenic Regions

The aim of this work was not to present yet another tool to annotate genomes, but instead a tool to be used in conjunction with other tools. As such the URs and StORFs identified by StORF-Reporter across the 179 genera from Ensembl Bacteria, can be interpreted and used in a number of different ways. However, once putative StORFs have been identified, the remaining regions in the collection of URs for any one genome are not without significance. For example, if we consider the interaction between bacteriophages and their hosts in isolation, it could be imagined that this continuous flow of genetic material, irrespective of whether it is coding or not, can be indicative of both the host and host-environment history (43). More recent studies have shown that it is not only phage-related and broken genetic material being passed, but also functional genes involved in host-related processes such as antibiotic resistance (44). The genesis of this genetic material and the mutations it accumulates can help decipher both past and future organism interactions. Traditional methods of homology detection via similarity searching do not work well here as many of these genomic elements have either never been characterised or are in a too fragmented state. However, the use of **UR-Extractor.py** does make it easier to study these cryptic regions as the resulting regions after the extraction of StORFs, could be classified as close to 'true Intergenic Regions (IRs)'. Further to this, these true IRs, should be the target for future studies aiming to complete prokaryote genome annotations.

cture(0,0)(-30,0)10 (-30,-5)(0,1)10 (-35,0)(1,0)30 (0,-30)10 (-5,-30)(1,0)10 (0,-35)(0,1)30 picturepicture(0,0)(30,0)10 (30,-5)(0,1)10 (35,0)(-1,0)30 (0,

cture(0,0)(-30,0)10 (-30,-5)(0,1)10 (-35,0)(1,0)30 (0,30)10 (-5,30)(1,0)10 (0,35)(0,-1)30 picturepicture(0,0)(30,0)10 (30,-5)(0,1)10 (35,0)(-1,0)30 (0,30

Supplementing Contemporary Annotations

The annotations provided by Ensembl Bacteria and Prodigal for the six model organisms can be taken to represent the the current state of prokaryotic genome annotation. To compare traditional genome annotation methods such as Prodigal to that of StORF-Reporter would be an unfair comparison. Furthermore, their intended purposes are at odds. Prodigal is a complex, multifaceted tool which uses a number of advanced computing techniques and involves the building of its own model trained on every genome sequence with which it is presented. This is so that it can adjust its parameters according to the features such as GC specific to each genome sequence. Additionally, Prodigal predicts genes for the entire length of the target genome.StORF-Reporter on the other hand is a one-size-fits-all tool which not only specifically studies URs, but also has parameters that are designed to detect the types of genes routinely missed by other tools (short, overlapping and genes which use alternative start codons). Therefore, as supported by our results that investigate StORF-Reporter's ability to identify genes not annotated by Prodigal or Ensembl, it is likely that one of the most beneficial ways to use StORF-Reporter is to generate additional (and likely missed) annotations for a genome (a 'GFF+', as discussed in the introduction). With the GFF_Adder tool, which is part of the ORForise package (8), StORFs can be used to supplement existing genome annotations. The *E. coli* pangenome study presents the opportunity for this process to deliver improvement to contemporary annotations without sacrificing the hard work already undertaken. For example, while the *E. coli* genomes presented in the paper "A comprehensive and high-quality collection of Escherichia coli genomes and their genes" (45) have undergone extensive curation they still contain URs capable of holding genes. Reannotating these genomes would undoubtedly lose much of the work and knowledge used by the authors in the process. Therefore, the ability to investigate only those regions of DNA without annotation and then combine the curated annotations with the additional StORF sequences is a clear advantage over traditional methods of reannotation.

Extending Pangenomes

It has been known for some time that a single genome is not enough to characterise the functional profile of a species as this speed and scope of prokaryotic speciation continues to highlight the need for the study of pangenomics. The *E. coli* pangenome could be argued to be the most studied and characterised of any prokaryotic organism (46). One recent study estimated the pangenomic gene collection at more than 13,000 genes which has "tremendous implications in terms of the diversity and pathogenesis of the species *E. coli* and its ability to colonize and cause disease in the human host." (46). However, the three studies of *E. coli* URs and StORFs in this chapter have shed light on the potential for yet more diversity of not only accessory but also core-genes.

In the 6 model organism study, *E. coli* StORFs reported the highest levels of sequence similarity to either SwissProt or intra-genome proteins (see Table 1). Those results, and the high levels of StORFs found in Ensembl-StORF and StORF-Only clusters in the *E. coli* pangenome has highlighted the possible true number of coding and pseudogenised genes missing from the annotations of one of the most studied organisms. The importance of these potential inter-genome genes is yet to be fully explored. While numerous theories could be deduced from their presence and distribution, fundamentally, they are likely candidates for the study of how a species' pangenome can adapt to ever-changing environments (46). These StORFs, functional or pseudogenised genes or not, are additional genetic blueprints an organism continues to hold onto (or is in the process of losing), against the pressures for genome streamlining. Additionally, some of these StORF-Only clusters were observed with higher levels of the COG group 'Information Storage & Processing', compared to Ensembl-Only (as seen in Supplementary Table 9), traditionally which have thought of as being core functions. Investigating this COG group specifically, there is another shift to the COG function 'Replication, recombination and repair' for StORF-Only clusters (see Supplementary Table 10). Interestingly this COG function covers genes involved in a number of essential processes and as such suggests the importance of a number of the StORF-Only clusters missing from the Ensembl genome annotations.

As with many of the results of this work, StORFs continue to present a complicated picture of their presence across the *E. coli* pangenome. Fundamentally, while the identification of these StORFs have provided additional genomic knowledge previously left in the dark corners of their genomes' URs, to truly use the nascent information they may harbor, much expanded experimental work is needed.

Extending Intra and Inter Genera Gene Collections

The results of the StORFs reported in the six model organisms and across the *E. coli* pangenomes have shown wide levels of diversity and spread. However, it could be possible that many of these StORFs are artifacts of genome assembly and species-specific structural elements and not primarily CDS genes. Therefore, the inter-genera analysis of the 6,223 genomes from Ensembl Bacteria was key to not only studying the presence of StORFs across further diverse species, but to also provide evidence to further validate them as putative CDS genes.

An EggNOG COG functional analysis of the inter-genera results identified that over half of the Ensembl-StORF clusters had a COG function annotation (52.71%) comparable to the Ensembl-only clusters (86.00%, see Supplementary Table 13) suggesting the StORFs in these clusters were real (but missed) genes. Interestingly, even though only 2.60% and 2.97% of the *E. coli* pangenome and inter-genera StORF-Only clusters (respectively) had COG annotations, the most common functions in both, were related to 'Information Storage & Processing'. Further to this, from that COG function, in both the cross-genera and *E. coli* pangenome StORF-Only clusters, it was 'Replication, recombination and repair' (L) which was the most common COG function. StORFs were found across many different genera at high sequence similarities ($>=95\%$) (see Supplementary Table 12). As with the *E. coli* pangenomic study, many StORFs from the same cluster were found in different regions across the genomes they were found in, or in the case of duplication, in different regions of the same genome. This adds to the evidence that StORFs can capture

cture(0,0)(-30,0)10 (-30,-5)(0,1)10 (-35,0)(1,0)30 (0,-30)10 (-5,-30)(1,0)10 (0,-35)(0,1)30 picturepicture(0,0)(30,0)10 (30,-5)(0,1)10 (35,0)(-1,0)30 (0,

cture(0,0)(-30,0)10 (-30,-5)(0,1)10 (-35,0)(1,0)30 (0,30)10 (-5,30)(1,0)10 (0,35)(0,-1)30 picturepicture(0,0)(30,0)10 (30,-5)(0,1)10 (35,0)(-1,0)30 (0,30)10 (-5,30)(1,0)10 (0,35)(0,-1)30 pictu

real CDS genes (even those without functional annotations). The presence of StORFs, sometimes duplicates found within the same genome but in different URs vastly spaced apart, adds credence to the likelihood that they are not structural elements or other genomic artifacts but instead functional (past or present) CDS genes. Fundamentally, this comparison clearly shows that not only are StORFs capturing known functional CDS genes, but also that their function may be unique to species.

As such, the observed variances in lengths of StORFs clustered together within and across the 179 genera in this study provide further justification of the need for species-agnostic gene prediction methodologies which do not rely on established model organisms or databases. Discovering the reason for their omission may help future tool advancements. However, the inadequacies in genome annotation tools and how they are used, as found by the ORForise platform, no doubt have a part to play. While these results may not fundamentally change the understanding of and structure of pangenomes or functional profiles of any one species, they could drastically change the landscape of the functional and phylogenetic interconnections between different genera.

Are Some StORFs Pseudogenised Genes?

The potential diversity of all possible protein sequences is vast (47). A typical protein length of about 300 amino acids can have more than 20^{300} different possible polypeptide chains, yet current protein databases only contain a tiny fraction of this, such as SwissProt, which contains a little over 500,000 protein sequences. Many of these are duplicates or very similar to each other and many of these are not experimentally validated. Adding to this complexity are the vast numbers of pseudogenes, which were likely functional proteins in the recent past, identified through the extensive genome sequencing of the last few decades. While mutation rates have long been studied in prokaryotes (48, 49), there are still unresolved questions surrounding the speed at which pseudogenes are created and subsequently removed from the genome. Some studies have shown that pseudogenes are more likely to be a result of failed horizontal gene transfer (50). Additionally, while archaea and most non-pathogenic bacteria exhibit greater retention of ancient gene remnants, obligate pathogenic bacteria tend to have younger pools of pseudogenes (50). As such, the types of genes pseudogenised and the rate at which they are mutated are highly variable between species and likely are related to environmental specific cues.

However, current methods fail to address this problem, therefore, the true extent of possible pseudogenisation through mutation is likely to remain unclear for the foreseeable future. The StORF-Reporter methodology has the potential to address this issue. As a StORF reports potential ORFs regardless of the likelihood of them being functional and it captures a portion of the 5' untranslated region of a CDS gene, it also has the potential to capture pseudogene-creating mutations, such as start codon or ribosomal binding sites mutations. As reported by Goodhead *et al* (51), of the 6 different process that a gene can be pseudogenised, a StORF has the potential to capture all of them. The most easily identified is when pseudogenisation occurs because of a premature stop. While a single StORF could only report one fragment of a gene pseudogenised by a premature stop codon, the combination of multiple StORFs has the potential to recover the entire sequence of the pseudogenised gene.

CONCLUSION

Genome annotation continues to be a developing field. With each new sequencing project, the proportion of uncharacterised sequences found can vary significantly depending on habitat (52), but often hundreds of novel genomes are discovered with little similarity to previously sequenced taxa (53) suggesting that there is much gene diversity yet to be discovered. Exacerbating this issue, both historic and systematic biases and errors mean that some types of genes will always be missed in these studies, resulting in many tools reporting the same untranslated regions in genomes. This might lead users to the false security of using a consensus of methods, but as we have demonstrated, the StORF-Reporter methodology can enhance current annotations, reporting missed ORFs in these regions.

One major limiting factor shared across many genome annotation analysis studies is the use of a genomic database such as Ensembl Bacteria as a ground truth (under the assumption that they are a complete reference). In this study, we challenged that assumption of completeness through the use of StORF-Reporter to investigate the URs in over six thousand Ensembl genome annotations. In direct contradiction to the idea of genome annotation completeness, a high number of UR StORFs were similar to known proteins. Furthermore, StORFs without similarity to Ensembl proteins, were often found across multiple genera with similar predicted functional profiles to "known" genes. While both these groups of StORFs are therefore prime candidates to be added into their respective genome annotations, the majority of StORFs across all genomes were without similarity to known proteins across any database examined. The processes which influence the emergence, distribution and subsequent loss of both novel and modified genes such as gene duplication, neofunctionalisation, 'genome streamlining' and purifying selection (54, 55, 56), suggests substantial pressure to 'lose' non-functional gene sequences. These factors make the resulting sequences which do persist (especially those found across multiple genera), regardless of similarity to known genes, more likely to be functionally important and/or active. Accessory genes, those most likely to be captured by StORFs (See Supplementary Table 3), are now thought to often carry out important and essential functions (1), often for adapting to environmental niches and changes (57). Investigating this relationship should be a priority for future work.

Conflict of interest statement. None declared.

REFERENCES

1. Itamar Sela, Yuri I Wolf, and Eugene V Koonin. Theory of prokaryotic genome evolution. *Proceedings of the National Academy of Sciences*, 113(41):11399–11407, 2016.
2. Ryan J Taft, Michael Pheasant, and John S Mattick. The relationship between non-protein-coding DNA and eukaryotic complexity. *Bioessays*, 29(3):288–299, 2007.

cture(0,0)(-30,0)10 (-30,-5)(0,1)10 (-35,0)(1,0)30 (0,-30)10 (-5,-30)(1,0)10 (0,-35)(0,1)30 picturepicture(0,0)(30,0)10 (30,-5)(0,1)10 (35,0)(-1,0)30 (0,-30)10 (-5,-30)(1,0)10 (0,-35)(0,1)30 p

cture(0,0)(-30,0)10 (-30,-5)(0,1)10 (-35,0)(1,0)30 (0,30)10 (-5,30)(1,0)10 (0,35)(0,-1)30 picturepicture(0,0)(30,0)10 (30,-5)(0,1)10 (35,0)(-1,0)30 (0,30)10 (-5,30)(1,0)10 (0,35)(0,-1)30 pictu

3. Matthew R Hemm, Brian J Paul, Thomas D Schneider, Gisela Storz, and Kenneth E Rudd. Small membrane proteins found by comparative genomics and ribosome binding site models. *Molecular Microbiology*, 70(6):1487–1501, 2008.
4. Jayavel Sridhar, Radhakrishnan Sabarinathan, Shanmugam Siva Balan, Ziauddin Ahamed Rafi, Paramasamy Gunasekaran, and Kanagaraj Sekar. Junker: an intergenic explorer for bacterial genomes. *Genomics, Proteomics & Bioinformatics*, 9(4-5):179–182, 2011.
5. Chen-Hsun Tsai, Rick Liao, Brendan Chou, Michael Palumbo, and Lydia M Contreras. Genome-wide analyses in bacteria show small-RNA enrichment for long and conserved intergenic regions. *Journal of bacteriology*, 197(1):40–50, 2015.
6. Daniel Luis Notari, Aurione Molin, Vanessa Davanzo, Douglas Picolotto, Helena Graziottin Ribeiro, and Scheila de Avila e Silva. IntergenicDB: a database for intergenic sequences. *Bioinformation*, 10(6):381, 2014.
7. Harry A Thorpe, Sion C Bayliss, Laurence D Hurst, and Edward J Feil. Comparative analyses of selection operating on nontranslated intergenic regions of diverse bacterial species. *Genetics*, 206(1):363–376, 2017.
8. Nicholas J Dimonaco, Wayne Aubrey, Kim Kenobi, Amanda Clare, and Christopher J Creevey. No one tool to rule them all: Prokaryotic gene prediction tool annotations are highly dependent on the organism of study. *Bioinformatics*, 12 2021.
9. L Dalgarno and J Shine. Conserved terminal sequence in 18s rrna may represent terminator anticodons. *Nature New Biology*, 245(148):261–262, 1973.
10. Douglas F Browning and Stephen JW Busby. The regulation of bacterial transcription initiation. *Nature Reviews Microbiology*, 2(1):57–65, 2004.
11. Thomas Dandekar, Berend Snel, Martijn Huynen, and Peer Bork. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends in Biochemical Sciences*, 23(9):324–328, 1998.
12. Herman Tse, James J Cai, Hoi-Wah Tsoi, Esther PT Lam, and Kwok-Yung Yuen. Natural selection retains overrepresented out-of-frame stop codons against frameshift peptides in prokaryotes. *BMC Genomics*, 11(1):1–13, 2010.
13. Andre Villegas and Andrew M Kropinski. An analysis of initiation codon utilization in the Domain Bacteria–concerns about the quality of bacterial genome annotation. *Microbiology*, 154(9):2559–2661, 2008.
14. Pavel V Baranov, John F Atkins, and Martina M Yordanova. Augmented genetic decoding: global, local and temporal alterations of decoding processes and codon meaning. *Nature Reviews Genetics*, 16(9):517–529, 2015.
15. Frida Belinky, Vladimir N Babenko, Igor B Rogozin, and Eugene V Koonin. Purifying and positive selection in the evolution of stop codons. *Scientific Reports*, 8(1):1–11, 2018.
16. Inna S Povolotskaya, Fyodor A Kondrashov, Alice Ledda, and Peter K Vlasov. Stop codons in bacteria are not selectively equivalent. *Biology Direct*, 7(1):1–13, 2012.
17. Patricia Sieber, Matthias Platzer, and Stefan Schuster. The definition of open reading frame revisited. *Trends in Genetics*, 34(3):167–170, 2018.
18. Karen Eilbeck, Suzanna E Lewis, Christopher J Mungall, Mark Yandell, Lincoln Stein, Richard Durbin, and Michael Ashburner. The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biology*, 6(5):R44, 2005.
19. Doug Hyatt, Gwo-Liang Chen, Philip F LoCascio, Miriam L Land, Frank W Larimer, and Loren J Hauser. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11(1):119, 2010.
20. Kevin L Howe, Bruno Contreras-Moreira, Nishadi De Silva, Gareth Maslen, Wasiu Akanni, James Allen, Jorge Alvarez-Jarreta, Matthieu Barba, Dan M Bolser, Lahcen Cambell, et al. Ensembl Genomes 2020 – enabling non-vertebrate genomic research. *Nucleic Acids Research*, 48(D1):D689–D695, 2020.
21. Alex Bateman, Maria-Jesus Martin, Sandra Orchard, Michele Magrane, Rahat Agivetova, Shadab Ahmad, Emanuele Alpi, Emily H Bowler-Barnett, Ramona Britto, Borisas Bursteinas, et al. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research*, 49(D1):D480–D489, 2020.
22. Andrew S. Warren, Jeremy Archuleta, Wu-chun Feng, and Joao Carlos Setubal. Missing genes in the annotation of prokaryotic genomes. *BMC Bioinformatics*, 11(1):131, 2010.
23. Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009.
24. Niv Sabath, Dan Graur, and Giddy Landan. Same-strand overlapping genes in bacteria: compositional determinants of phase bias. *Biology Direct*, 3(1):36, 2008.
25. Tit-Yee Wong, Sanjit Fernandes, Naby Sankhon, Patrick P Leong, Jimmy Kuo, and Jong-Kang Liu. Role of premature stop codons in bacterial evolution. *Journal of bacteriology*, 190(20):6718–6725, 2008.
26. Natalia N Ivanova, Patrick Schwientek, H James Tripp, Christian Rinke, Amrita Pati, Marcel Huntemann, Axel Visel, Tanja Woyke, Nikos C Kyrpides, and Edward M Rubin. Stop codon reassignments in the wild. *Science*, 344(6186):909–913, 2014.
27. Fred Sherman, John W Stewart, and Susumu Tsunasawa. Methionine or not methionine at the beginning of a protein. *Bioessays*, 3(1):27–31, 1985.
28. Alexey V Lobanov, Anton A Turanov, Dolph L Hatfield, and Vadim N Gladyshev. Dual functions of codons in the genetic code. *Critical Reviews in Biochemistry and Molecular Biology*, 45(4):257–265, 2010.
29. Limin Fu, Beifang Niu, Zhengwei Zhu, Sitao Wu, and Weizhong Li. Cd-hit: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23):3150–3152, 2012.
30. Benjamin Buchfink, Chao Xie, and Daniel H Huson. Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, 12(1):59–60, 2015.
31. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research*, 47(D1):D506–D515, 2019.
32. Carlos P Cantalapiedra, Ana Hernández-Plaza, Ivica Letunic, Peer Bork, and Jaime Huerta-Cepas. eggNOG-mapper v2: Functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol Biol Evol*, 38(12):5825–5829, 2021.
33. Rohan Maddamsetti, Philip J Hatcher, Anna G Green, Barry L Williams, Debora S Marks, and Richard E Lenski. Core genes evolve rapidly in the long-term evolution experiment with escherichia coli. *Genome biology and evolution*, 9(4):1072–1083, 2017.
34. Fabian Sievers and Desmond G Higgins. Clustal omega for making accurate alignments of many protein sequences. *Protein Science*, 27(1):135–145, 2018.
35. Jayati Sengupta, Rajendra K Agrawal, and Joachim Frank. Visualization of protein s1 within the 30s ribosomal subunit and its interaction with messenger rna. *Proceedings of the National Academy of Sciences*, 98(21):11991–11996, 2001.
36. Morgan N Price, Paramvir S Dehal, and Adam P Arkin. Fasttree 2–approximately maximum-likelihood trees for large alignments. *PloS one*, 5(3), 2010.
37. Ivica Letunic and Peer Bork. Interactive tree of life (itol) v5: an online tool for phylogenetic tree display and annotation. *Nucleic acids research*, 49(W1):W293–W296, 2021.
38. Sankha Ghatak, Zachary A King, Anand Sastry, and Bernhard O Palsson. The y-ome defines the 35% of escherichia coli genes that lack experimental evidence of function. *Nucleic acids research*, 47(5):2446–2454, 2019.
39. Ryan Charles Pink, Kate Wicks, Daniel Paul Caley, Emma Kathleen Punch, Laura Jacobs, and David Raul Francisco Carter. Pseudogenes: pseudo-functional or key regulators in health and disease? *Rna*, 17(5):792–798, 2011.
40. Igor B Rogozin, Kira S Makarova, Darren A Natale, Alexey N Spiridonov, Roman L Tatusov, Yuri I Wolf, Jodie Yin, and Eugene V Koonin. Congruent evolution of different classes of non-coding DNA in prokaryotic genomes. *Nucleic Acids Research*, 30(19):4264–4271, 2002.
41. Mark Borodovsky and James McIninch. GENMARK: Parallel gene recognition for both DNA strands. *Computers & Chemistry*, 17(2):123–133, 1993.
42. Michael Lynch. Streamlining and simplification of microbial genome architecture. *Annu. Rev. Microbiol.*, 60:327–349, 2006.
43. Roger W Hendrix. Bacteriophage genomics. *Current Opinion in Microbiology*, 6(5):506–511, 2003.
44. Anna Colavecchio, Brigitte Cadieux, Amanda Lo, and Lawrence D Goodridge. Bacteriophages contribute to the spread of antibiotic resistance genes among foodborne pathogens of the enterobacteriaceae family–a review. *Frontiers in Microbiology*, 8:1108, 2017.
45. Gal Horesh, Grace A Blackwell, Gerry Tonkin-Hill, Jukka Corander, Eva Heinz, and Nicholas R Thomson. A comprehensive and high-quality collection of Escherichia coli genomes and their genes. *Microbial Genomics*, 7(2), 2021.
46. David A Rasko, MJ Rosovitz, Garry SA Myers, Emmanuel F Mongodin, W Florian Fricke, Pawel Gajer, Jonathan Crabtree, Mohammed Sebaihia, Nicholas R Thomson, Roy Chaudhuri, et al. The pangenome structure of Escherichia coli: comparative genomic analysis of E. coli commensal and pathogenic isolates. *Journal of Bacteriology*, 190(20):6881–6893, 2008.

cture(0,0)(-30,0)10 (-30,-5)(0,1)10 (-35,0)(1,0)30 (0,-30)10 (-5,-30)(1,0)10 (0,-35)(0,1)30 picturepicture(0,0)(30,0)10 (30,-5)(0,1)10 (35,0)(-1,0)30 (0,-30)10 (-5,-30)(1,0)10 (0,-35)(0,1)30 p

cture(0,0)(-30,0)10 (-30,-5)(0,1)10 (-35,0)(1,0)30 (0,30)10 (-5,30)(1,0)10 (0,35)(0,-1)30 picturepicture(0,0)(30,0)10 (30,-5)(0,1)10 (35,0)(-1,0)30 (0,30

47. Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. The shape and structure of proteins. In *Molecular Biology of the Cell. 4th edition*. Garland Science, 2002.
48. Salvador E Luria and Max Delbrück. Mutations of bacteria from virus sensitivity to virus resistance. *Genetics*, 28(6):491, 1943.
49. William A Rosche and Patricia L Foster. Determining mutation rates in bacterial populations. *Methods*, 20(1):4–17, 2000.
50. Yang Liu, Paul M Harrison, Victor Kunin, and Mark Gerstein. Comprehensive analysis of pseudogenes in prokaryotes: widespread gene decay and failure of putative horizontally transferred genes. *Genome Biology*, 5(9):1–11, 2004.
51. Ian Goodhead and Alistair C Darby. Taking the pseudo out of pseudogenes. *Current Opinion in Microbiology*, 23:102–109, 2015.
52. Zheng Zhang, Jianing Wang, Jinlan Wang, Jingjing Wang, and Yuezhong Li. Estimate of the sequenced proportion of the global prokaryotic genome. *Microbiome*, 8(1):1–9, 2020.
53. Mark Kowarsky, Joan Camunas-Soler, Michael Kertesz, Iwijn De Vlaminck, Winston Koh, Wenying Pan, Lance Martin, Norma F Neff, Jennifer Okamoto, Ronald J Wong, et al. Numerous uncharacterized and highly divergent microbes which colonize humans are revealed by circulating cell-free DNA. *Proceedings of the National Academy of Sciences*, 114(36):9623–9628, 2017.
54. Susumu Ohno. *Evolution by gene duplication*. Springer Science & Business Media, 2013.
55. Anthony Levasseur and Pierre Pontarotti. The role of duplications in the evolution of genomes highlights the need for evolutionary-based approaches in comparative genomics. *Biology Direct*, 6(1):1–12, 2011.
56. Stephen J Giovannoni, J Cameron Thrash, and Ben Temperton. Implications of streamlining theory for microbial ecology. *The ISME Journal*, 8(8):1553–1565, 2014.
57. Jian Jiao, Meng Ni, Biliang Zhang, Ziding Zhang, J Peter W Young, Ting-Fung Chan, Wen Xin Chen, Hon-Ming Lam, and Chang Fu Tian. Coordinated regulation of core and accessory genes in the multipartite genome of Sinorhizobium fredii. *PLoS Genetics*, 14(5):e1007428, 2018.

cture(0,0)(-30,0)10 (-30,-5)(0,1)10 (-35,0)(1,0)30 (0,-30)10 (-5,-30)(1,0)10 (0,-35)(0,1)30 picturepicture(0,0)(30,0)10 (30,-5)(0,1)10 (35,0)(-1,0)30 (0,