Hundreds of out-of-frame remodelled gene families in the E. coli pangenome

Andrew K Watson, Philippe Lopez, Eric Bapteste

Hundreds of out-of-frame remodelled gene families in the *E. coli* pangenome

Andrew K Watson, Philippe Lopez, Eric Bapteste*

Institut de Systématique, Evolution, Biodiversité (ISYEB), Sorbonne Université, CNRS, Museum National d'Histoire Naturelle, EPHE, Université des Antilles, 7, quai Saint Bernard, 75005 Paris, France

*Corresponding author: Eric Bapteste eric.bapteste@mnhn.fr

Abstract

All genomes include gene families with very limited taxonomic distributions that potentially represent new genes and innovations in protein-coding sequence, raising questions on the origins of such genes. Some of these genes are hypothesised to have formed *de novo*, from non-coding sequences, and recent work has begun to elucidate the processes by which *de novo* gene formation can occur. A special case of *de novo* gene formation, overprinting, describes the origin of new genes from non-coding alternative reading frames of existing open reading frames (ORFs). We argue that additionally, out-of-frame gene fission/fusion events of alternative reading frames of ORFs and out-of-frame lateral gene transfers could contribute to the origin of new gene families. To demonstrate this, we developed an original pattern-search in sequence similarity networks, enhancing the use of these graphs, commonly used to detect in-frame remodelled genes. We applied this approach to gene families in 524 complete genomes of *Escherichia coli*. We identified 767 gene families whose evolutionary history likely included at least one out-of-frame remodelling event. These genes with out-of-frame components represent ~2.5% of all genes in the *E. coli* pangenome, suggesting that alternative reading frames of existing ORFs can contribute to a significant proportion of *de novo* genes in bacteria.

Introduction

The origins of new genes and ORFans

One of the most striking findings arising from a range of recent comparative genomics and pangenome studies has been the extent to which the acquisition or creation of "new" genes contributes to the evolution of different lineages (Kaessmann 2010). These include remarkable sets of genes called genomic ORFans that appear to be unique to individual genomes or lineages (Fischer and Eisenberg 1999). Thus, explaining the origin of taxonomically restricted gene families is an important aspect of understanding the evolution of genes and genomes. These gene families may have arisen by previously described divergent evolutionary processes, such as the accumulation of sufficient non-synonymous substitutions that makes it impossible to detect sequence similarity between novel genes and their homologs. Moreover, gene families can have complex evolutionary histories, in addition to substitutions, duplications, lateral gene transfers, and fusion and fission events,

which altogether complicate the understanding of gene origins (Tautz and Domazet-Lošo 2011; McLysaght and Hurst 2016). Furthermore, some genes with limited taxonomic distribution appear to have originated *de novo* in some lineages.

The idea that genes can be created *de novo* from non-coding sequences had historically been resisted, due to the low probability that a random string of amino acids could acquire function (Jacob 1977). At the same time, the first genes hypothesised to have been created by overprinting (discussed below) were identified in viral genomes (Barrell et al. 1976; Brown and Smith 1977). Since that time, the frequency of *de novo* gene formation as opposed to traditional routes for gene creation by modification (e.g. following gene duplication events and fusion or fission events), as estimated by systematic screens attempting to identify *de novo* genes, has remained controversial (Guerzoni and McLysaght 2016; Moyers and Zhang 2016; Domazet-Lošo et al. 2017; Moyers and Zhang 2017; Casola 2018), and new approaches to improve the ability to distinguish truly "novel" genes are under constant development (Casola 2018; Jain et al. 2019; Vakirlis and McLysaght 2019). Nonetheless, carefully characterised examples of genes hypothesised to have formed *de novo* have been documented (Johnson et al. 2001; Levine et al. 2006; Cai et al. 2008; Zhou et al. 2008; Knowles and McLysaght 2009) and suggested to be subjected to purifying selection, implying that these novel genes are functional (Xu and Zhang 2016). Further, a number of hypotheses have been developed to describe how *de novo* genes might form (Levine et al. 2006; Kaessmann 2010; Zhao et al. 2014; McLysaght and Hurst 2016).

Overprinting in the origin of new genes

Overprinting has been described as a special case of *de novo* gene creation. A distinction is drawn between *de novo* gene formation and overprinting in that de novo gene formation describes the origin of genes from previously untranscribed regions, while overprinting describes the origin of genes from alternative, previously untranslated, reading frames of existing ORFs. Overprinting can occur by the acquisition of a new out-of-frame start codon, leading to the creation of a completely new proto-gene (Ohno 1984; Grassé 1977), or by the extension of existing ORFs through the acquisition of upstream start codons or the loss of downstream stop codons, such that the new extended region of the ORF overlaps with an alternative, untranslated reading frame of a neighbouring gene (Rancurel et al. 2009). In the short term, both scenarios would create a pair of overlapping ORFs. Either the overlap of an original parent ORF and a new proto-gene, likely to be short owing to the random nature of its protein-coding sequence, or the overlap between a new ORF created by the extension of an existing ORF and a neighbouring ORF. If either of these new genes

3

persist and continue to be expressed in their host lineages, then, with time and the accumulation of changes, they could become new genes. Crucially, both the parent ORF and the newly formed genes originate from the same DNA sequence. However, the use of alternative, previously untranslated reading frames means that all or part of the protein sequences they encode would share no similarity with their parent ORFs. Duplication events occurring either before or after such overprinting events, coupled with differential loss and/or retention of the parent ORF and of the "novel" ORFs, could then lead to the physical separation of the new gene from its parent. For example, in Drosophila 56 genes were identified where two different alleles with variation in the position of a stop codon were maintained in a genome, demonstrating the principle of differential extension of genes following duplication events (Lee and Reinhardt 2012)

Traditionally, cases of overprinting are identified based on searches for pairs of overlapping protein-coding genes in genomes. Overlapping protein-coding gene pairs are relatively common in viral genomes, where genes hypothesised to have been formed by overprinting with a broad range of accessory functions have been characterised, including pairs of genes overlapping both on the same strand (sense) or opposite strands (antisense) (Sabath et al. 2012; Barrell et al. 1976; Brown and Smith 1977; Pavesi 2006; Rancurel et al. 2009; Carter et al. 2013). While a very short overlap of <15 nucleotides between protein-coding gene pairs, and particularly same-strand overlap, is considered common in prokaryotes (Fukuda et al. 1999; Fukuda et al. 2003; Johnson and Chisholm 2004), longer pairs of overlapping genes of the type associated with overprinting have been reported more rarely (McVeigh et al. 2000; Behrens et al. 2002; Delaye et al. 2008; Fellner et al. 2014; Fellner et al. 2015; Hücker, Vanderhaeghen, Abellan-Schneyder, Wecko, et al. 2018; Hücker, Vanderhaeghen, Abellan-Schneyder, Scherer, et al. 2018; Vanderhaeghen et al. 2018). However, commonly used genome annotation tools often repress the annotation of an embedded gene that overlaps over its entire length with a longer gene, meaning that the true number of overlapping genes could be systematically underrepresented in annotated genomes (Neuhaus et al. 2010). Thus, the contribution of overprinting to genome evolution in prokaryotes may be significantly underestimated. The results of recent systematic ribosomal profiling studies (Ingolia et al. 2009) have supported this idea, with translatome characterisation providing additional evidence for the translation of antisense RNAs previously assumed to be "non-coding" in bacteria (Friedman et al. 2017; Smith et al. 2019; Weaver et al. 2019; Ardern et al. 2020), including 27 in E. coli K-12 and hundreds in *Mycobacterium tuberculosis*.

4

Here, we expanded the use of an existing network-based method, originally developed to identify the signatures of in-frame gene remodelling (Pathmanathan et al. 2018), to also detect protein-coding genes that may have formed from parts of out-of-frame existing coding sequences. Importantly, this method does not rely on direct observation of overlapping gene pairs. This approach systematically searches for protein sequence similarity between annotated ORFs and the predicted protein sequences from the alternative reading frames of other annotated ORFs using sequence similarity networks (Corel et al. 2016; Méheust et al. 2016; Méheust, Bhattacharya, et al. 2018; Watson et al. 2019). It relies on the automated detection of a novel pattern within sequence similarity networks, compatible with the origin of new genes in genomes through at least one out-of-frame gene remodelling event.

Proposing out-of-frame remodelling events as an additional, novel route to gene creation

Fusion and fission events can create new genes combining functional units in different ways, "tinkering" with existing genetic material with the potential to produce an emergent function (Jacob 1977; Pathmanathan et al. 2018). Typically, gene fusions are proposed to play a major role in the origin of new genes, with early estimates suggesting a contribution of between 27% and 64% to multidomain protein evolution in bacteria (Pasek et al. 2006). A strong preference for fusion over fission events has been predicted (Snel et al. 2000; Kummerfeld and Teichmann 2005; Marsh and Teichmann 2010), though some reports suggest that rates of gene fission are underestimated (Leonard and Richards 2012). Fusion and fission events have historically been detected by parsing the results of protein sequence similarity searches (Enright et al. 1999; Marcotte et al. 1999), based on identifying protein sequences (composite sequences) that have partial and non-overlapping similarity to at least two other protein sequences from distinct gene families (component sequences) (Jachiet et al. 2013; Pathmanathan et al. 2018) (Figure 1).

At the DNA level, an archetypical fusion between parent genes A and B combine components (e.g. gene domains or entire gene sequences) from each gene perfectly in-frame with one another. Such in-frame fusions prevent the introduction of frameshifts that would alter the reading frame of either of its components, and thus result in fully translated protein sequences encoded by the in-frame composite genes. However, fusion could also occur out-of-frame, a case where the original reading frame of at least one component (e.g. a gene domain, or an almost entire gene sequence) is altered. The component in an

5

alternative reading frame is unlikely to encode an existing protein-coding gene, but rather a random protein sequence including multiple stop codons. Thus, out-of-frame fusions (and, likewise fissions) are expected to be deleterious and counter-selected without additional insertion/deletion events to revert that particular out-of-frame gene fragment to its original reading frame. Yet, in some cases, out-of-frame fusions could be tolerated, and the subsequent accumulation of changes could lead to the successful formation of new genes and peptides. This idea is analogous to that presented by (Stewart and Rogers 2019), who identified 51 cases in which chromosomal rearrangements in Drosophila combined the 5' regions of existing ORFs with previously untranscribed regions of the genome, leading to the creation of new genes that included novel peptides. Given that prokaryotic genomes are chimeric, in that they include genes with varied evolutionary histories due to gene transfers (Gogarten et al. 2002; Bapteste et al. 2009; Dagan and Martin 2009), and that genes acquired by gene transfer participate in traditional fusion and fission events (Wolf et al. 2000; Méheust et al. 2016; Méheust, Bhattacharya, et al. 2018; Méheust, Watson, et al. 2018), we further propose that out-of-frame fusion events could eventually combine genetic components with different phylogenetic origins. Moreover, in theory, integration of full-sized laterally transferred genes could occur out-of-frame but still lead to the production of a new gene. Importantly, all of these theoretical out-of-frame processes (overprinting, out-of-frame fusion/fission within genomes, and out-of-frame partial or full-sized LGT inside a host genome) produce a detectable pattern in sequence similarity networks. They connect nodes corresponding to an out-of-frame "component" gene with an in-frame ORF. Gene remodelling is a term used to describe how new genes can be created by modifications and rearrangements to existing genes, including by fusion and fission or by overprinting, meaning that two evolutionarily related genes, one partly descending from the other, are not homologs. As such, the gene families identified in this study are henceforth referred to as "Out-of-frame Remodelled" gene families (ORgene families).

Here, we used original analyses of sequence similarity networks and provided evidence for a significant role of out-of-frame processes in gene remodelling in *E. coli*. We generated and analysed a novel, inclusive type of sequence similarity network that integrated, not only ORFs encoded proteins, but also protein sequences predicted from all reading frames of all annotated ORFs from 524 complete *Escherichia coli* genome assemblies. *E. coli* was chosen for this study as an established model system for which there is extensive associated metadata, and a bacterial species in which some strongly supported examples of overlapping protein-coding reading frames and overprinting have already been identified (McVeigh et al. 2000; Behrens et al. 2002; Delaye et al. 2008; Fellner
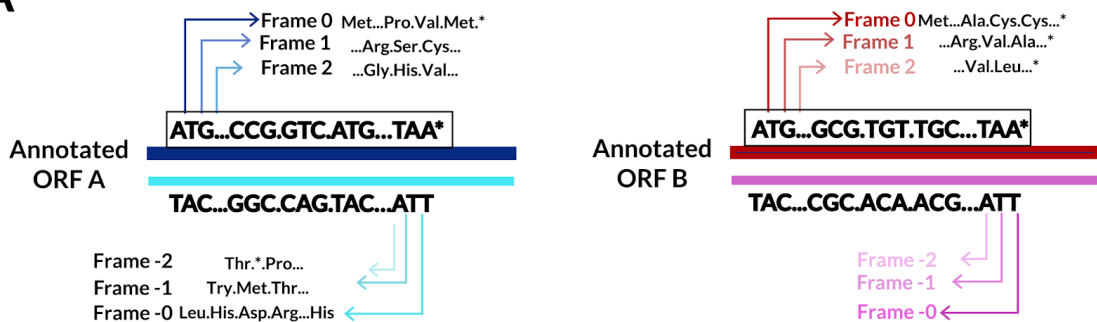
6

et al. 2014; Fellner et al. 2015; Hücker, Vanderhaeghen, Abellan-Schneyder, Wecko, et al. 2018; Hücker, Vanderhaeghen, Abellan-Schneyder, Scherer, et al. 2018; Vanderhaeghen et al. 2018). Remarkably, our approach identified 767 gene families, conserved in at least two host genomes, matching the network pattern associated with out-of-frame remodelling processes, suggesting that these processes may be implicated in a significant proportion of new gene creation.
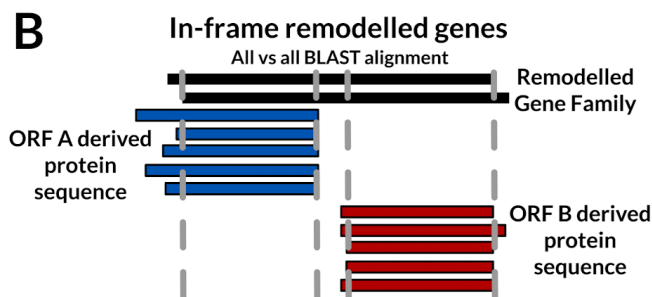
Results & Discussion

Definition and detection of ORgenes

Using a combination of CD-HIT and six-frame all-vs-all protein sequence similarity searches (see materials and methods), we clustered 2,568,313 protein sequences derived from the in-frame translation of annotated open reading frames from 524 *E. coli* genomes into 31,277 gene families. We then investigated the processes through which these families originated. 1,983 gene families (6.3% of the pangenome) were predicted to have originated as remodelled gene families (as defined in Figure 1 and materials and methods). Within these remodelled gene families, 1,216 (61.3%) had similarity to two or more in-frame components from ORF derived protein sequences (Figure 1B), consistent with the proposed role for in-frame gene remodelling events in molecular evolution (Enright et al. 1999; Pasek et al. 2006; Jachiet et al. 2014). Remarkably, however, the remaining 767 remodelled gene families included at least one out-of-frame component, i.e. a genetic segment that only matches with an alternative reading frame derived from an existing ORF sequence (Figure 1C). Henceforth, we refer to such remodelled genes as ORgenes (for Out-of-frame Remodelled genes). An ORgene is a gene that results from the addition of at least one out of frame segment or that produces a novel remodelled gene by out-of-frame fission or overprinting. We hypothesise that these genes may occur by a combination of overprinting (SI Figure 1), gene fusion/fission events (SI Figure 2), simpler frameshifts or out-of-frame LGT (SI Figure 3). Overall, these 767 ORgene families make up ~2.5% of all 31,277 identified *E. coli* gene families, suggesting that out-of-frame events could be a significant contributor to the evolution of protein-coding genes in *E. coli*.
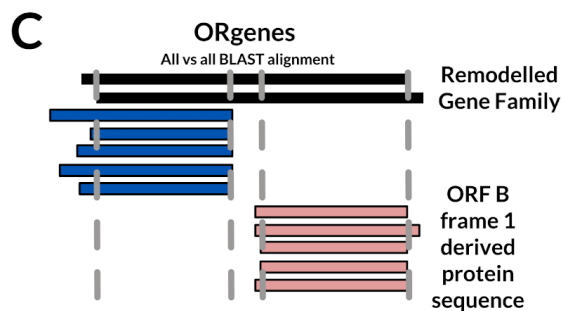
A

Frame 0   Met...Pro.Val.Met.*
Frame 1   ...Arg.Ser.Cys...
Frame 2   ...Gly.His.Val...

ATG...CCG.GTC.ATG...TAA*

Annotated ORF A

TAC...GGC.CAG.TAC...ATT

Frame -2   Thr.*.Pro...
Frame -1   Try.Met.Thr...
Frame -0   Leu.His.Asp.Arg...His

Frame 0   Met...Ala.Cys.Cys...*
Frame 1   ...Arg.Val.Ala...*
Frame 2   ...Val.Leu...*

ATG...GCG.TGT.TGC...TAA*

Annotated ORF B

TAC...CGC.ACA.ACG...ATT

Frame -2
Frame -1
Frame -0

B   In-frame remodelled genes

All vs all BLAST alignment

Remodelled Gene Family

ORF A derived protein sequence

ORF B derived protein sequence

Remodelled gene family

ORF A derived component gene family

ORF B derived component gene family

ATG...CCG.GTC.ATG...GCG.TGT.TGC...TAA*

Met...Pro.Val.Met...Ala.Cys.Cys...*

C   ORgenes

All vs all BLAST alignment

Remodelled Gene Family

ORF A derived protein sequence

ORF B frame 1 derived protein sequence

Remodelled gene family

ORF A derived component gene family

Frame 1 of ORFB derived component gene family

ATG...CCG.GTC.ATG...CGT.GTT.GCX...TAA*

Met...Pro.Val.Met...Arg.Val.Ala...*

D

ORF-ALT
160 gene families

ALT-ORF
220 gene families

ALT-ALT
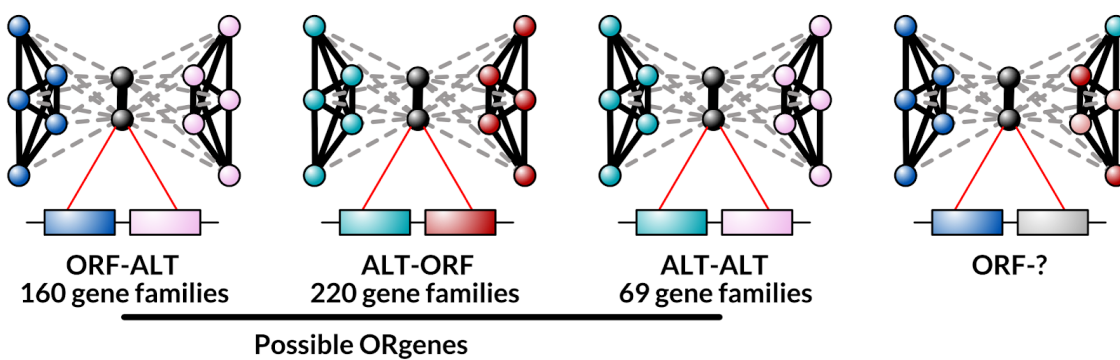69 gene families

ORF-?

Possible ORgenes

Figure 1: Defining and detecting ORgenes using sequence similarity networks. A) A mock example showing two different genes, ORFA (Blue) and ORFB (Red), and all six of their potential reading frames (where Frame 0 is the annotated protein-coding reading frame). B) Detecting a traditional remodelled gene. A remodelled gene or gene family is detected in sequence similarity networks based on its partial similarity to two or more other genes (or components), where the region of partial similarity does not overlap. In this case the remodelled gene family (in black) has partial similarity to described ORFs in family A (in blue) and in family B (red). The region of similarity between ORFA with the remodelled gene family, and between ORFB and the remodelled gene family does not overlap. The example remodelled sequence at the bottom of this panel indicates the result of in-frame remodelling involving ORFA and ORFB. C) In an ORgene, one part of the remodelled gene family has partial similarity to sequences derived from an "alternative-frame" of another gene family rather than the natural reading frame. In this case the C-terminal end of the remodelled gene family has partial similarity to the sequence encoded by frame 1 of ORFB. The example remodelled sequence at the bottom of this panel indicates the result of out-of-frame remodelling involving ORFA and ORFB. D) Possible types of ORgenes defined in this study are orientated with the most N-terminal component on the left and the most C-terminal component on the right. In an "ORF-ALT" gene, the C-terminal component is derived from an alternative frame. In an "ALT-ORF" gene, the N-terminal component is derived from an alternative frame, and in an "ALT-ALT" gene both detected components are derived from an alternative frame.

Of 767 predicted ORgene families, 709 are simple cases in the sense that they only combine two distinct components, of which at least one is out-of-frame with respect to existing genes or domains. In 260 of the ORgene families with two distinct components, the "ALT" derived component includes sequences derived from several different alternative reading frames and cannot be assigned to a single frame of origin. The remaining 449 ORgene families in which a clear frame of origin can be assigned to the "ALT" component will be the subject of future discussion. The majority (380) of these simple ORgene families include one ORF component, and one out-of-frame component (Figure 1D). In terms of Nter-Cter orientation, there is a slight but significant preference for an ALT-ORF orientation of components (220 gene families; column 1 of the heatmap in Figure 2), with ORF components mapping closer to the C-terminal end of the remodelled gene, as compared to the opposite ORF-ALT orientation (160 gene families; row 1 of the heatmap in Figure 2) (Chi2 test, P. value $2x10^{-3}$).

The C-terminal location of alternative-frame derived components in some ORgenes might be compatible with the origin of these genes by out-of-frame fusion in the C-terminal end of an existing ORF. Intuitively, a C-terminal position of the ALT region would possibly provide a good structure for the maintenance of the ORgene, given that the existing ORF's start codon and transcriptional/translational control would be present in the upstream region. Similar to the idea that fortuitous transcription can lead to *de novo* gene formation (Cai et al. 2008; Kaessmann 2010), the integration of an out-of-frame sequence at the C-terminus of an existing gene and its subsequent expression, if not immediately deleterious, might lead to new gene formation, 'by extension' (Bornberg-Bauer et al. 2015). An analogous mechanism of C-terminal extension of existing ORFs by chromosomal rearrangement, placing previously untranscribed regions downstream of existing ORFs, was identified as a potential route for the creation of 51 new genes by chromosomal rearrangement in *Drosophila yakuba* (Stewart and Rogers 2019). Similarly, recent studies have highlighted the existence of population level variations in the position of the stop codon in genes in Drosophila which would create a C-terminal extension. While many such variations were considered deleterious, a small subset spread rapidly to become the dominant allele in a population, including several examples that included relatively long extensions to protein coding sequences (Lee and Reinhardt 2012).

However, most ORgenes were found in the ALT-ORF orientation, andmore complex steps may be required for their maintenance within the genome than those proposed for an ORF-ALT orientation. While we can only speculate as to the significance of this finding, it is possible that adding an ALT region at the N-terminus on an existing ORF may be an unanticipated way to change a gene transcriptional control, or to provide a genuinely novel gene with a genuinely novel expression pattern. Previous studies have suggested new genes tend to have lower expression levels (Donoghue et al. 2011; Palmieri et al. 2014; Zhao et al. 2014), potentially owing to the deleterious effects of the expression of new peptides (e.g. accumulation of misfolded proteins (Monsellier and Chiti 2007; Koonin and Wolf 2010)). The prevalence of an ALT derived component at the N-terminus may indicate that ORgenes preserving the same transcriptional control as its parent ORFs gene, yet unable to fulfil the same functions, may be counterselected. We found no significant differences between the levels of transcription (as measured by RPKM) or translation (as measured RCV) between ORF-ALT and ALT-ORF orientated genes in the reanalysis of RNAseq and RiboSeq data from *E. coli* MG1655, O157:H7 str. Sakai and O157:H7 str. EDL933 (Wang et al. 2015; Hücker et al. 2017; Neuhaus et al. 2017)(SI Figure 5; SI Table 1). The only significant difference in expression shared across all genomes was between the

10

level of transcription between in-frame remodelled genes and other genes (Adjusted P. value 0.0), and this was matched by a significant difference in RCV in both O157:H7 strains (SI Table 1). Thus, ALT-ORF ORgenes are not expressed at measurably different levels to other genes in the dataset, despite the apparent hurdles of acquiring new start codons and transcriptional start sites.

The remaining 69 predicted ORgene families with two domains have an ALT-ALT component architecture with two different predicted components of alternative frame origin (Figure 1D; Figure 2). While 12 of these ALT-ALT ORgenes have a relatively narrow host distribution, i.e. being found in less than three different genomes, there are exceptions to this. For instance, the predicted inner-membrane protein YMFA (F19715; SI Table 2), an ALT-ALT gene, is found in 359 *E.coli* genomes. Likewise, two ALT-ALT genes are universally conserved across the dataset, found in 524 genomes. These gene families encode members of ancient gene families, with a broad distribution outside of *E. coli*: *a* predicted Thiosulphate transport protein of the ABC transport family (F41228; SI Table 2) and a predicted oligopeptidase of the M3 peptidase family (F27662; SI Table 2). We investigated whether the distribution of the source ORFs for the ALT components were similarly broadly distributed. For the ABC transport family (F41228), the source of both ALT components are narrowly distributed, found in only 3 or 11 different genomes respectively. However, for both YMFA and the M3 peptidase family, the source of one ALT component had a narrow distribution (2-4 genomes), while the source of the other had a much broader distribution (437 and 521 genomes respectively). In the case of YMFA, the broadly distributed ALT component is found in more genomes than the ORgene itself, and while they coincide in 344 genomes, the ORgene is found without the source of its ALT component in 15 genomes, and vice versa the source of the ALT component is found without the ORgene in 93 genomes. The annotation of members of these ancient families as "ALT-ALT" ORgenes suggests that in at least some cases, ORgenes are (as they should be) detected due to out-of-frame fission rather than fusion events in the host genomes. Similar processes could explain the annotation "ALT-ORF" components. In these cases, it may be the fissioned out-of-frame components of these ancient ORgene families that go on to produce novel ORFs and peptides in *E. coli*, rather than ORgenes themselves, or a combination of overprinting and fission.
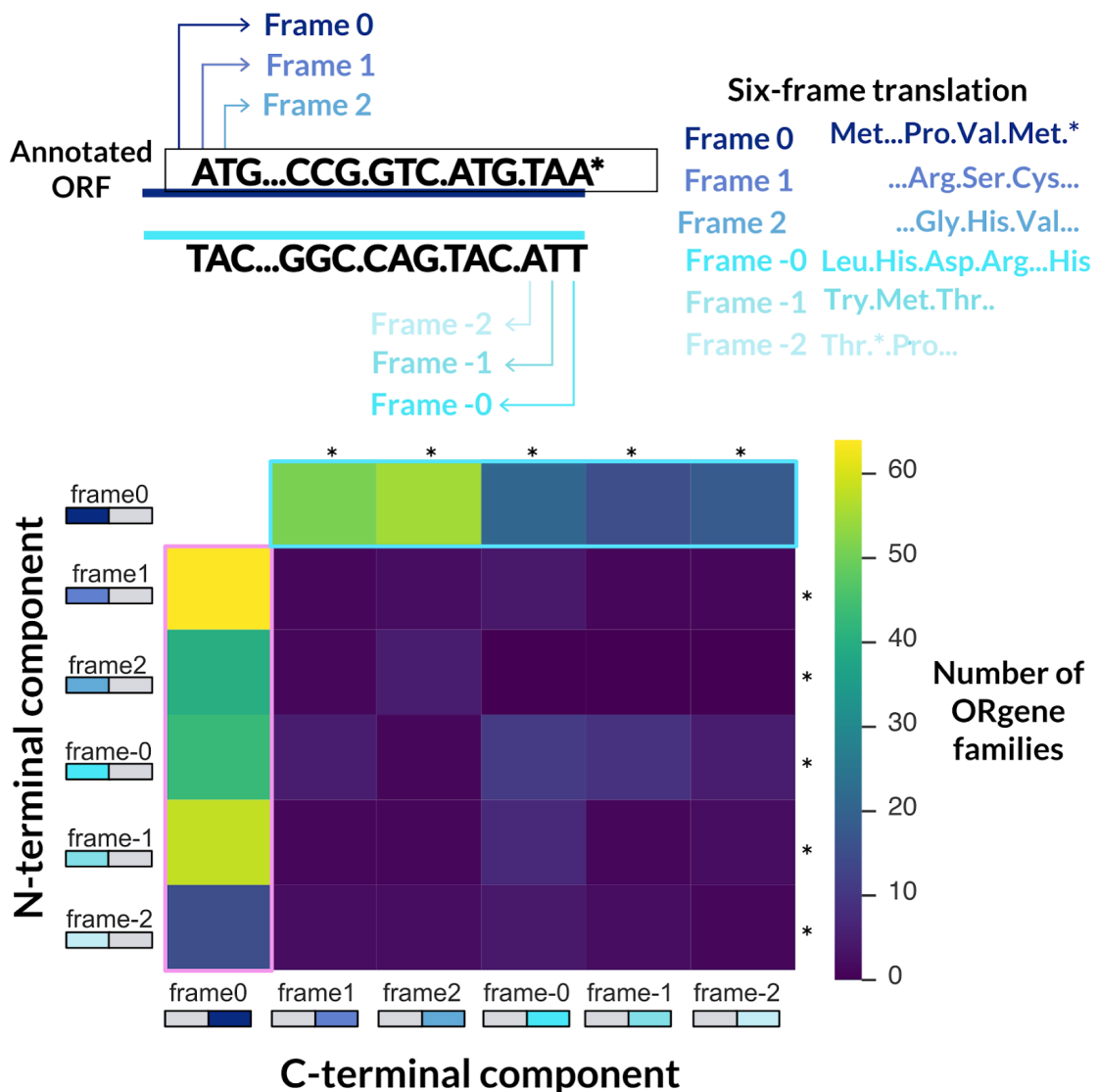
11

Figure 2: Strong biases observed in the composition of ORgenes. The component combinations in simple two-component remodelled gene families are shown in a heatmap. Top shows a short example ORF sequence (as seen in Figure 1) translated into 6 reading frames. Rows indicate the most N-terminal component, and columns indicate the most C-terminal component. The blank "frame0-frame0" position at the top left of the grid represents the 1,216 "traditional" in-frame remodelled genes. The majority of ORgenes include at least one "frame 0" component, corresponding to an annotated ORF, and are boxed in cyan and magenta. ORF-ALT genes, where the "frame0" component is the most N-terminal, are highlighted by a cyan box. ALT-ORF genes are highlighted by a magenta box. An asterisk marks rows and columns derived from an alternative reading frame.

ALT frame derived regions of ORgenes are short and disordered

Given our hypothesis that the "ALT" regions of ORgenes were not previously protein coding, we would predict that their contribution to ORgenes would be shorter than that of "ALT" components, since they are likely to initially encode short peptides interrupted by start-codons. Comparing the proportion of an ORgene covered by the ORF region with that covered by the ALT region supports this idea (Figure 3A). By only considering ORF-ALT and ALT-ORF in ORgenes with two different domains, we directly compared the lengths of the ORF derived segment of the ORgene to the ALT derived segment of the same ORgenes (Figure 3B and 3C). This confirms that ALT regions tend to be significantly shorter and cover less than the total length of ORgenes than ORF regions (Wilcoxon signed-rank test; P. values=0.0 for each comparison).

Previous analyses have suggested both overlapping genes (Rancurel et al. 2009) and young genes may have a higher intrinsic structural disorder (Wilson et al. 2017; Willis and Masel 2018), with the suggestion that new genes are more likely to successfully form from non-coding sequences pre-adapted to gene formation (Masel 2006). As such, we investigated whether ALT regions encoded potential disordered protein regions using IUPRED2 and ANCHOR2 scores generated by IUPRED2A (Mészáros et al. 2018; Erdős and Dosztányi 2020). For both scores, a cutoff of >0.5 is used to indicate disordered regions. Significantly more ALT derived ORgene segments were found to be disordered than ORF-derived segments based on IUPRED2 scores (Figure 3D; Chi2 test P. value 1.44e-95), with a total of 684 ALT derived segments of ORgenes identified as disordered compared to 110 ORF derived segments (from a total of 10,214 ORgenes), and ANCHOR2 scores (Figure 3E Chi2 test P. value 4.69e-34), with 397 ALT derived segments predicted as including disordered binding domains compared to 94 ORF sequences. We also used SEG-HCA (Faure and Callebaut 2013) to predict whether either ORF or ALT regions were more likely to overlap with H2CD regions of ORgenes, which are predicted to be foldable regions (Faure and Callebaut 2013; Bitard-Feildel and Callebaut 2017). In this case, ORF regions were significantly more likely to include entire H2CD regions (1,124 regions in total) than ALT regions (399 regions in total) of ORgenes (Chi2 test P. value 4.1x10[-78]), however the presence of foldable elements in some ALT regions raises the possibility that out-of-frame events are involved in the origin of novel foldable domains.

Finally, we confirmed the translated status of component genes that contributed the ALT frame to ORgenes detected in our analysis, using their original annotated open reading frame. 293 of these genes were translated in one of the three previously discussed

13

RNASeq/RiboSeq datasets, representing 67% of all ALT component sources found in datasets. This supports our argument that the ALT component contributing to ORgenes come from an alternative reading frame of genuine ORFs rather than misannotated untranslated regions of the genome.
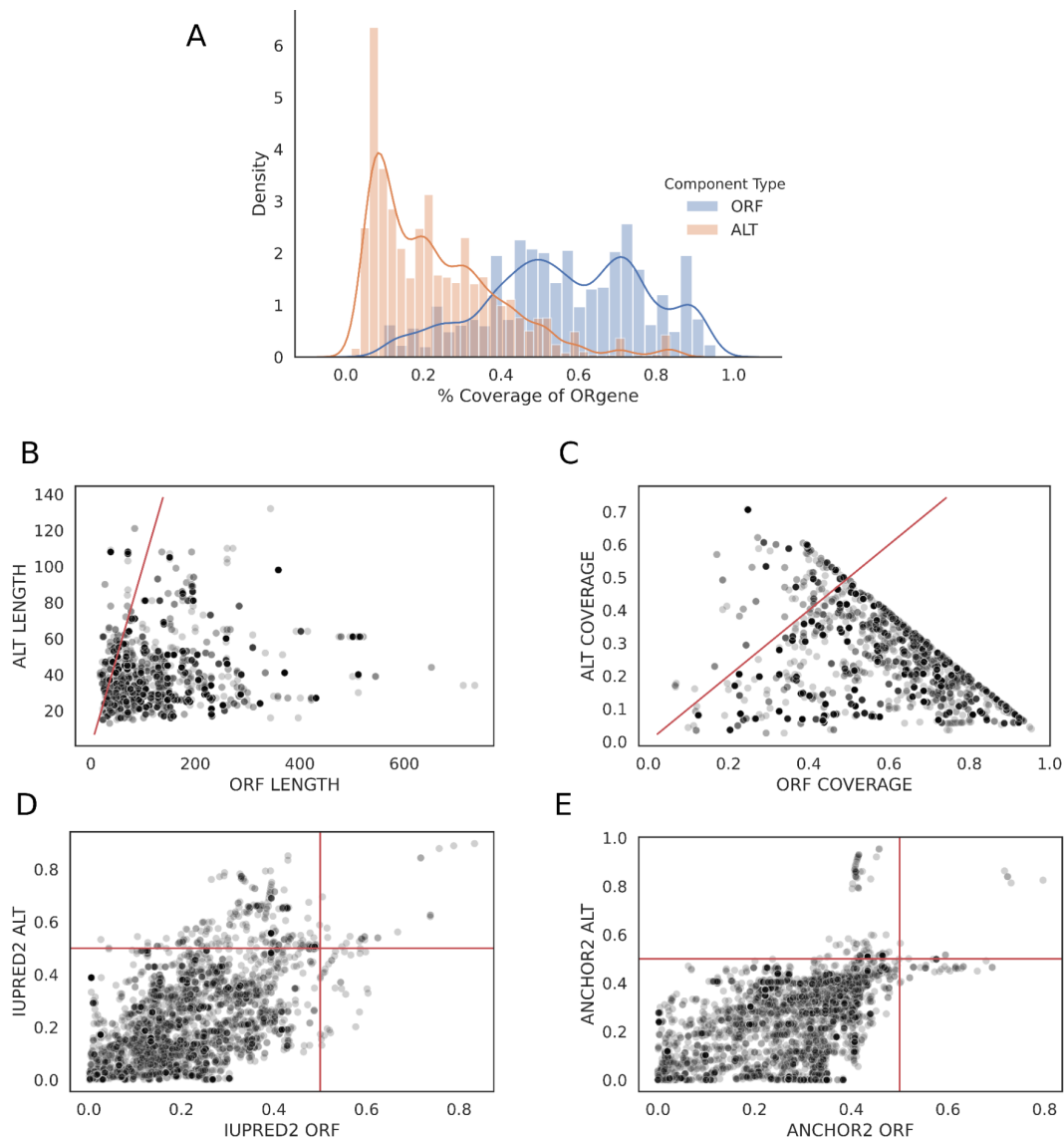

Figure 3: The ALT frame derived components of ORgenes are short and disordered. A) Comparing the proportion of ORgenes covered by ORF derived regions to ALT derived regions for all ORgenes in the dataset. B) For all ORF-ALT and ALT-ORF genes, the length of ORF derived regions (x-axis) compared to the length of ALT derived regions (y-axis) from the same ORgene. C) The proportion of an ORgene that corresponds to an ORF region (x-axis) compared to the proportion corresponding to an ALT region (y-axis) in ORF-ALT and ALT-ORF genes. In B and C, the red line indicates x=y, where both ORF and ALT regions

14

are of equal length or coverage. D and E) The average D) IUPRED2 score and E) ANCHOR2 score of all residues within an entire ORF region (x-axis) or an entire ALT region (y-axis), giving an indication as to whether these regions include D) intrinsic disordered regions or E) disordered binding regions. Here, red lines indicate the threshold (0.5) over which regions are considered disordered for these metrics.

Evidence that ORgenes are *bona fide* genes

We analysed the taxonomic distribution of ORgenes across *E. coli* genomes to determine whether and which of these genes affected by out-of-frame processes had been particularly evolutionarily successful in this lineage. If ORgenes are recent innovations, we expect their taxonomic distribution to be limited, however this limited distribution also means that it is difficult to rule out genome annotation or assembly errors, such as indels leading to frameshifts in annotated genes, as factors in their detection. Predicted ORgenes and/or their components that are found in multiple *E. coli* genomes are less likely to originate from independent annotation errors, as their detection would require a shared annotation error. As such, only remodelled gene families with at least two genes, and whose direct neighbours in the network are gene families that also included at least two genes, were considered in all analyses (see Methods). Out of 767 ORgene families fulfilling this condition, 115 are found in three or fewer genomes, compared to 156 traditional in-frame remodelled gene families, and ~33% of all gene families (10,091) within our sample of *E. coli* genomes. This is a very stringent condition, and it is worth noting that the 1,294 singleton ORgene families excluded by the filters used in this study may include *bona fide* remodelled genes. If singleton ORgenes were retained, up to 8.3 % of the *E. coli* pangenome could be affected by out-of-frame events, which at first sight supports the idea that the ORgenes may include *E. coli* specific innovations, detected as part of ongoing gene creation.

If sequencing errors have significantly impacted our predictions, we might expect to identify lower-quality genomes that are enriched in predicted ORgenes. Reassuringly however, the numbers of predicted in-frame remodelled genes and of ORgene families per *E. coli* genome are normally distributed and correlate strongly with the sizes of these genomes (as measured by the number of genes), with no clear outliers of "low quality" (as assessed by sequencing depth) (Figure 4). Moreover, 430 ORgenes belonging to 167 ORgene families were identified as transcribed and translated (defined by RCV >0.35) in the genomes of *E. coli* MG1655, O157:H7 str. Sakai and O157:H7 str. EDL933. Additionally, of the 439 genes that acted as sources for ALT components found in ORgenes and present in these three genomes, 293 (67%) were identified as translated. The expression and

translation of these genes supports their annotation as real protein coding ORFs, but does not help us to distinguish between ORgenes and misannotated remodelled genes that may include indels due to errors in the genome assembly. The alignment of illumina data associated with an RNAseq project (Sastry et al. 2019) to the *E. coli str. K-12 substr. MG1655* genome and subsequent variant calling identified no predicted indels in any of the predicted ORgenes, reinforcing our confidence that our stringent predictions identified *bona fide* genes rather than annotation errors.
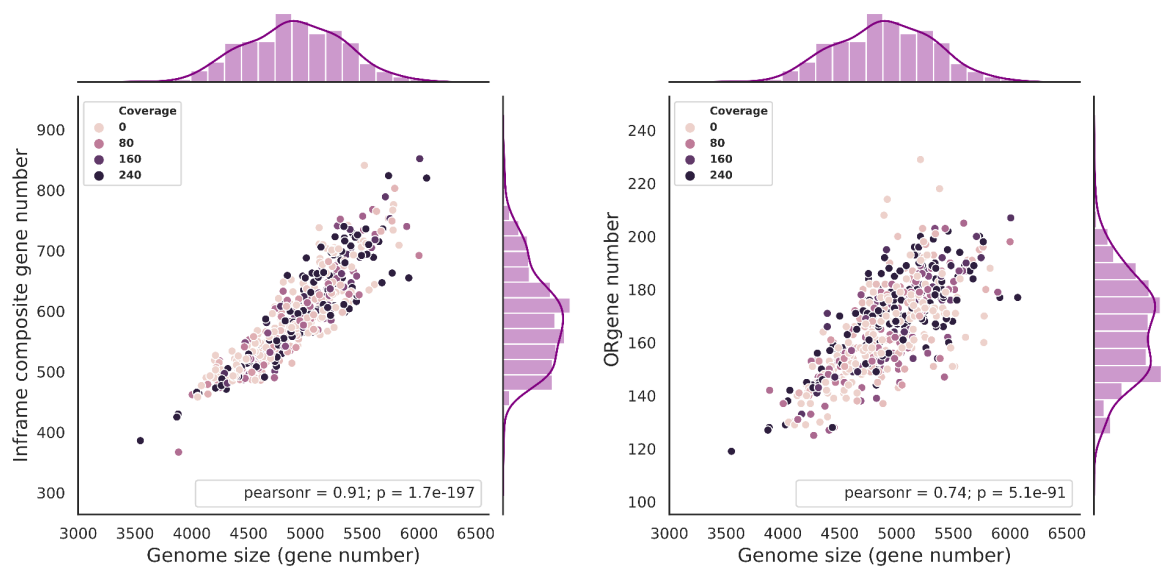

Figure 4: The distribution of remodelled gene families supports their annotation as *bona fide* genes. The number of predicted remodelled genes per *E. coli* genome, for genomes in which sequencing coverage was reported in the NCBI database. For each genome in the dataset, the number of genes per genome (x-axis) is compared to the number of in-frame remodelled gene families (left panel) or ORgene families (right panel) in those genomes (y-axis). Colours of points represent the reported sequencing coverage of those genomes. There is a strong positive correlation between genome size and the number of predicted remodelled genes.

Functional biases in ORgene families and their components

Based on EggNog (Huerta-Cepas et al. 2016) functional annotations of all predicted remodelled genes, unsurprisingly, many traditional in-frame (530; 43% of all in-frame remodelled gene families) and ORgene families (480; 63% of all ORgene families) have

unknown roles (Figure 5A; SI Table 3). These genes either have no hits in the EggNog database, or they only hit genes described as having a poorly characterised function. ORgene families are enriched in gene families with no hits in the EggNog database (350 ORgene families; 45.6% of all ORgene families; FDR adjusted P. value 9.8×10^{-56}) when compared to in-frame remodelled gene families (205 gene families; 16.8%), while depleted in genes with hits in the EggNog database of poorly characterised function (FDR adjusted P. value 2.45×10^{-2}), indicating that ORgenes remain largely uncharacterised, a logical finding for novel genes (Figure 5). Still, a subset of five ORgene families include genes that have been identified as essential in *E. coli (Baba et al. 2006; Yamazaki et al. 2008; Goodall et al. 2018)* SI Table 2. These examples include the peptide release factors A and B, and the *E. coli* SSB (single-stranded binding protein). As with the previously highlighted and broadly distributed ALT-ALT gene families, these five essential ORgene families belong to ancient gene families, suggesting that these families are detected as ORgenes because their ALT component gave birth to new ORFs in *E. coli* by an out-of-frame process (e.g. out-of-frame gene fission or overprinting of the ORgene, essential, gene family).

Interestingly, functionally characterised ORgenes show singular trends with respect to in-frame remodelled genes. There are significantly fewer ORgene families with predicted roles in cellular processes and signalling (FDR adjusted P. value 8.04×10^{-3}) and information storage and processing (FDR adjusted P. value 5.28×10^{-3}) than traditional in-frame remodelled genes, particularly in COG category L (Replication, recombination and repair; FDR adjusted P. value $3.96. \times 10^{-7}$) (Figure 5). Thus, ORgenes display unusual functional biases compared to traditional in-frame remodelled genes.
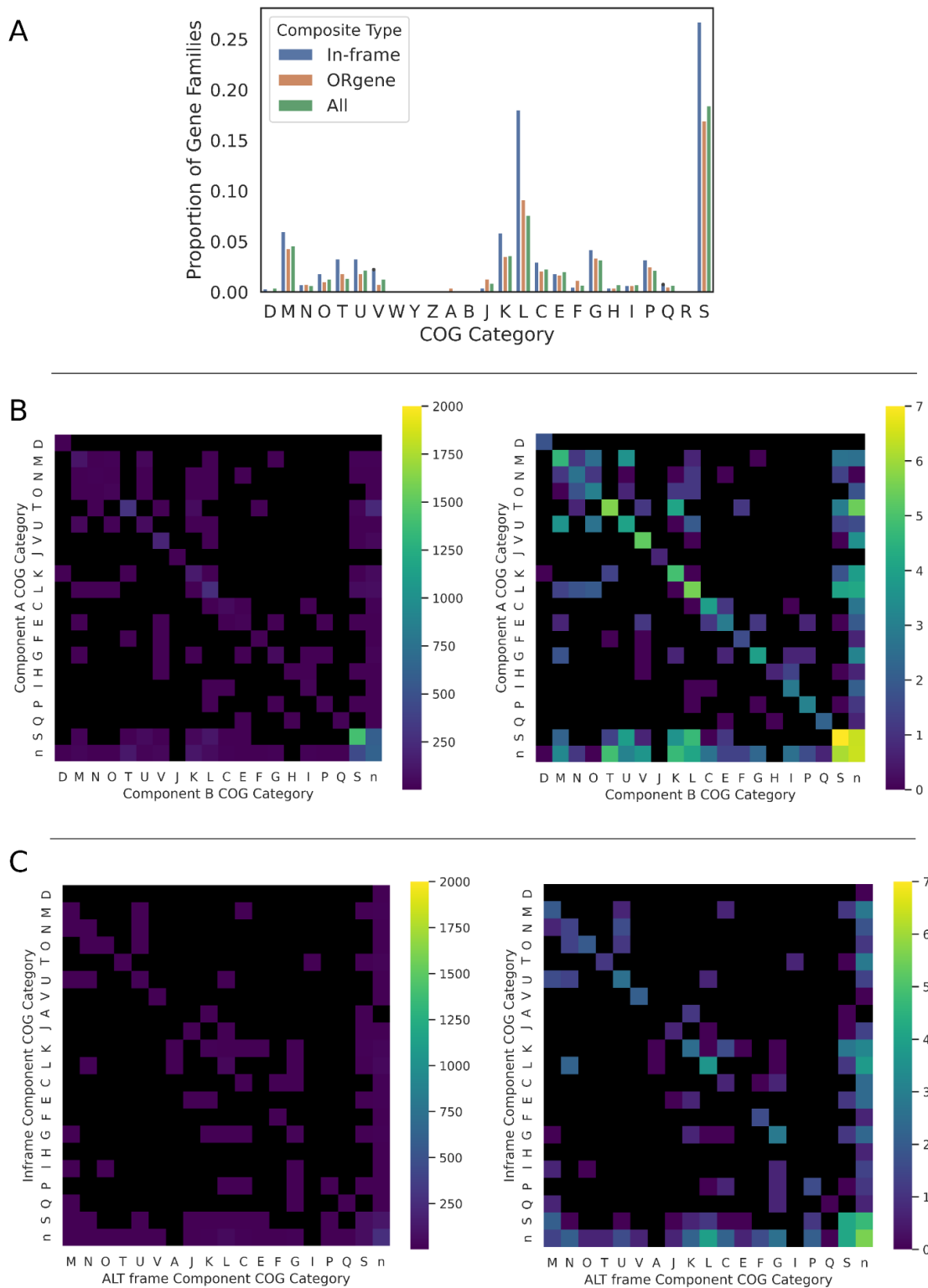
Figure 5: Predicted remodelled genes show unusual biases in functional predictions. Panel A: The proportion of gene families assigned to a particular COG functional category from all traditional in-frame (blue) and ORgene families (orange) to all gene families in the dataset (green). Panel B: Heatmaps indicating the predicted functions (COG categories) of

the components of traditional in-frame remodelled gene families with a total of two predicted components, A and B (N-terminal and C-terminal). Panel C: For all ORgenes, the predicted functions of the in-frame/ORF component of ORgene families (rows), and of the "parent" ORF of the ALT-frame derived component (columns). For panels B and C, the left heatmap indicates the raw number of gene families, while right indicates the log of that number. Black squares indicate zero values.

For ORgenes composed of an ALT (out-of-frame) component and of an ORF (in-frame) component, we analysed the predicted functional consistency of the origins of these two components; for the functional annotation of the ALT components, using that of the parent ORF from which it was derived. Previous studies on in-frame remodelled genes have supported a tendency for the combination of components with similar functions (i.e. fusion is likely to combine two components that already have similar functions, or fission is likely to split a gene into two components with similar functions) (Yanai et al. 2001). We confirmed this trend for the predicted in-frame remodelled genes in our analysis (Figure 5; Panel B), that showed a strong preference for the combination of functionally-like components. In ORgenes, we might expect less correlation between the predicted function of the ORF (frame 0) and that of the parent ORF for the ALT component, since a parent ORF of an ALT sequence and this ALT sequence encode very different proteins as a result of a frameshift. However, we still observe a similar preference towards "like-like" functional combinations between the parent of the "alternative frame" sequence and the function of the ORF derived components of a given ORgene (Figure 5; Panel C). We interpret this functional consistency as reflecting a positional bias. In a similar way that fusion/fission events preferentially occur between neighbouring genes, we might expect the same to be true of out-of-frame fusion/fission events. The organisation of the bacterial genome into operons of functionally related genes would then lend itself to the combination of components from functionally related source genes.

Evolutionary origins of ORgene families

Remarkably, at least 37 ORgene families have similarity to transposable element related proteins (SI Table 2). These results are consistent with the ability of mobile genetic elements to jump into genomes, an ability which has previously been shown to contribute to the formation of new fusion genes (Bennetzen 2000; Lai et al. 2005; Jiang et al. 2011) and has already been associated with the origins of new genes (Feschotte 2008; Jangam et al. 2017; Joly-Lopez and Bureau 2018), including the proposed *de novo* exogenisation of non-coding regions in eukaryotes (Cordaux et al. 2006; Schmitz and Brosius 2011) and

overprinting (McVeigh et al. 2000; Balabanov et al. 2012). Further, 53 ORgene families are predicted to be phage associated (SI Table 2). This result is consistent with the common reports of overlapping gene pairs and overprinting in viral genomes (Sabath et al. 2012; Rancurel et al. 2009). Previous studies have highlighted the important role that horizontal gene transfer from viruses can play in the evolution of *E. coli* (and other bacterial species) (Frazão et al. 2019). These 53 ORgenes may have originated from overlapping sequences integrated into the bacterial chromosome via phage integration. If so, our work highlights a potentially important under-appreciated impact of phage mediated gene transfer on bacterial evolution: the introduction of new genes potentially formed by out-of-frame remodelling processes in viruses to the bacterial genome. Overall, mobile genetic elements and gene transfer may be involved in the detection of at least 12% of predicted ORgene families (95 of 767 gene families).

Overall however, structural genomic analyses of ORgenes suggest that many ORgenes originate by overprinting. In the described processes for creating new genes by overprinting (outlined in SI Figure 1) it is implicit that at some point in the evolutionary history of a remodelled gene, it must overlap with another gene. Thus, the simplest explanation for the observation of overlap between any ORgene in a gene family and the source of its ALT component in at least one genome is overprinting. This was true for at least one ORgene in 317 ORgene families (~41% of all ORgene families), of which 178 are ORF-ALT or ALT-ORF ORgene families. The ORF-ALT orientation of 57 of these gene families is compatible with the gain of a new out-of-frame stop codon, while the ALT-ORF orientation of 121 ORgene families is compatible with the gain of a new out-of-frame start codon (SI Figure 1). This is possibly an underestimate, as differential loss and retention of genes following ORgene formation may mean that an overlap is never directly observed (SI Figure 4). The proportion of ORgenes in a family that overlap with an ALT component varies across ORgene families, but is low (<20%) in the majority of cases. There are only 9 gene families in which all ORgenes overlap with an ALT component (SI Figure 6). To investigate overlaps between ORgenes and ALT components further, we define three types of overlap. First, and previously highlighted as most common in prokaryotic genomes (Rogozin et al. 2002; Fukuda et al. 2003), are unidirectional (same-strand overlaps). Second, there are convergent overlaps (opposite strand overlaps at the C-terminus), and third, divergent overlaps (opposite strand overlaps at the N-terminus). We also quantified the length of the overlap between ORgenes and ORFs associated with the ALT components, comparing these with all overlapping gene pairs in the sampled *E. coli* genomes.

In *E. coli* genomes, there is a strong preference for unidirectional overlaps (418,760 gene pairs) compared to convergent (53,472 gene pairs) and divergent (20,696 gene pairs) overlaps. This remains true for longer overlaps, e.g. for the 2,800 gene pairs whose overlap covers >50% of the length of the shortest sequence in the pair (Table 1): 1,164 of these overlaps are unidirectional, 864 are convergent and 772 are divergent, consistent with observations from other prokaryotic genomes (Rogozin et al. 2002; Fukuda et al. 2003) (Table 1). Despite the general rarity of long overlaps, in total 317 of the 767 predicted ORgene families (and 3,196 gene pairs in total) include at least one gene that, in at least one *E. coli* genome, directly overlaps with the source of the ALT sequence (Table 1). For 107 ORgene families (and 226 gene pairs), the overlap covers >50% of the length of the shortest sequence in the pair, accounting for ~8% of all long overlaps observed in the *E. coli* pangenome (Table 1). In both cases, the preference for unidirectional overlaps is retained. The pronounced physical overlap between many ORgenes and their ORFs associated with their ALT components supports their origin by overprinting.

Moreover, we also defined the overlaps between ORgenes and ORFs associated with an ALT component by reading phases, describing how the codon positions of these overlapping genes align against one another, which in turn can impact the degree of mutual evolutionary constraint on the overlapping gene pair (Krakauer 2000; Rogozin et al. 2002). For example, in a study of convergent overlaps in bacterial genomes, a bias for the 123:132 phase has been identified, meaning that, for overlapping genes A and B, the second codon position of gene A aligns with the third and degenerate codon position in gene B (Rogozin et al. 2002). A broader study encompassing both the most common tandem (unidirectional) overlaps, and antiparallel (both convergent and divergent) overlaps observed a strong preference for 123:312 phase overlaps within the tandem overlaps, and a preference for both 123:132 and 123:213 phase overlaps within the antiparallel overlaps, with these occurring at similar frequencies. Most experimentally supported examples of convergent overlapping genes pairs that have provided evidence for overprinting in bacteria are in the 123:213 phase (Delaye et al. 2008; Tunca et al. 2009; Fellner et al. 2015; Hücker, Vanderhaeghen, Abellan-Schneyder, Scherer, et al. 2018; Hücker, Vanderhaeghen, Abellan-Schneyder, Wecko, et al. 2018). Likewise, the phases of longer overlaps between ORgenes and the ORFs associated with their ALT components match these biases associated with overlapping genes, with a preference for 123:312 phase unidirectional overlaps (70 gene pairs) and the 123:132 phase (51 gene pairs) for antiparallel (convergent and divergent overlaps). Taken together, the presence of overlaps observed between ORgenes and the ORFs associated with ALT components, as well as the types and the

21

translation phase of their overlaps, is consistent with the possibility that a subset of ORgenes originated by overprinting, from overlapping genes following usual trends in *E. coli* genomes.

Overlap Type	Phase	All *E. coli* Overlapping Gene Pairs	ORgene Overlapping Gene Pairs	All *E. coli* Long Overlaps	ORgene Long Overlaps
Unidirectional	"123:123"	152 (0.03%)	NA	92 (3.3%)	NA
	"123:231"	304,494 (62%)	761 (23.9%)	502 (18%)	48 (20.3%)
	"123:312"	114,114 (23%)	1656 (52%)	570 (20.4%)	70 (29.7%)
Convergent	"123:321"	14,956 (3%)	122 (3.8%)	386 (13.8%)	23 (9.7%)
	"123:132"	8,435 (1.7%)	68 (2.1%)	360 (12.9%)	14 (5.9%)
	"123:213"	30,081 (6.1%)	68 (2.1%)	118 (4.2%)	13 (5.5%)
Divergent	"123:321"	3,657 (0.7%)	78 (2.4%)	67 (2.4%)	6 (2.5%)
	"123:132"	7,350 (1.5%)	84 (2.6%)	414 (14.8%)	37 (15.7%)
	"123:213"	9,689 (2%)	359 (11.3%)	291 (10.4%)	25 (10.6%)

Table 1: The types and phases of overlaps between pairs of genes in the *E. coli* pangenome, and between ORgenes and the source of their ALT components. % values are the proportion of all overlapping genes in that column.

Conclusion

Our work introduced a dedicated network method to test for the presence of a remarkable set of gene families, conserved in at least two genomes, whose evolutionary history involved out-of-frame events, in addition to more conventional in-frame remodelled genes. Applying this method to 524 genomes from *E. coli*, we found that overall, 6.3% of the gene families in the *E. coli* pangenome appeared to be remodelled in that lineage, testifying to the highly dynamic nature of prokaryotic genomes in terms of their gene composition. Moreover, we unravelled a common, unsuspected novel class of remodelled genes, the ORgenes, in *E. coli* pangenomes, amounting to 2.5% of the total pangenome. Importantly, ORgene families are distinct from the more extensively studied class of in-frame remodelled genes, as both kinds of remodelled genes show different biases in their functional compositions, and differences in their mode of origin. Specifically, our analyses of functional biases, and of positional and translation phase overlaps, support the idea that overprinting was involved in the origins of a significant subset of ORgene families, with 40% of these families overlapping with their ALT derived component in at least one genome. Therefore,

22

our study supports growing evidence that overprinting and overlapping genes may play a greater role in gene creation in bacterial genomes than previously anticipated. Further, the identification of at least 12% ORgene families of putative phage and transposon origins encourages further analysis of the role for these mobile genetic elements in the supply of new genes in bacterial evolution.

Methods

Sequence similarity network construction and analysis

The dataset used in this study included sequences from 524 complete *E. coli* genomes from the NCBI database (SI Table 1). These were all *E. coli* genomes whose assembly level was annotated as "complete" in the NCBI database and had >500 annotated protein-coding ORFs when the dataset was assembled in 2019. The predicted protein sequences from all annotated ORFs in these genomes (not including annotated pseudogenes) were clustered using CD-HIT to reduce redundancy. Two genes were clustered if they had 100% sequence identity covering over 90% of the length of both genes in the dataset. The representative sequence for each cluster was the longest, by default. The DNA sequences corresponding to the representative ORF for each cluster were then translated into 6 open reading frames, reading through any stop codons. Protein sequences from all six frames were used as both the query and database for an all-vs-all diamond BLASTp search (Buchfink et al. 2014) using the "more-sensitive" search mode, with a minimum reported E value of <1e-5, and a minimum sequence identity of 30%. The results of the sequence similarity search were filtered using cleanBlastp (part of the compositeSearch package (Pathmanathan et al. 2018)) to remove self-hits.

The sequence similarity results were parsed using CompositeSearch (Pathmanathan et al. 2018) to identify potential remodelled genes, using default parameters except setting the minimum number of sequences for a component to two, meaning that "singletons" will not be classed as components of remodelled genes in an effort to reduce the impact of false positives produced by annotation errors in a single genome. In the first step, gene families are identified as connected components in the sequence similarity network with thresholds of an E. value <1e-5, a mutual coverage of >80% and a sequence identity of >30%. CompositeSearch then uses this family assignment and the full sequence similarity network to identify potential remodelled genes and remodelled gene families (Pathmanathan et al. 2018).

We parsed the CompositeSearch results to identify every putative remodelled gene family that only included sequences derived from the coding reading frame (frame 0) of annotated ORFs. Traditional in-frame remodelled genes present component families in which all sequences are derived from the coding reading frame (frame 0) of annotated ORFs. We parsed the output of CompositeSearch to identify the new class of remodelled genes, where all protein sequences from at least one component of the 'composite' gene were derived from an "alternate" reading frame, i.e. not the frame 0 of an annotated protein sequence, but the frames -2,-1, -0, +1, or +2 of an annotated protein sequence, which we called ORgenes. Both in-frame and ORgene families were filtered for those in which both composite families and predicted component families include at least two genes (excluding singletons), and where the relationship between remodelled genes and component families are supported by edges between at least two remodelled genes (in 99.6% of the cases from two different genomes) and two components in the sequence similarity network.

Gene family annotation

Transcriptome data from *Escherichia coli str. K-12 substr. MG1655* (NCBI accession: NC_000913, corresponding to assembly GCF_000005845.2 in this dataset) from 278 RNAseq projects that included 154 different experimental conditions was acquired from (Sastry et al. 2019). This includes all transcription data for genes >100nt in length that have >10 fragments per million-mapped reads across all samples. Further, all sequencing data from the 30 samples associated with bioproject PRJNA504479 were acquired from NCBI. Each sample includes high quality RNA sequencing data from *E. coli str. K-12 substr. MG1655* grown under different conditions and sequenced on the Illumina HiSeq 4000 (Sastry et al. 2019). Data from all samples were mapped to the *E. coli str. K-12 substr. MG1655* genome using bowtie2 and the default parameters for the "very-sensitive" alignment mode (Langmead and Salzberg 2012). From a total of ~334 million paired reads, 99.44% were successfully aligned to the *E. coli* genome. All reads mapping to predicted ORgenes and 200 base flanking regions were extracted using bedtools (Quinlan and Hall 2010). Bcftools was used for variant calling and bcf/vcf filtering (Li 2011; Danecek et al. 2021), with a maximum read depth of 2,500, a ploidy of 1, and a minimum quality of 20.

Published ribosome profiling and RNAseq data from three studies focusing on *E. coli* strains MG1655 (Wang et al. 2015), O157:H7 str. Sakai (Hücker et al. 2017) and O157:H7 str. EDL933 (Neuhaus et al. 2017) grown with LB were obtained from NCBI. The genomes of all of these strains are included in our analysis of ORgenes. Reads were prepared for analysis according to an existing protocol for the analysis of ribosome profiling data

24

(Zehentner et al. 2020). Briefly, reads were trimmed using cutadapt (Martin 2011), aligned to the genomes of the appropriate strains using bowtie2 (Langmead and Salzberg 2012) and Bedtools (Quinlan and Hall 2010) was used to remove reads mapping to rRNA and tRNA from each dataset. Raw counts of reads mapping to annotated ORFs in each genome were calculated using HTSeq (Anders et al. 2015) and normalised to RPKM values using a custom python script. Translation was predicted using the Ribosome Coverage Value (RCV) (RPKM translatome / RPKM transcriptome) with an RCV >0.35 previously described as indicative of translation (Hücker et al. 2017). Log transformed RPKM and RCV values were used to test for significant differences between the transcription and translation of different categories of genes in each genome (ORgene categories outlined in Figure 1D, in-frame remodelled genes, and all other genes in the genome). The Kruskal-Wallis H-test (non-parametric ANOVA) was first used to test for any overall effect of gene category on gene transcription or translation. Then, the Mann-Whitney U Test with Bonferroni correction was used for pairwise comparison of transcription or translation levels between different gene categories (Additional Table 2).

To compare the properties of ORF compared to ALT regions of ORF-ALT and ALT-ORF ORgenes, their sequences were extracted from the dataset based on coordinates identified in compositeSearch. Intrinsic disorder was predicted using IUPRED2A in "long" mode (Mészáros et al. 2018; Erdős and Dosztányi 2020), with ANCHOR2 scores enabled. The average score of all residues in an entire ORF or ALT region, respectively, was used to estimate the disorder propensity of that region. SEG-HCA with default settings for individual sequences was used to identify sequence domains with a high density of hydrophobic clusters (H2CD), indicating that the domain may be foldable (Faure and Callebaut 2013; Bitard-Feildel and Callebaut 2017). We defined ORF or ALT regions as foldable if they included all residues in a predicted foldable domain (H2CD).

Functional predictions for protein sequences from annotated *E. coli* ORFs were made using EggNOG mapped v2 (with default parameters in "diamond" mode) (Huerta-Cepas et al. 2016). The set of putatively "essential" genes in *E. coli* was taken from (Goodall et al. 2018), including predictions from the Keio collection (Baba et al. 2006) and PEC database (Yamazaki et al. 2008). Enrichment or depletion of functional categories in ORgenes compared to traditional in-frame remodelled genes was assessed using Fisher's exact test (Fisher 1922) implemented in the SciPy python module, with Benjamini/Hochberg False Discovery Rate correction (Benjamini and Hochberg 1995) used to adjust P-values, implemented in the statsmodels python module.

25

Availability of data and materials

All data used in this study is publicly available from the NCBI genome assembly and SRR databases. The identifiers of all genomes included in the study are listed in SI Table 1. The DNA sequence of all predicted ORFs from these genomes is available at https://figshare.com/s/faebf854a7f04b48801d. Protein sequences of all gene families, including ORgenes and inframe composites, are available at https://figshare.com/s/7530e0323288dc006cd2. A standalone script used to differentiate between in-frame and ORgene families in this dataset is available at https://figshare.com/s/0d0ab7744e141b874648.

Acknowledgements

Bibliography

Anders S, Pyl PT, Huber W. 2015. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31:166–169.

Ardern Z, Neuhaus K, Scherer S. 2020. Are Antisense Proteins in Prokaryotes Functional? *Front Mol Biosci* 7:187.

Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, Datsenko KA, Tomita M, Wanner BL, Mori H. 2006. Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol. Syst. Biol.* 2:2006.0008.

Balabanov VP, Kotova VY, Kholodii GY, Mindlin SZ, Zavilgelsky GB. 2012. A novel gene, ardD, determines antirestriction activity of the non-conjugative transposon Tn5053 and is located antisense within the tniA gene. *FEMS Microbiol. Lett.* 337:55–60.

Bapteste E, O'Malley MA, Beiko RG, Ereshefsky M, Gogarten JP, Franklin-Hall L, Lapointe FJ, Dupre J, Dagan T, Boucher Y, et al. 2009. Prokaryotic evolution and the tree of life are two different things. *Biol. Direct* 4:34.

Barrell BG, Air GM, Hutchison CA. 1976. Overlapping genes in bacteriophage φX174. *Nature* 264:34–41.

Behrens M, Sheikh J, Nataro JP. 2002. Regulation of the overlapping pic/set locus in

Shigella flexneri and enteroaggregative Escherichia coli. *Infect. Immun.* 70:2915–2925.

Benjamini Y, Hochberg Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Series B Stat. Methodol.* 57:289–300.

Bennetzen JL. 2000. Transposable element contributions to plant gene and genome evolution. *Plant Mol. Biol.* 42:251–269.

Bitard-Feildel T, Callebaut I. 2017. Exploring the dark foldable proteome by considering hydrophobic amino acids topology. *Sci. Rep.* 7:41425.

Bornberg-Bauer E, Schmitz J, Heberlein M. 2015. Emergence of de novo proteins from "dark genomic matter" by "grow slow and moult." *Biochem. Soc. Trans.* 43:867–873.

Brown NL, Smith M. 1977. The sequence of a region of bacteriophage φX174 DNA coding for parts of genes A and B. *J. Mol. Biol.* 116:1–28.

Buchfink B, Xie C, Huson DH. 2014. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* 12:59–60.

Cai J, Zhao R, Jiang H, Wang W. 2008. De novo origination of a new protein-coding gene in Saccharomyces cerevisiae. *Genetics* 179:487–496.

Carter JJ, Daugherty MD, Qi X, Bheda-Malge A, Wipf GC, Robinson K, Roman A, Malik HS, Galloway DA. 2013. Identification of an overprinting gene in Merkel cell polyomavirus provides evolutionary insight into the birth of viral genes. *Proc. Natl. Acad. Sci. U. S. A.* 110:12744–12749.

Casola C. 2018. From De Novo to "De Nono": The Majority of Novel Protein-Coding Genes Identified with Phylostratigraphy Are Old Genes or Recent Duplicates. *Genome Biol. Evol.* 10:2906–2918.

Cordaux R, Udit S, Batzer MA, Feschotte C. 2006. Birth of a chimeric primate gene by capture of the transposase gene from a mobile element. *Proc. Natl. Acad. Sci. U. S. A.* 103:8101–8106.

Corel E, Lopez P, Méheust R, Bapteste E. 2016. Network-Thinking: Graphs to Analyze Microbial Complexity and Evolution. *Trends Microbiol.* 24:224–237.

Dagan T, Martin W. 2009. Getting a better picture of microbial evolution en route to a network of genomes. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 364:2187–2196.

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, et al. 2021. Twelve years of SAMtools and BCFtools. *Gigascience* [Internet] 10. Available from: http://dx.doi.org/10.1093/gigascience/giab008

Delaye L, DeLuna A, Lazcano A, Becerra A. 2008. The origin of a novel gene through overprinting in Escherichia coli. *BMC Evol. Biol.* 8:31.

Domazet-Lošo T, Carvunis A-R, Albà MM, Šestak MS, Bakaric R, Neme R, Tautz D. 2017. No Evidence for Phylostratigraphic Bias Impacting Inferences on Patterns of Gene Emergence and Evolution. *Mol. Biol. Evol.* 34:843–856.

Donoghue MT, Keshavaiah C, Swamidatta SH, Spillane C. 2011. Evolutionary origins of

27

Brassicaceae specific genes in Arabidopsis thaliana. *BMC Evol. Biol.* 11:47.

Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA. 1999. Protein interaction maps for complete genomes based on gene fusion events. *Nature* 402:86–90.

Erdős G, Dosztányi Z. 2020. Analyzing protein disorder with IUPred2A. *Curr. Protoc. Bioinformatics* 70:e99.

Faure G, Callebaut I. 2013. Comprehensive repertoire of foldable regions within whole genomes. *PLoS Comput. Biol.* 9:e1003280.

Fellner L, Bechtel N, Witting MA, Simon S, Schmitt-Kopplin P, Keim D, Scherer S, Neuhaus K. 2014. Phenotype of htgA (mbiA), a recently evolved orphan gene of Escherichia coli and Shigella, completely overlapping in antisense to yaaW. *FEMS Microbiol. Lett.* 350:57–64.

Fellner L, Simon S, Scherling C, Witting M, Schober S, Polte C, Schmitt-Kopplin P, Keim DA, Scherer S, Neuhaus K. 2015. Evidence for the recent origin of a bacterial protein-coding, overlapping orphan gene by evolutionary overprinting. *BMC Evol. Biol.* [Internet] 15. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4683798/

Feschotte C. 2008. Transposable elements and the evolution of regulatory networks. *Nat. Rev. Genet.* 9:397–405.

Fischer D, Eisenberg D. 1999. Finding families for genomic ORFans. *Bioinformatics* 15:759–762.

Fisher RA. 1922. On the Interpretation of χ 2 from Contingency Tables, and the Calculation of P. Available from: https://zenodo.org/record/1449484

Frazão N, Sousa A, Lässig M, Gordo I. 2019. Horizontal gene transfer overrides mutation in Escherichia coli colonizing the mammalian gut. *Proc. Natl. Acad. Sci. U. S. A.* 116:17906–17915.

Friedman RC, Kalkhof S, Doppelt-Azeroual O, Mueller SA, Chovancová M, von Bergen M, Schwikowski B. 2017. Common and phylogenetically widespread coding for peptides by bacterial small RNAs. *BMC Genomics* 18:553.

Fukuda Y, Nakayama Y, Tomita M. 2003. On dynamics of overlapping genes in bacterial genomes. *Gene* 323:181–187.

Fukuda Y, Washio T, Tomita M. 1999. Comparative study of overlapping genes in the genomes of Mycoplasma genitalium and Mycoplasma pneumoniae. *Nucleic Acids Res.* 27:1847–1853.

Gogarten JP, Doolittle WF, Lawrence JG. 2002. Prokaryotic evolution in light of gene transfer. *Mol. Biol. Evol.* 19:2226–2238.

Goodall ECA, Robinson A, Johnston IG, Jabbari S, Turner KA, Cunningham AF, Lund PA, Cole JA, Henderson IR. 2018. The Essential Genome of Escherichia coli K-12. *MBio* 9:e02096–17.

Grassé PP. 1977. Evolution of Living Organisms.

Guerzoni D, McLysaght A. 2016. De Novo Genes Arise at a Slow but Steady Rate along the

Primate Lineage and Have Been Subject to Incomplete Lineage Sorting. *Genome Biol. Evol.* 8:1222–1232.

Hücker SM, Ardern Z, Goldberg T, Schafferhans A, Bernhofer M, Vestergaard G, Nelson CW, Schloter M, Rost B, Scherer S, et al. 2017. Discovery of numerous novel small genes in the intergenic regions of the Escherichia coli O157:H7 Sakai genome. *PLoS One* 12:e0184119.

Hücker SM, Vanderhaeghen S, Abellan-Schneyder I, Scherer S, Neuhaus K. 2018. The Novel Anaerobiosis-Responsive Overlapping Gene ano Is Overlapping Antisense to the Annotated Gene ECs2385 of Escherichia coli O157:H7 Sakai. *Front. Microbiol.* 9:931.

Hücker SM, Vanderhaeghen S, Abellan-Schneyder I, Wecko R, Simon S, Scherer S, Neuhaus K. 2018. A novel short L-arginine responsive protein-coding gene (laoB) antiparallel overlapping to a CadC-like transcriptional regulator in Escherichia coli O157:H7 Sakai originated by overprinting. *BMC Evol. Biol.* 18:21.

Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, Rattei T, Mende DR, Sunagawa S, Kuhn M, et al. 2016. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.* 44:D286–D293.

Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS. 2009. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* 324:218–223.

Jachiet P-AA, Colson P, Lopez P, Bapteste E. 2014. Extensive gene remodeling in the viral world: new evidence for nongradual evolution in the mobilome network. *Genome Biol. Evol.* 6:2195–2205.

Jachiet P-A, Pogorelcnik R, Berry A, Lopez P, Bapteste E. 2013. MosaicFinder: identification of fused gene families in sequence similarity networks. *Bioinformatics* 29:837–844.

Jacob F. 1977. Evolution and tinkering. *Science* 196:1161–1166.

Jain A, Perisa D, Fliedner F, von Haeseler A, Ebersberger I. 2019. The Evolutionary Traceability of a Protein. *Genome Biol. Evol.* 11:531–545.

Jangam D, Feschotte C, Betrán E. 2017. Transposable Element Domestication As an Adaptation to Evolutionary Conflicts. *Trends Genet.* 33:817–831.

Jiang N, Ferguson AA, Slotkin RK, Lisch D. 2011. Pack-Mutator-like transposable elements (Pack-MULEs) induce directional modification of genes through biased insertion and DNA acquisition. *Proc. Natl. Acad. Sci. U. S. A.* 108:1537–1542.

Johnson ME, Viggiano L, Bailey JA, Abdul-Rauf M, Goodwin G, Rocchi M, Eichler EE. 2001. Positive selection of a gene family during the emergence of humans and African apes. *Nature* 413:514–519.

Johnson ZI, Chisholm SW. 2004. Properties of overlapping genes are conserved across microbial genomes. *Genome Res.* 14:2268–2272.

Joly-Lopez Z, Bureau TE. 2018. Exaptation of transposable element coding sequences. *Curr. Opin. Genet. Dev.* 49:34–42.

Kaessmann H. 2010. Origins, evolution, and phenotypic impact of new genes. *Genome Res.* 20:1313–1326.

Knowles DG, McLysaght A. 2009. Recent de novo origin of human protein-coding genes. *Genome Res.* 19:1752–1759.

Koonin EV, Wolf YI. 2010. Constraints and plasticity in genome and molecular-phenome evolution. *Nat. Rev. Genet.* 11:487–498.

Krakauer DC. 2000. Stability and evolution of overlapping genes. *Evolution* 54:731–739.

Kummerfeld SK, Teichmann SA. 2005. Relative rates of gene fusion and fission in multi-domain proteins. *Trends Genet.* 21:25–30.

Lai J, Li Y, Messing J, Dooner HK. 2005. Gene movement by Helitron transposons contributes to the haplotype variability of maize. *Proc. Natl. Acad. Sci. U. S. A.* 102:9068–9073.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9:357–359.

Lee YCG, Reinhardt JA. 2012. Widespread polymorphism in the positions of stop codons in Drosophila melanogaster. *Genome Biol. Evol.* 4:533–549.

Leonard G, Richards TA. 2012. Genome-scale comparative analysis of gene fusions, gene fissions, and the fungal tree of life. *Proceedings of the National Academy of Sciences* 109:21402–21407.

Levine MT, Jones CD, Kern AD, Lindfors HA, Begun DJ. 2006. Novel genes derived from noncoding DNA in Drosophila melanogaster are frequently X-linked and exhibit testis-biased expression. *Proc. Natl. Acad. Sci. U. S. A.* 103:9935–9939.

Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27:2987–2993.

Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D. 1999. Detecting protein function and protein-protein interactions from genome sequences. *Science* 285:751–753.

Marsh JA, Teichmann SA. 2010. How do proteins gain new domains? *Genome Biol.* 11:126.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* 17:10–12.

Masel J. 2006. Cryptic genetic variation is enriched for potential adaptations. *Genetics* 172:1985–1991.

McLysaght A, Hurst LD. 2016. Open questions in the study of de novo genes: what, how and why. *Nat. Rev. Genet.* 17:567–578.

McVeigh A, Fasano A, Scott DA, Jelacic S, Moseley SL, Robertson DC, Savarino SJ. 2000. IS1414, an Escherichia coli insertion sequence with a heat-stable enterotoxin gene embedded in a transposase-like gene. *Infect. Immun.* 68:5710–5715.

Méheust R, Bhattacharya D, Pathmanathan JS, McInerney JO, Lopez P, Bapteste E. 2018. Formation of chimeric genes with essential functions at the origin of eukaryotes. *BMC Biol.* 16:30.

Méheust R, Watson AK, Lapointe F-J, Papke RT, Lopez P, Bapteste E. 2018. Hundreds of novel composite genes and chimeric genes with bacterial origins contributed to haloarchaeal evolution. *Genome Biol.* 19:75.

Méheust R, Zelzion E, Bhattacharya D, Lopez P, Bapteste E. 2016. Protein networks identify novel symbiogenetic genes resulting from plastid endosymbiosis. *Proc. Natl. Acad. Sci. U. S. A.* 113:3579–3584.

Mészáros B, Erdős G, Dosztányi Z. 2018. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res.* 46:W329–W337.

Monsellier E, Chiti F. 2007. Prevention of amyloid-like aggregation as a driving force of protein evolution. *EMBO Rep.* 8:737–742.

Moyers BA, Zhang J. 2016. Evaluating Phylostratigraphic Evidence for Widespread De Novo Gene Birth in Genome Evolution. *Mol. Biol. Evol.* 33:1245–1256.

Moyers BA, Zhang J. 2017. Further Simulations and Analyses Demonstrate Open Problems of Phylostratigraphy. *Genome Biol. Evol.* 9:1519–1527.

Neuhaus K, Landstorfer R, Simon S, Schober S, Wright PR, Smith C, Backofen R, Wecko R, Keim DA, Scherer S. 2017. Differentiation of ncRNAs from small mRNAs in Escherichia coli O157:H7 EDL933 (EHEC) by combined RNAseq and RIBOseq - ryhB encodes the regulatory RNA RyhB and a peptide, RyhP. *BMC Genomics* 18:216.

Neuhaus K, Oelke D, Fürst D, Scherer S, Keim DA. 2010. Towards Automatic Detecting of Overlapping Genes - Clustered BLAST Analysis of Viral Genomes. In: Pizzuti C, Ritchie MD, Giacobini M, editors. Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics. Lecture Notes in Computer Science. Springer Berlin Heidelberg. p. 228–239.

Ohno S. 1984. Birth of a unique enzyme from an alternative reading frame of the preexisted, internally repetitious coding sequence. *Proc. Natl. Acad. Sci. U. S. A.* 81:2421–2425.

Palmieri N, Kosiol C, Schlötterer C. 2014. The life cycle of Drosophila orphan genes. *Elife* 3:e01311.

Pasek S, Risler J-L, Brézellec P. 2006. Gene fusion/fission is a major contributor to evolution of multi-domain bacterial proteins. *Bioinformatics* 22:1418–1423.

Pathmanathan JS, Lopez P, Lapointe F-J, Bapteste E. 2018. CompositeSearch: A Generalized Network Approach for Composite Gene Families Detection. *Mol. Biol. Evol.* 35:252–255.

Pavesi A. 2006. Origin and evolution of overlapping genes in the family Microviridae. *J. Gen. Virol.* 87:1013–1017.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26:841–842.

Rancurel C, Khosravi M, Dunker AK, Romero PR, Karlin D. 2009. Overlapping genes produce proteins with unusual sequence properties and offer insight into de novo protein creation. *J. Virol.* 83:10719–10736.

Rogozin IB, Spiridonov AN, Sorokin AV, Wolf YI, Jordan IK, Tatusov RL, Koonin EV. 2002. Purifying and directional selection in overlapping prokaryotic genes. *Trends Genet.* 18:228–232.

Sabath N, Wagner A, Karlin D. 2012. Evolution of Viral Proteins Originated De Novo by Overprinting. *Mol. Biol. Evol.* 29:3767–3780.

Sastry AV, Gao Y, Szubin R, Hefner Y, Xu S, Kim D, Choudhary KS, Yang L, King ZA, Palsson BO. 2019. The Escherichia coli transcriptome mostly consists of independently regulated modules. *Nat. Commun.* 10:5536.

Schmitz J, Brosius J. 2011. Exonization of transposed elements: A challenge and opportunity for evolution. *Biochimie* 93:1928–1934.

Smith C, Canestrari JG, Wang J, Derbyshire KM, Gray TA, Wade JT. 2019. Pervasive Translation in Mycobacterium tuberculosis. *bioRxiv* [Internet]:665208. Available from: https://www.biorxiv.org/content/10.1101/665208v1

Snel B, Bork P, Huynen M. 2000. Genome evolution. Gene fusion versus gene fission. *Trends Genet.* 16:9–11.

Stewart NB, Rogers RL. 2019. Chromosomal rearrangements as a source of new gene formation in Drosophila yakuba. *PLoS Genet.* 15:e1008314.

Tautz D, Domazet-Lošo T. 2011. The evolutionary origin of orphan genes. *Nat. Rev. Genet.* 12:692–702.

Tunca S, Barreiro C, Coque J-JR, Martín JF. 2009. Two overlapping antiparallel genes encoding the iron regulator DmdR1 and the Adm proteins control sidephore and antibiotic biosynthesis in Streptomyces coelicolor A3(2). *FEBS J.* 276:4814–4827.

Vakirlis N, McLysaght A. 2019. Computational Prediction of De Novo Emerged Protein-Coding Genes. *Methods Mol. Biol.* 1851:63–81.

Vanderhaeghen S, Zehentner B, Scherer S, Neuhaus K, Ardern Z. 2018. The novel EHEC gene asa overlaps the TEGT transporter gene in antisense and is regulated by NaCl and growth phase. *Sci. Rep.* 8:17875.

Wang J, Rennie W, Liu C, Carmack CS, Prévost K, Caron M-P, Massé E, Ding Y, Wade JT. 2015. Identification of bacterial sRNA regulatory targets using ribosome profiling. *Nucleic Acids Res.* 43:10308–10320.

Watson AK, Lannes R, Pathmanathan JS, Méheust R, Karkar S, Colson P, Corel E, Lopez P, Bapteste E. 2019. The Methodology Behind Network Thinking: Graphs to Analyze Microbial Complexity and Evolution. *Methods Mol. Biol.* 1910:271–308.

Weaver J, Mohammad F, Buskirk AR, Storz G. 2019. Identifying Small Proteins by Ribosome Profiling with Stalled Initiation Complexes. *MBio* [Internet] 10. Available from: http://dx.doi.org/10.1128/mBio.02819-18

Willis S, Masel J. 2018. Gene Birth Contributes to Structural Disorder Encoded by

Overlapping Genes. *Genetics* 210:303–313.

Wilson BA, Foy SG, Neme R, Masel J. 2017. Young Genes are Highly Disordered as Predicted by the Preadaptation Hypothesis of De Novo Gene Birth. *Nat Ecol Evol* 1:0146–0146.

Wolf YI, Kondrashov AS, Koonin EV. 2000. Interkingdom gene fusions. *Genome Biol.* 1:research0013.1–13.13.

Xu J, Zhang J. 2016. Are Human Translated Pseudogenes Functional? *Mol. Biol. Evol.* 33:755–760.

Yamazaki Y, Niki H, Kato J-I. 2008. Profiling of Escherichia coli Chromosome Database. In: Osterman AL, Gerdes SY, editors. Microbial Gene Essentiality: Protocols and Bioinformatics. Methods in Molecular Biology™. Totowa, NJ: Humana Press. p. 385–389.

Yanai I, Derti A, DeLisi C. 2001. Genes linked by fusion events are generally of the same functional category: A systematic analysis of 30 microbial genomes. *Proceedings of the National Academy of Sciences* [Internet] 98. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC35447

Zehentner B, Ardern Z, Kreitmeier M, Scherer S, Neuhaus K. 2020. A Novel pH-Regulated, Unusual 603 bp Overlapping Protein Coding Gene pop Is Encoded Antisense to ompA in Escherichia coli O157:H7 (EHEC). *Front. Microbiol.* 11:377.

Zhao L, Saelao P, Jones CD, Begun DJ. 2014. Origin and spread of de novo genes in Drosophila melanogaster populations. *Science* 343:769–772.

Zhou Q, Zhang G, Zhang Y, Xu S, Zhao R, Zhan Z, Li X, Ding Y, Yang S, Wang W. 2008. On the origin of new genes in Drosophila. *Genome Res.* 18:1446–1455.