Overlapping genes in natural and engineered genomes

Bradley W. Wright[1,2], Mark P. Molloy[3] and Paul R. Jaschke[1]✉

Abstract | Modern genome-scale methods that identify new genes, such as proteogenomics and ribosome profiling, have revealed, to the surprise of many, that overlap in genes, open reading frames and even coding sequences is widespread and functionally integrated into prokaryotic, eukaryotic and viral genomes. In parallel, the constraints that overlapping regions place on genome sequences and their evolution can be harnessed in bioengineering to build more robust synthetic strains and constructs. With a focus on overlapping protein-coding and RNA-coding genes, this Review examines their discovery, topology and biogenesis in the context of their genome biology. We highlight exciting new uses for sequence overlap to control translation, compress synthetic genetic constructs, and protect against mutation.

Coding sequences
(CDSs). A continuous stretch of nucleotides that are bounded by a start and stop codon and undergo translation.

Overlapping genes
In eukaryotes, a gene overlap is when at least one nucleotide on either the same or opposite strand is shared between the primary transcripts of two or more genes. In prokaryotes and viruses, it is when at least two different coding sequences share a nucleotide either on the same or opposite strands.

When the first DNA genome was sequenced by Frederick Sanger in 1977, the results solved a perplexing mystery that had bothered scientists for some time. Previous analysis of the proteins produced by bacteriophage φX174 during infection seemed to require coding sequences (CDSs) longer than the measured length of the phage genome[1]. The mystery was solved when analysis of the genome sequence revealed extensive overlap between coding regions, with the internal scaffolding gene overlapping the genome replication gene and the lysis gene embedded entirely within the external scaffolding gene[1,2]. The compressed nature of these viral genes led to the conclusion that hidden within the genome could be other undiscovered sites of polypeptide synthesis[2]. Further refinement of the φX174 gene model showed an alternative start site within the genome replication gene A that produced a truncated protein with an identical CDS to the C-terminus of the A protein but holding a distinct function[3,4]. Thus, overlapping genes have been observed from the very beginning of sequencing and genomics. Since then, overlapping genes, and more specifically open reading frames (ORFs) and CDSs, have become a common genetic feature described during viral genome annotation[5], including within the SARS-CoV-2 genome[6]. However, until recently, their true abundance and importance was overlooked outside of the realm of viral genomics[7] and their discovery and annotation within cellular genomes have generally been treated as unique and idiosyncratic.

Today, we are seeing a renaissance of the field owing to the rapid advancement of genome-scale protein and RNA measurement tools and increasingly advanced prediction algorithms (BOX 1), which have collectively revealed an abundance of overlapping genes and ORFs within cellular genomes. Recent work on the human genome has placed estimates of overlapping features much higher than previously thought[8,9], encompassing 26% of all protein-coding genes[10]. This estimate will likely increase in the future as small ORFs (sORFs) encoding microproteins are increasingly being found in the human genome within previously annotated genes[11–13].

In this Review, we define a gene overlap in eukaryotes when at least one nucleotide is shared between the outermost boundaries of the primary transcripts of two or more genes, such that a DNA base mutation at the point of overlap would affect transcripts of all genes involved in the overlap (FIG. 1a, top). Thus, overlapping genes as defined here include 5′ and 3′ untranslated regions (UTRs) as well as introns. Overlapping ORFs and CDSs, which are components of genes, are distinctly defined here as when the overlap occurs in a sequence region of two or more genes that encode protein in the mature transcript such that a DNA base mutation at the point of overlap would alter a codon and potentially the protein sequence of one or more members of the overlap. We define a gene overlap in prokaryotes and viruses as when the CDSs of two genes share a nucleotide either on the same or opposite strands (FIG. 1a, bottom). These definitions are compatible with a recently updated, community-driven effort to create consensus classifications of non-canonical ORFs, of which overlaps are one example[14].

Here, we review overlapping genes as fundamental features of both cellular and viral genomes. We first discuss the diverse topologies and functions of overlapping genes in natural genomes across prokaryotes, eukaryotes and viruses. We then highlight their importance for synthetic biology approaches, as bioengineers are both faced with disentangling CDSs to refactor gene clusters and whole genomes and inspired to implement these

[1]*Department of Molecular Sciences, Macquarie University, Sydney, NSW, Australia.*

[2]*Cecil H. and Ida Green Center for Reproductive Biology Sciences, University of Texas Southwestern Medical Center, Dallas, TX, USA.*

[3]*Bowel Cancer and Biomarker Laboratory, School of Medical Sciences, The University of Sydney, Sydney, NSW, Australia.*

✉*e-mail: paul.jaschke@mq.edu.au*

https://doi.org/10.1038/s41576-021-00417-w

Box 1 | **Identifying overlapping genes and ORFs**

Genome annotation is the bedrock against which genome-scale measurements are compared, with most bioinformatics pipelines today annotating genomes through a combination of sequence alignments and hidden Markov modelling. However, many of these standardized methods may be inappropriate for the discovery of overlapping genes because they are reliant on already curated genes, where overlapping genes are poorly represented and contain atypical sequence composition[40,41,176]. For example, the RAST[177] pipeline uses both ab initio (GLIMMER) and sequence homology steps (SEED genome database) to annotate genomes[178] but markedly penalizes overlaps between predicted open reading frames (ORFs), which potentially misses vital features[177]. Furthermore, genome annotation standards are biased against feature overlaps, especially genes “completely contained in another gene”[179]. The solution may be custom algorithms tailored for overlap mapping that have been created specifically for viral genome annotations (for example, OLGenie[180]) and annotation pipelines based on hidden Markov models trained on databases of experimentally confirmed overlapping genes[181]. Some tools, such as Glimmer3 and BG7, are more tolerant of overlapping ORFs by retaining candidate ORFs even if they overlap other predicted ORFs[182,183]. New annotation databases, such as OpenProt[184], are being created in response to the growing realization that eukaryotic gene models need to include polycistronic transcripts with non-AUG initiation sites[185].

Proteogenomic methods, including bottom-up proteomics and ribosome profiling, in combination with DNA sequencing and perturbation, have been critical for the identification of overlapping genes. Mass spectrometry-based proteomic techniques are used mainly to confirm the expression of gene products based on genomic sequence annotation and are notionally limited by the quality of annotations. Most commonly, proteomics is performed using shotgun tandem mass spectrometry, whereby proteolytic peptide digests are ionized and sequenced based on peptide fragment ion mass-to-charge ratios, thus providing primary evidence of translated gene products. However, for large-scale studies, MS data must be computationally matched to in silico digests of the theoretical proteome. Unbiased six-frame genome translations can be used to maximize the proteome ‘search space’ but are rarely implemented due to expanded computational analysis time and high false-discovery rates[186]. In addition, recent studies have shown unexpectedly strong non-AUG translation initiation[187,188], which are not accounted for in standard six-frame AUG translations. N-terminal peptide enrichment strategies can be used to identify sites of translation initiation, regardless of start codon used[189,190], but the database needs to already include these candidates. Despite these considerations, proteomic measurements can be powerful, with one study identifying 1,259 alternative proteins produced from previously annotated human transcripts[191].

Complementary to mass spectrometry proteomics, ribosome profiling (Ribo-Seq) is a method that involves capturing ribosomes as they decode mRNA and sequencing the section of the transcript bound by the ribosome[192]. In particular, the translation initiation site Ribo-Seq variant, which uses inhibitors to pause ribosomes on the start codon, has revealed an abundance of new translation initiation sites within transcripts in prokaryotic[29], eukaryotic[11] and viral genomes[193–195].

RNA sequencing alone can also identify genomic regions with overlapping transcripts. For example, 180,000 alternate ORFs within previously annotated coding regions were found in humans[66], and a transcription start site profiling study in *Helicobacter pylori* identified pervasive transcription on the opposite strand of canonical genes (that is, antisense transcription)[196].

Overlapping ORFs discovered using the above methods have been verified using a variety of reverse genetics approaches, including CRISPR–Cas9 and catalytically dead Cas9 (dCas9) disruption[11,12,65], as well as an attempt at proof-by-synthesis to establish the absence of any undiscovered overlapping genes[197].

features in synthetic genetic constructs to control protein expression and slow evolution. We limit our discussion to protein-coding and RNA-coding regions within genomes that partially or completely overlap at least one other gene. For information on ORFs localized entirely within 5′ or 3′ UTRs, which itself is a rapidly evolving field, we direct readers to other works[15,16].

Overlapping gene topology and function

Studying overlapping genes across cellular and viral genomes reveals different patterns of overlap topologies that vary in frequency between prokaryotes and eukaryotes[8,17]. The reasons for these observed patterns are either more frequent biogenesis of certain types, evolutionary selection for retention of certain topologies or a combination of the two. At the moment, no consensus exists for the relative importance of these two factors, that is, creation versus retention. Overlap is thought to arise from at least six mechanisms that result in one gene becoming entangled with another, either through sequence extension[9,18,19], re-arrangement of existing genes[20,21], or de novo gene and ORF creation within an existing gene[22].

Three directional overlap topologies are possible (FIG. 1b). Unidirectional overlaps (→→) occur between genes encoded on the same strand and may be further categorized according to the reading frame for overlapping ORFs. The remaining two topologies occur between genes on opposite strands and are called convergent (→←) and divergent (←→) (FIG. 1b). Unidirectional overlaps are more frequent in genomes of viruses and bacteria[5,17], whereas the divergent and convergent overlaps are more frequent in eukaryote genomes[10,23]. The way the two genes interact can be described as either overlapped, with only part of each gene sequence occupying the same genomic region, or nested (FIG. 1c), whereby the entire extent of one gene is enclosed within the borders of a larger gene. The relationship between overlapping and nested genes has been described in other ways, including ‘internal–external’[20] or ‘mother–daughter’ genes[24].

The different ways that genes are defined in prokaryotes and eukaryotes in the literature has possibly biased estimates of the prevalent types of overlaps between these groups. For example, in prokaryotic and virus literature, gene overlaps are only considered when the CDSs of the genes overlap[5,17], whereas in eukaryotic literature overlaps are more often considered between the primary transcript boundaries[10,25] (FIG. 1a). The effect of these different definitions is that certain types of overlap seem to be more prevalent in eukaryotes versus prokaryotes but, if the same definitions were used for both, these apparent differences could in fact disappear. For instance, overlapping CDSs have certain constraints on relative reading frame and sequence composition[26,27] that overlaps between 5′ and 3′ UTR do not. Within the limitations posed by the way overlapping genes are described in the literature, we compare and discuss prokaryotic and eukaryotic gene overlap from both their idiosyncratic aspects as well as their similarities, where present.

Prokaryotes. Overlapping CDSs within prokaryotic genomes have been reported in both bacteria[28–30] and archaea[31] and, on average, 27% of CDSs in these groups are involved in at least one instance of overlap[19]. Across prokaryotes, the frequency of CDS overlap within a genome seems to be constant regardless of genome size[17,32], although certain groups can deviate sharply from this pattern. For example, intracellular microbial parasites show a weak correlation between genome size and the number of overlapping CDSs[33].

In prokaryotic genomes, 84% of CDS overlaps are unidirectional[17] (→→) and produced through start codon or stop codon loss, resulting in one member of a

Open reading frames
(ORFs). A continuous stretch of nucleotides, on genome or transcript, that are bounded by a start and stop codon.

Small ORFs
(sORFs). ORFs that are equal to or less than 300 nt in length.

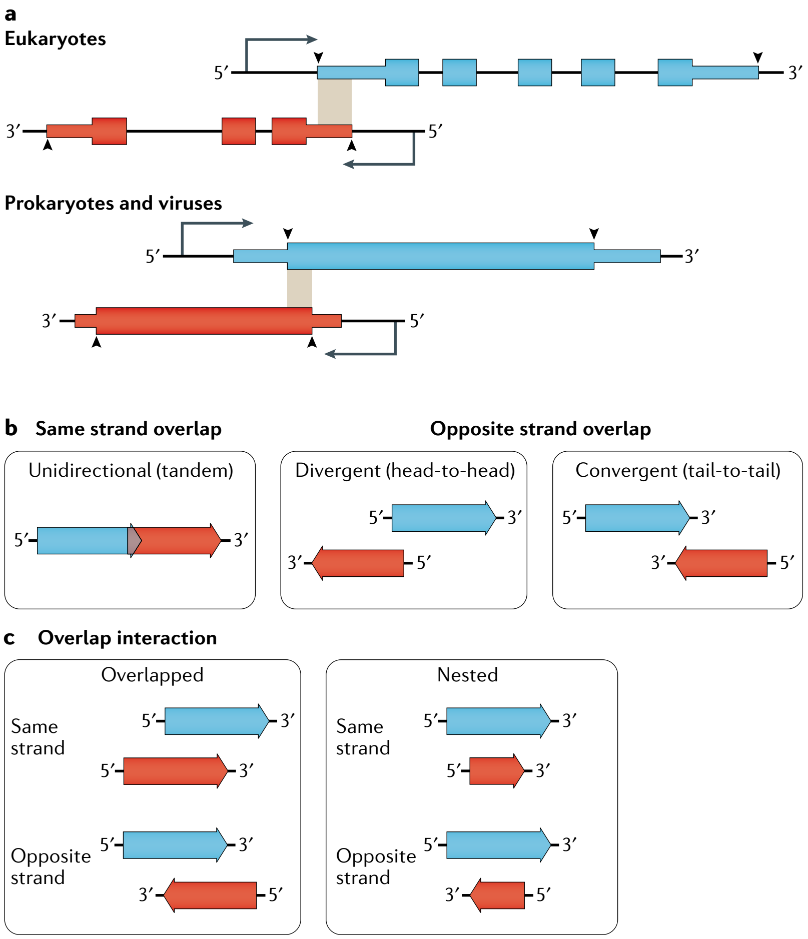

Fig. 1 | **Overlapping gene definition and topologies. a** | Gene overlap definitions differ between prokaryotes and eukaryotes. (Top) Eukaryote overlaps are most frequently defined as overlaps between the boundaries of the primary transcript, shown here in the shaded region. Often, the overlap is only between the 5′ untranslated region (UTR) or 3′ UTR of both transcripts (5′ UTR overlap shown)[10,170]. (Bottom) In contrast, prokaryote and virus genes are only considered to overlap if their coding sequences overlap[5,27]. Thin boxes denote 5′ and 3′ UTRs while thick boxes are coding sequences. Arrowheads indicate the extent of the consensus definition of gene boundaries within studies referenced in this review. **b** | Genes and open reading frames (ORFs) can be overlapped in one of three topologies. Unidirectional (also called tandem) overlaps occur between genes and ORFs on the same strand. Divergent (also called head-to-head) overlaps occur between genes and ORFs on opposite strands that overlap at their 5′-ends. Convergent (also called tail-to-tail) overlaps occur between genes and ORFs on opposite strands that overlap at the 3′-ends[27]. **c** | Gene and ORF interactions can be either overlapped, where only limited portions of each gene or ORF are overlapping, or nested, where the entire sequence of one partner falls within the boundaries of the other.

Primary transcripts
A transcribed RNA molecule, containing both exons and introns, prior to undergoing post-transcriptional processing to yield a final, mature transcript.

Overlapping ORFs
When at least one nucleotide on either the same or opposite strand is shared between two sequences that consist of a length divisible by three and begin with a translation start codon and end at a stop codon.

pair of adjacent non-overlapped CDSs expanding their coding sequence into their adjacent partner (FIG. 2a,b). Sequence analysis shows that stop codon loss of the upstream partner is the most frequent mechanism for unidirectional overlap creation[32,34]. Start codon loss of the downstream partner and de novo start codon creation within an existing CDS (FIG. 2c) also generate unidirectional overlaps[18,32]. Over 98% of currently identified unidirectional overlaps are less than 60-bp long, with the vast majority of these short overlaps either 1 bp or 4 bp overlapping start and stop codons (TA[A]TG, TG[A]TG, or [ATGA])[17,35]. This overlap motif may be intimately tied to prokaryotic operons, where clusters of related genes are under the regulatory control of a single promoter, and overlapping start and stop codons of their respective CDSs may facilitate enhanced regulatory control through translational coupling between adjacent partners[36].

Convergent (→←) and divergent (←→) overlaps (FIG. 1b) are observed at lower frequencies in prokaryotes compared with eukaryotes, and similar to unidirectional overlaps, are biased towards short overlap lengths[35]. Short convergent overlaps are strongly biased towards 4-bp stop codon overlaps owing to the incompatibility of forward-strand stop codons (TAA, TAG, TGA) with reverse-strand stop codons (TTA, CTA, TCA) in any other configuration[37]. Divergent overlaps (FIG. 1b) do not have strong phase biases but are substantially rarer than convergent overlaps[38], which is likely due to the presence of critical sequence structures in the 5′-end of CDSs that impose additional evolutionary constraints on the successful retention of these overlap topologies.

It is currently unclear whether the commonness of short tandem start–stop overlaps compared to long nested overlaps (FIG. 1b) is a result of biology or merely reflects our ease to detect them. Despite increasing numbers of fully nested CDSs within prokaryotes being discovered due to a convergence of proteomic and ribosome profiling methods (BOX 1), the idea that many more long nested overlaps within prokaryotes remain to be discovered is contentious[19,35] and genome annotation pipelines are biased against their existance[39]. The unusual sequence characteristics of long overlapping CDSs may have also contributed to the difficulty of their discovery, resulting in undercounting[40,41]. One reason put forward to explain why long nested overlaps should be rare includes the evolutionary burden of maintaining larger overlaps, although evidence to the contrary showing positive selection at overlaps[27,42,43] shows that this explanation may be too simplistic. Selection for long convergent overlaps has been shown to have a strong reading frame bias and it has been suggested that retention involves positive selection at the birth of the overlap, followed by purifying selection afterwards[27]. Recently, an overlapping protein-encoding CDS with extensive 603 bp overlap has been discovered embedded in the highly conserved *ompA* gene in enterohaemorrhagic *Escherichia coli*[44], showing that, with improved measurement tools, more of these long nested overlaps may be discovered[42].

While the precise selective forces governing the retention of long unidirectional CDS overlaps in prokaryotes are unknown, the selective forces governing the retention of some short stop–start overlaps likely act through their enhancing effect on gene expression[36] (FIG. 3a,b). Furthermore, overlapping CDS frequency is higher in fast-growing thermophilic organisms, which suggests that genome streamlining is an adaptive strategy for fast growth at high temperatures[45,46]. Mechanistically, overlaps between start and stop codons of adjacent unidirectional CDSs provide additional benefits for translational coupling[47–50] and ribosome re-initiation[48,50] (FIG. 3a,b) in addition to benefits already provided by operons[51,52]. The menaquinone biosynthesis pathway in *E. coli* is an example of multiple gene members connected via overlapped

Operons
Regions in the genome of prokaryotes that encode multiple adjacent CDSs under the control of a single promoter.

stop–start sites within a single operon across all three reading frames (FIG. 4a).

Functional entanglement of overlapped CDSs can act on their retention over evolutionary selection beyond gene expression levels. For example, the overlapping *drrA*/*drrB* genes encode an efflux pump for the anticancer agent doxorubicin in the production strain *Streptomyces peucetius*. When the overlap was disrupted, the expression levels of DrrA and DrrB proteins remained unchanged and membrane trafficking was unaffected but functional assembly of the protein complex was lost[47]. Correct protein complex assemply has been revealed to be spatially regulated at the translation level for genes linked in operons, which may explain the DrrA/DrrB finding[47,53]. Overlap functions such as this are likely to be prevalent for overlapped CDSs given the functional assortment of genes involved in overlap[54].

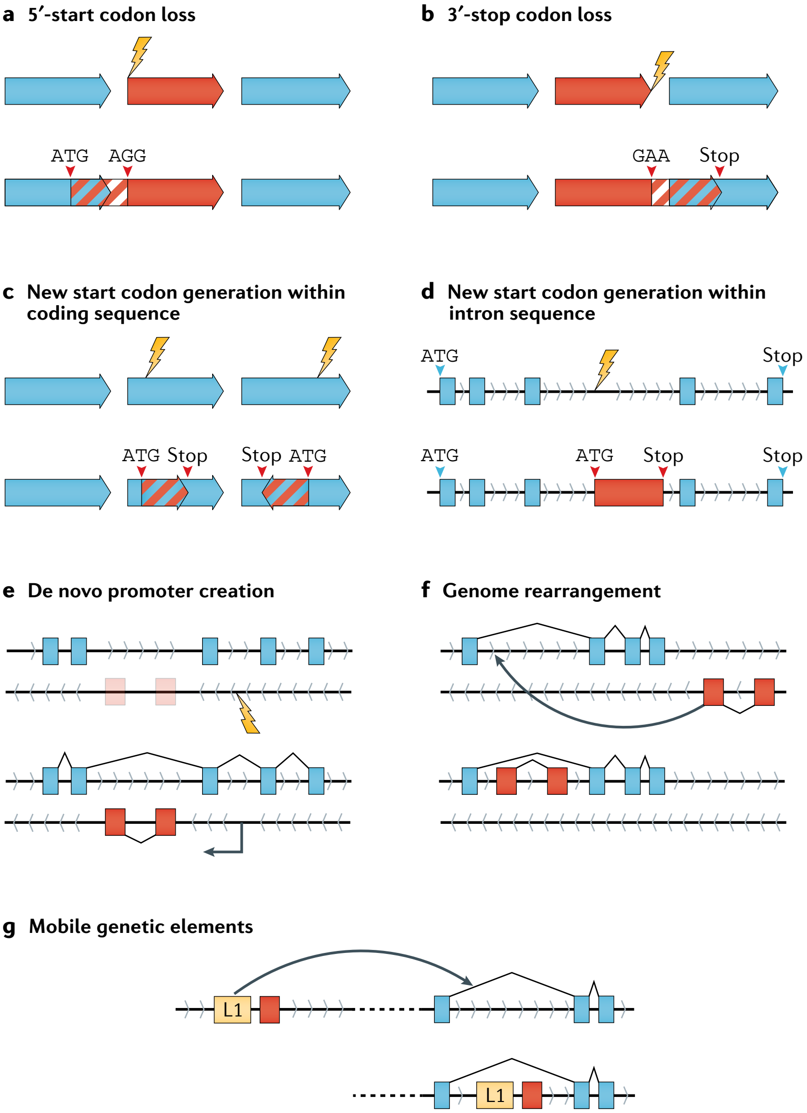
a 5′-start codon loss
ATG
AGG
b 3′-stop codon loss
GAA
Stop
c New start codon generation within coding sequence
ATG Stop
Stop ATG
d New start codon generation within intron sequence
ATG
Stop
ATG
ATG
Stop
Stop
e De novo promoter creation
f Genome rearrangement
g Mobile genetic elements
L1
L1

Fig. 2 | **Mechanisms of gene and ORF overlap creation.** New overlaps can be created through a range of mechanisms and likely require numerous complementary developments to produce the appropriate sequence context for retention of gene or open reading frame (ORF) functionality. **a** | Mutations removing the start codon of a downstream ORF may result in the next available upstream start codon being utilized, which could be within an upstream ORF[18]. **b** | Mutational loss of a stop codon may result in the extension of an ORF. Similar to start codon loss, the next available stop codon may be utilized, which could be within a downstream ORF[19]. **c** | De novo generation of an ORF may begin with the creation of a start codon within an existing coding region through mutation and, in conjunction with a downstream stop codon, produces an overlapping ORF[18]. **d** | Non-coding intron sequences may acquire a start codon through mutation and, in conjunction with a downstream stop codon, produce a nested ORF[20]. **e** | Mutations that result in the de novo development of a sequence capable of recruiting transcriptional machinery (such as a promoter or enhancer) may result in a new overlapping gene[171]. **f** | Genome rearrangements, such as inversions and translocations, may result in distant non-overlapping genes becoming overlapped. This mechanism has been seen within human cancers. **g** | Mobile genetic elements carrying genes (such as transposons or proviral genes) may localize to within a gene, generating a new gene overlap[172,173].

Eukaryotes. In eukaryotic genomes, the prevalence of overlapping genes is difficult to assess because of the inconsistent nomenclature that is used to describe the relationship between the genes, their 5′ or 3′ UTRs, and CDSs. Unlike prokaryotes, classifications and studies of overlapping genes in eukaryotes are as varied as their genome size and complexity. The predominant type of overlap is convergent[8,10,23] (FIG. 1b), although generalization within eukaryotes is less useful given their genome diversity, which ranges from unicellular eukaryotes with compact, intron-poor genomes to complex, multicellular eukaryotes with expanded genomes and high intron densities[55,56].

Most overlapping genes in eukaryotes are classified as such because their 5′ or 3′ UTRs overlap[57]. Of those with overlaps between the start and stop codon boundaries of either member (FIG. 1a), introns provide an additional non-coding location for gene transcripts to overlap. When an entire ORF is contained within an overlapping gene's intron it is referred to as intron nesting[20,58]. True exon–exon overlaps make up the minority of transcript overlap in eukaryotes[8,23] but new technologies (BOX 1) suggest that they may be more common than currently appreciated[11,12].

Nested gene overlaps in eukaryotes occur most frequently within an intron of the larger partner as is the case for three antisense nested genes, *EVI2A*, *EVI2B* and *OMG*, within intron 27b of the human *NF1* gene (FIG. 4b). Nested overlaps are thought to be created through four processes: (1) mobilization of a distal gene into the intron of another gene (for example, through retrotransposition), (2) de novo creation of an ORF within an intron of an existing gene, (3) one ORF is internalized after an adjacent gene acquires additional exons and (4) two external genes flanking another gene fuse, thus internalizing the other gene[20] (FIG. 2). The introns that harbour nested genes are considerably longer than other introns, suggesting acquisition of an existing gene through retrotransposition, among other mechanisms, is a dominant process rather than de novo evolution[21,59]. However, evidence from metazoans shows that several de novo genes have emerged from introns in that lineage[20,60]. The extent of the nesting can vary from an internal gene with a single exon residing within the intron of an external gene (for example, *H2BFS* within *HSF2BP* in humans[21]) to multiple layered 'Russian doll-like' nestings in *Drosophila melanogaster*[20].

Eukaryotic overlapping protein-coding genes are implicated in lineage-specific groups. For example,

Translational coupling
The interdependence of translation efficiency of overlapping CDSs, in particular for those with overlapping start–stop codons.

Positive selection
The process whereby the frequency of an allele in a population increases as a result of an increase of fitness of their carrier.

Purifying selection
The process whereby harmful alleles are eliminated from a population. Also known as negative selection.

Retrotransposition
The movement of genetic information from one genomic location to another through an RNA intermediary.

the majority of vertebrate genes with overlapping transcripts are not conserved across species[9,57] likely because overlapping genes tend to be young and frequently lost during evolutionary time[57]. A broad study of five well-described metazoan genomes (*Caenorhabditis elegans*, *Danio rerio*, *Drosophila melanogaster*, *Mus musculus* and human) found that, for protein-coding genes, transcript overlap is selected against and mainly species specific and the majority of new overlaps are in terminal non-coding exons[25]. Overlap between opposite strand exons containing coding sequence is also lineage specific, with the mammalian genes *THRA* (which encodes thyroid hormone receptor alpha) and *NR1D1* (which encodes nuclear receptor subfamily 1 group D member 1) displaying convergent overlap in the coding sequence portion of their 3′ exons, whereas marsupials seem to have lost this feature since their divergent evolution over 90 million years ago. This change results in an absence, during marsupial development, of the TRα2 protein, a variant of the receptor unable to bind the hormone[61].

Although rare, eukaryotes contain genes with CDS overlaps[8,9,62,63] as well as overlaps that span exon–intron boundaries[57,64]. A community-driven roadmap on translated ORFs has proposed that these overlapping

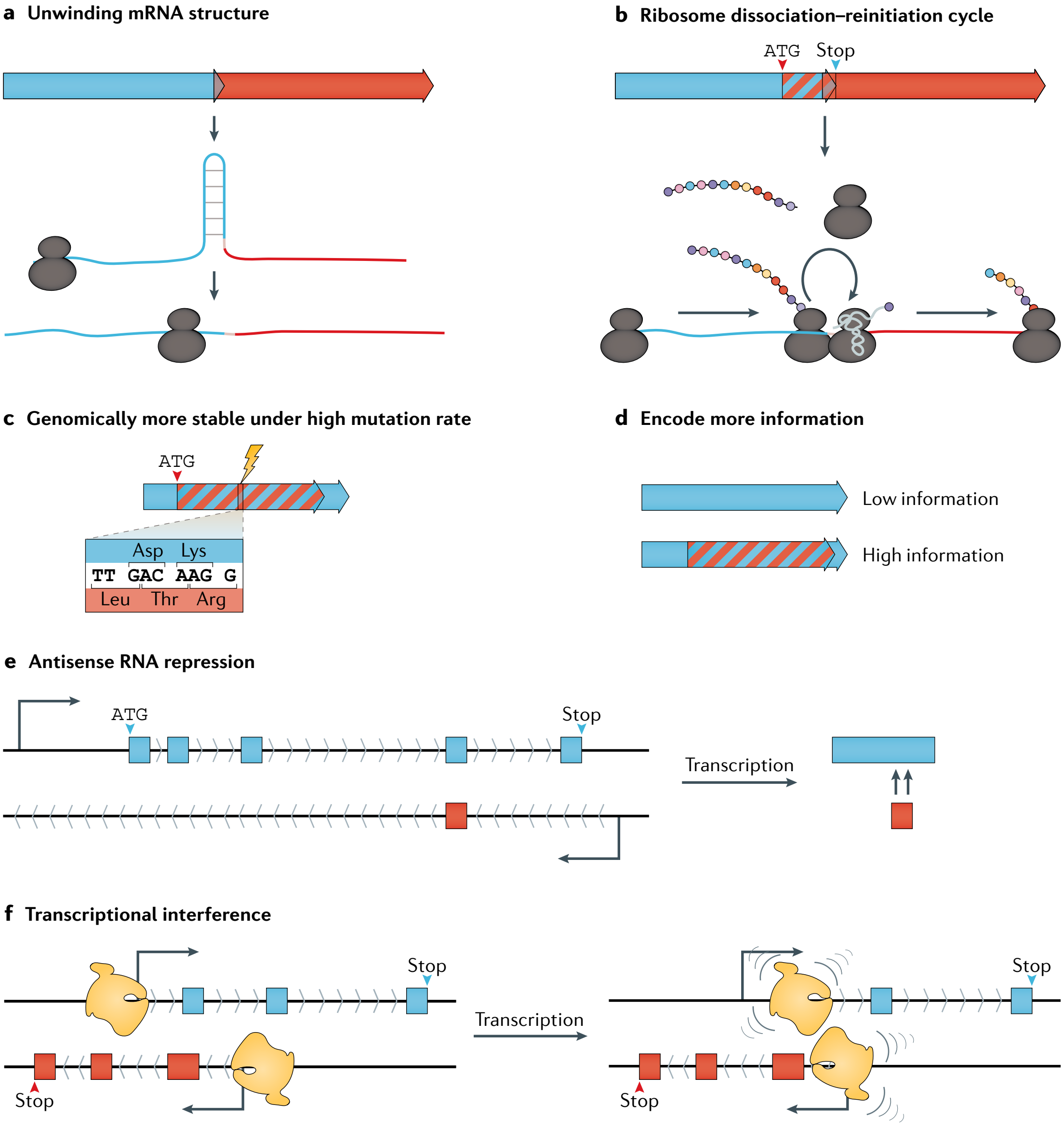
a Unwinding mRNA structure
b Ribosome dissociation–reinitiation cycle
ATG
Stop
c Genomically more stable under high mutation rate
ATG
Asp Lys
TT GAC AAG G
Leu Thr Arg
d Encode more information
Low information
High information
e Antisense RNA repression
ATG
Stop
Transcription
f Transcriptional interference
Stop
Transcription
Stop
Stop
Stop

Fig. 3 | **Selective pressures involved in retaining gene and ORF overlaps. a,b** | Overlapping start and stop codons cause translation coupling between unidirectional overlapping open reading frames (ORFs) through unwinding of mRNA secondary structure around the ribosome binding site and start codon and by enhancing ribosome re-initiation[48]. **c** | Overlapping sequence regions cause mutations to affect more than one ORF, increasing fitness cost and preserving overlapped sequences under mutational pressure[71,75]. **d** | Encoding more ORFs in the same sequence region allows genetic novelty with reduced genome changes, which is particularly advantageous for viruses that have spatial constraints on genome size[76,77]. **e** | Sense–antisense gene and ORF overlap is frequently involved with gene expression regulation, including non-coding RNA and long non-coding RNA[96]. **f** | Transcriptional tuning from convergent overlapping genes and ORFs as a result of interactions between RNA polymerase collisions (transcriptional interference[174,175]).

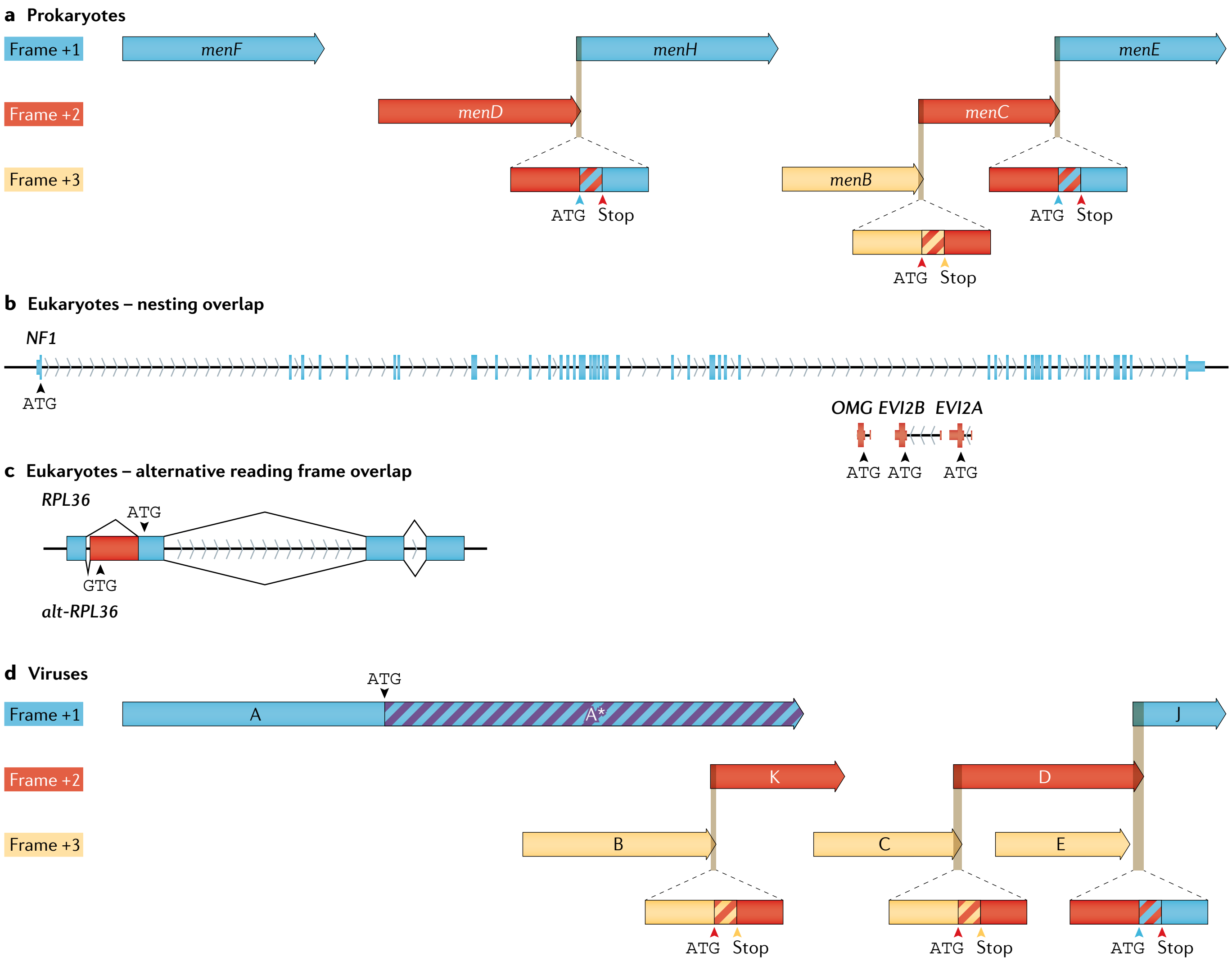

Fig. 4 | **Gene and ORF overlap across prokaryotes, eukaryotes and their viruses. a** | *Escherichia coli* menaquinone biosynthesis operon contains three short stop–start coding sequence (CDS) overlaps. **b** | The large human gene *NF1* and internal nested protein-coding ORFs *OMG*, *EVI2B* and *EVI2A* are located within *NF1* introns. **c** | Recently described *alt-RPL36* (bottom) overlaps the human ribosomal protein gene *RPL36* (REF.[65]) through an out-of-frame GTG start codon within a 5′-extended *RPL36* exon present on *RPL36* transcript variant 2. The *alt-RPL36* CDS generates a longer protein with an entirely different sequence from *RPL36* (REF.[65]). **d** | The virus φX174 contains overlaps in all three reading frames: three short unidirectional stop–start CDS overlaps, two nested CDSs, and one in-frame start generating an N-terminally trunkated protein.

CDSs be annotated as novel genes despite the shared locus[14]. The recently described *alt-RPL36* ORF[65] (FIG. 4c) is one such example of a gene possessing two distinct and functional CDSs overlapping the same genomic sequence. These alternative ORFs[66] are often functionally related and implicated in a range of human diseases[12,67]. For example, the cyclin-dependent kinase inhibitor 2A (p16INK4a) and tumour suppressor ARF, which regulate the tumour suppressors retinoblastoma protein (RB) and p53 transcription factor[68], are produced as alternatively spliced transcripts from what is now considered the same gene (*CDKN2A*), even though the proteins do not share sequence or structural similarity, and the *E1b* exon that produces the ARF protein is ~20 kb upstream of the other *CDKN2A* exons[68]. Similarly, a recently discovered nested overlapping ORF within the *FUS* ORF (alt-FUS) is associated with neurodegeneration[69] and alt-Ataxin is mutated in spinocerebellar ataxia type 1 (REF.[64]).

Viruses. The topology of overlapping genes in viruses is determined both by the host cell type as well as by constraints unique to viruses. Despite viruses having diverse genomes (RNA or DNA in single-stranded or double-stranded form) and lifestyles, overlapping CDSs are found across all known virus groups[5,70]. The proportion of viruses with overlapping CDSs within their genomes varies from double-stranded RNA viruses having fewer than a quarter to almost three-quarters of *retroviridae* (single-stranded RNA using reverse transcriptase) and single-stranded DNA genomes containing overlapping CDSs[5]. Segmented viruses, those with the genome split into separate pieces and packaged either all

Alternative ORFs
Also called non-canonical ORFs, are ORFs that occupy a shared sequence region with a canonical CDS, often in a different reading frame.

in the same capsid or in separate capsids, are more likely to contain an overlap than non-segmented viruses[5]. The retention of overlapping CDSs in viruses has been attributed to enabling evolutionary rate reduction and increasing mutational robustness[71,72] as well as being a result of capsid size limitations[73].

The role of overlapping genes in reducing the rate of viral evolution has been most intensively examined in RNA viruses, which have higher mutation rates, smaller genomes and less CDS overlap than DNA viruses of comparable length[5,73,74]. Studies have supported the notion that CDS overlap increases hypersensitivity to mutation (as a mutation on average would affect more than one CDS)[26] but that genome (or population) mutational robustness is increased overall[71] (FIG. 3c). This has been eloquently demonstrated with the overlapping *rev* and *tat* genes of the RNA virus HIV1 (REF.[75]). Functional segregation is observed between the overlapped regions, facilitating the purging of possible deleterious mutations; that is, important nucleotide or amino acid regions of one gene overlap regions subject to fewer constraints in the other[75].

Thus, given that gene overlap regions are likely protective and increase fitness, why then do RNA viruses have fewer overlapping genes than DNA viruses with lower mutation rates and less restrictive genome sizes?[5,73] The answer may lie in the balancing of different selection pressures. For instance, the lower mutation rate of DNA viruses facilitates greater genomic novelty and evolutionary exploration within a structurally constrained genome and may therefore be the primary driver of gene overlaps[76,77] (FIG. 3d). By contrast, in RNA viruses, overlaps may primarily be a means for maintaining mutational robustness in the face of higher mutational rates (FIG. 3c)[71,75] as exemplified with the population fitness advantage conferred by the *rev* and *tat* overlap of HIV1 (REF.[75]).

Virus capsid size restrictions driving the evolution of gene overlaps has been a focal point of investigation due to early observations of dramatic viability loss in viruses with genomes engineered to be longer than wild type[78]. For instance, increasing the single-stranded DNA genome length of ΦX174 by >1% results in almost complete loss of infectivity[79]. This is thought to be the result of the strict physical constraints imposed by the finite capsid volume and, as such, any evolutionary innovation must be facilitated in the existing sequence space (FIG. 3d) rather than by increasing genome length. This idea is supported by work with adeno-associated viruses as gene delivery vectors, where viral packaging is constrained by genetic cargo size limits[80], necessitating the use of multiple vectors to deliver large human genes such as *CFTR*[81]. Studies have shown a strong prevalence of overlapping CDS births in the +2 frame over the +3 frame[40,77], which is likely due to two factors: mutational bias, whereby start codons are more prevalent in the +2 reading frame relative to known CDSs[40,74], and recent evidence suggesting that the sequence of known CDSs in the +2/−2 reading frames preserves key physicochemical properties of the original sequence[82].

The seemingly simple relationship between genome and capsid has also been questioned. Combined structural and genomic data have shown that most viruses do not fully utilize the available internal space of the capsid[76]. Furthermore, viruses are highly biased towards short overlaps, with the vast majority less than 50 nt (REF.[5]) in length, overall negatively correlated with genome length[70], with absolute nucleotide overlap summed across the genome rarely exceeding 1,500 nt (REF.[76]). This distribution of overlap length within viruses points towards overlaps being favoured for several different reasons, with short CDS overlaps enabling translational coupling, whereas long overlaps being retained mainly when they generate genetic novelty that increases fitness. For example, a 4-nt (ATGA) stop–start overlap within a *Totivirus* directs coupled translation of the CDSs[83], whereas a 276-nt overlap in phage ΦX174 between its recently evolved lysis gene E and scaffolding gene D (FIG. 4d) enables the phage to lyse its host and release virions more efficiently[1,2].

Overlap of ncRNA with protein-coding genes

Another important and highly abundant type of overlap within genomes is between non-coding RNA (ncRNA) genes and those of protein-coding genes. Shared sequence overlap may be between the mature ncRNA transcript region and the CDS region of mature protein-coding mRNA or it may only occur between 5′ and/or 3′ UTR regions of the transcripts.

In prokaryotic genomes, ncRNAs are an increasingly identified feature[84], with *cis*-encoded antisense RNA regulation being a major player in physiological responses[84,85]. Examples of these pairings have demonstrated tight-knit regulation of expression of the protein-coding gene such as in type I toxin/antitoxin systems[86] and in Mg^{2+} tolerance and virulence[87]. Interestingly, examples of unusually long antisense RNA have also been found, which likely hold greater regulatory control functions (such as regulation of entire operons) and have acquired their own designation as 'excludons'[84,88]. Overlapping regulatory RNAs embedded within the coding sequence of bacterial genes can act in diverse regulatory roles[89–91]. Evidence is also emerging that ncRNAs in prokaryotes can contain protein-coding ORFs[92,93]. For more information on prokaryotic overlapping ncRNAs, we refer readers to another review[84].

In eukaryotes, the sense–antisense overlapping transcripts are called *cis*-natural antisense transcripts (cis-NATs) and this type of overlap topology is frequently found in eukaryotic genomes in convergent or divergent relationships (FIG. 1b). Cis-NATs have regulatory functions at the RNA level[25,94] and the most frequent combination is one protein-coding transcript paired with an antisense non-protein-coding transcript[95] enabling enhanced transcriptional and post-transcriptional gene regulation[96] (FIG. 3e,f). The regulatory roles of cis-NATs span major biological functions[97] but can be generalized into protein expression regulation[98], splice site masking[99,100], double-stranded RNA-dependent mechanisms[101,102] and chromatin remodelling[103,104]. Furthermore, due to the *cis*-acting mechanism and shared genetic loci, the evolutionary trajectories of both genes are closely entwined[105,106]. As such, interesting questions surround their evolution and acquisition,

Non-coding RNA
(ncRNA). A strand of RNA that has been transcribed from DNA but does not undergo translation. This RNA will typically have a regulatory function.

Cis-natural antisense transcripts
(cis-NATs). Transcribed products from the DNA strand complementary to a region harbouring a sense transcript of either protein-coding or non-coding genes.

Refactoring
The reorganization of biological systems or pathways with the goal of improving the ease and predictability of future engineering efforts. The process often involves removing CDS overlaps and changing regulatory sequences.

Biosynthetic gene clusters
A physically clustered group of two or more genes that encode a biosynthetic pathway for the production of a specialized metabolite.

such as whether one member of the pair arose de novo through the acquisition of a promoter or by other mechanisms (FIG. 2). Recently, some overlapping ncRNA antisense transcripts have been found to also encode proteins[11,107], further increasing the complexity and constraints of these overlapping interactions.

Many cis-NATs have been associated with human disease, including cancer progression[108–111]. For example, the convergently overlapping *WDR83* and *DHPS* genes both encode proteins; together, RNA duplexing of their 3′ UTRs results in the concordant increase in their transcript stability and protein expression, ultimately resulting in increased cell proliferation in gastric cancer cells[102]. In a subpopulation of patients with α-thalassaemia, the disorder is caused by a chromosomal deletion that creates a new gene overlap between *HBA2* and *LUC7L*, resulting in antisense transcripts from *LUC7L* silencing the otherwise intact copy of *HBA2* through CpG island methylation[112].

An emerging feature of many ncRNAs is the presence of internal translationally active sequences termed sORFs. These sORFs are commonly defined as an ORF that spans no more than 300 nt that, owing to these small lengths, have lain hidden within previously described ncRNA transcripts[113]. In humans, 30% of sORF-derived proteins (also called microproteins) identified by mass spectrometry were mapped internally to annotated genes[114]. Subsequent studies have expanded this number using a variety of methods[115,116], including recent work that systematically uncovered hundreds of sORFs. The sORFs were found overlapping both internal sequences as well as the start codons of annotated ORFs[11]. Investigations into the functionality of the overlapping sORFs have implicated many in human disease pathology[107,117]. Furthermore, it is likely that many of the ncRNAs found to possess sORFs are in fact misannotated and should be re-defined as mRNA; however, there are examples of RNA that possess dual functionality (non-coding and coding), thereby complicating classifications[118,119]. More information on this developing area can be found in recent reviews[120,121], including the in-motion and recent community-driven initiative to comprehensively define and catalogue these classes of non-canonical ORFs in major databases[14].

Overlapping genes in bioengineering

As we have outlined, gene overlaps in natural genomes are complex and their true number is only beginning to emerge. However, in synthetic biology, the re-engineering of natural genomes is well under way. Synthetic biology uses raw genetic material from diverse sources within heterologous systems to create new metabolic pathways[122], enzymatic activities[123], orthogonal transcription[124,125] and translation initiation systems[126,127], and complex genetic devices[128,129]. As such, the functional characteristics of overlapped genetic elements are becoming increasingly important to understand. Furthermore, the field of synthetic genomics is rapidly rebuilding entire genomes from the ground up (for example, *E. coli*[130] or the yeast *Saccharomyces cerevisiae*[131]), with important choices to be made during the design stage for how to deal with overlapping sequences[132].

Refactoring overlapping genes. Genome refactoring is a process of reorganizing gene architecture by reformatting the underlying sequences while maintaining functionality[133]. With the aim of increasing modularity, refactoring is often used to remove overlaps between genes so each is encoded on a separate piece of DNA. The effects of removing overlaps by encoding CDSs into their own distinct sequence regions may disrupt regulatory elements, such as promoters, or important RNA secondary structure elements as well as translational coupling from stop–start overlaps. Genome refactoring was pioneered with the bacteriophage T7 (REF.[133]) but is now commonly applied to biosynthetic gene clusters, where the aim is to exert transcriptional and translational control over the cluster in a heterologous host[122].

Over the past 15 years, a number of genome engineering projects that modified overlapping CDSs and gene 5′ or 3′ UTRs have resulted in losses in viability and efficiency in the final bioengineered product[133–137]. For example, removing CDS overlaps in the bacteriophage T7 resulted in infectious virus yet significantly reduced fitness[133]. Subsequent work using serial passaging and selection for high growth rate over 100 generations was able to show substantial fitness increases similar to pre-adapted wild-type levels[138]. Similarly, a project to ‘decompress’ bacteriophage φX174 had the explicit aim to test the essentiality of CDS overlaps[134]. While coding potential was retained (FIG. 5a), this refactoring led to numerous phenotypic defects, including a substantial reduction in burst size and lower attachment efficiency, along with large changes in levels of several essential assembly and replication proteins produced during the infection cycle[139].

The first complete refactoring of a complex biosynthetic cluster involving overlapping CDSs involved moving the nitrogen fixation cluster of *Klebsiella oxytoca* into *E. coli*[140]. This process involved rebuilding the entire gene cluster from the bottom up, with the removal of non-essential CDSs, codon optimization and disruption of six CDS overlaps (FIG. 5b). In a subsequent, larger project, the group refactored the *Salmonella* pathogenicity island 1 to isolate and control production of the type III secretion system[141]. The refactoring disrupted eight CDS overlaps potentially involved in translational coupling and totalling 90 bp in length. Interestingly, the team discovered that the *spaO* gene contained an in-frame alternative start site at a GTG codon, essentially an in-frame overlapped CDS[141]. In both the nitrogen fixation cluster and the type III secretion system, potential functional deficiencies caused by the removal of CDS overlaps and translational coupling were compensated through careful empirical tuning of the individual ribosome binding sites (RBSs) and transcriptional regulation[140,141].

Other smaller-scale refactoring projects have targeted overlapping CDSs specifically to remove engineering limitations. For example, the gene overlaps in the *dbz* operon in *Rhodococcus erythropolis*, which is used to remove sulfur and upgrade petroleum, were removed to relieve a bottleneck in the efficiency of the process.

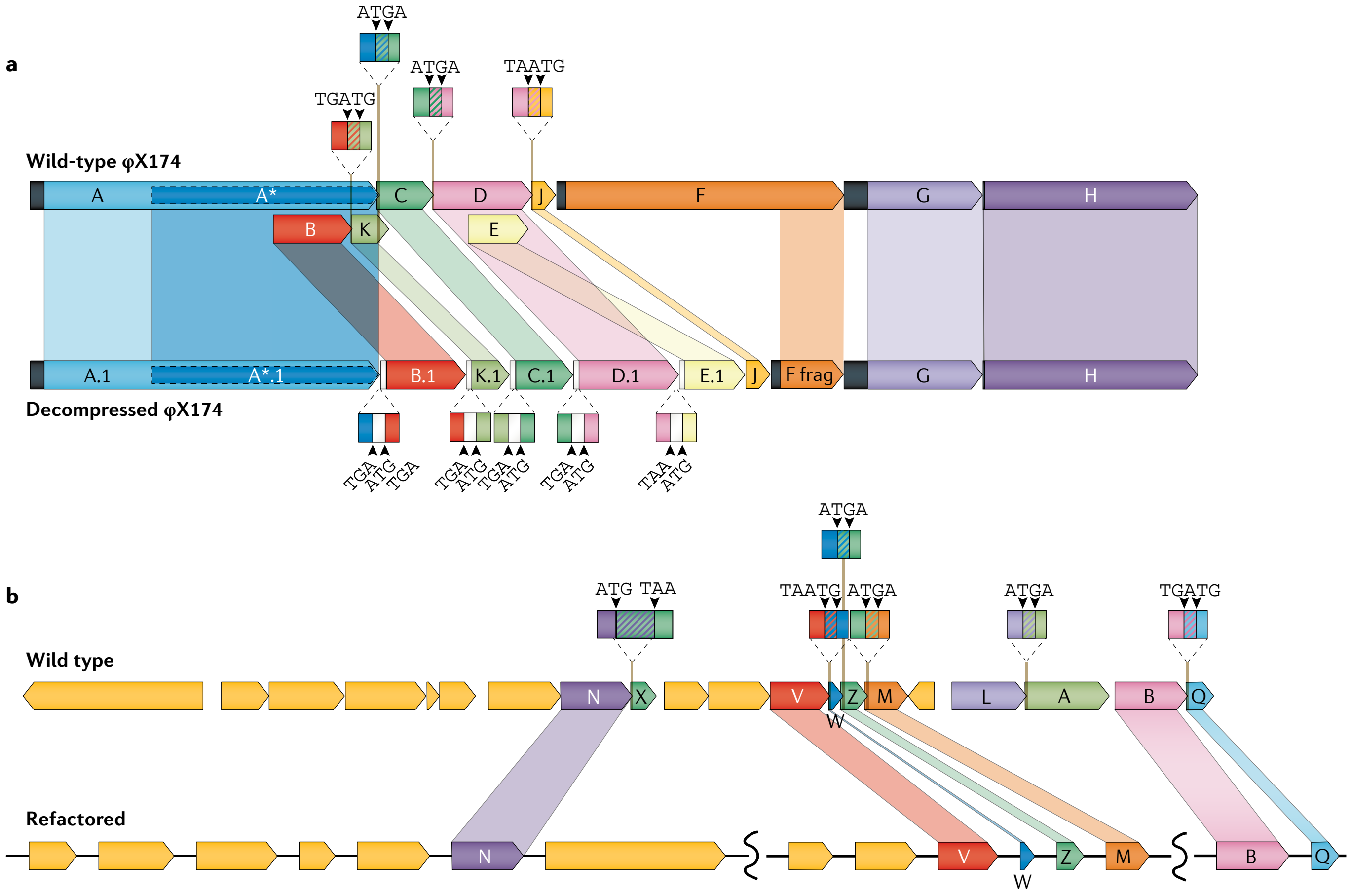

Fig. 5 | **Disruptions to overlapping genes within a refactored phage genome and a complex biosynthetic gene cluster. a** | Creation of φX174.1f, also known as decompressed φX174, disrupted four unidirectional stop–start coding sequence (CDS) overlaps and two fully nested overlapping CDSs[134]. **b** | Refactoring the nitrogen fixation cluster from *Klebsiella oxytoca* disrupted four stop–start CDS overlaps and CDS overlaps varying from 1–14 bp (REF.[140]).

Through rational design targeting the rate-limiting enzyme of this operon (DszB), removal of the overlap of the start and stop codons of *dszA* and *dszB* CDSs resulted in a 12-fold increase in desulfurization activity over the wild-type operon[142]. Similarly, M13 phage CDSs *VII* and *IX* are naturally overlapped, limiting our ability to use the P9 protein for phage display. Removal of the CDS overlap solved this problem, although it resulted in a 1.4-fold decrease in phage infectivity[143].

Beyond refactoring gene clusters, entire cellular genomes have now been refactored. During the design of the synthetic yeast chromosomes in the Yeast2.0 project, 15 instances of ORF overlap were identified where the desired TAG>TAA stop codon swap would have altered the codons of a verified ORF; however, details of how each instance was specifically addressed was not reported[144]. An *E. coli* genome engineering project to replace all 321 instances of TAG stop codons with TAA encountered several instances of CDS overlap where replacement might affect one of the partners. The first instance was the convergent overlapping *yegV* and *yegW* CDSs (both contained TAG stop codons in the overlapped region). Fortuitously, conversion of both overlapping TAGs to TAA conserved amino acid identity of the opposite CDS[145].

A more extensive refactoring project to create an *E. coli* with a 61-codon genome via the removal of two sense (TCG and TCA) and one stop codon (TAG) encountered 91 instances of where these codons occurred in a region overlapping two CDSs[132]. If the overlapping CDSs were convergent, either silent mutations were incorporated or, if otherwise unavailable, the CDSs were separated by duplicating the overlap region followed by their independent recoding. In instances of unidirectional overlap, the CDSs were separated by duplicating the overlap region plus 20 bp upstream for a synthetic insert. At the start of this insert, an in-frame stop codon (TAA) was added to terminate translation from the original RBS. The result of this sophisticated refactoring process produced a viable *E. coli* albeit with a doubling time 1.6x longer than the parent strain under standard conditions[132]. Due to the vast number of changes across the genome, it is not possible at this time to attribute the slowed growth rate to CDS overlap disruption, although translational coupling between unidirectional overlaps would likely be disrupted by the RBS duplication protocol.

Conversely, some studies have taken a more cautious approach towards overlapping genes. For example, in the construction of the widely used *E. coli* K-12 single knockout library (Keio collection), deletions of dual coding regions were avoided by conserving overlap regions[146]. Similarly, in the minimal *Mycoplasma mycoides* genome, instances in which a retained CDS (essential or quasi-essential) was partially overlapped with a CDS to be deleted (non-essential) resulted in the overlapping region being retained[136].

Applications of engineered overlapping genes. With increasing recognition that gene overlaps are functionally important and play vital roles within natural organisms, the construction of new overlapping genes has begun to be exploited in bioengineering. Theoretical work has previously shown that the genetic code is flexible enough to accommodate artificial overlap of protein domains[147,148], and even artificial proteins[149], often with the stated aim to protect the overlapping CDS from genetic drift[150] in similar ways to that found in viruses[71,75,151].

Recently, two methods for generating artificial CDS overlaps between a gene of interest and an essential gene have been described and empirically tested[152,153]. The Constraining Adaptive Mutations using Engineered Overlapping Sequences (CAMEOS) method[152] searches for available overlaps between the CDSs of an essential gene and a gene of interest to be shielded from mutation[152] (FIG. 6a). The algorithm uses a two-step process that relies on pre-existing or newly computed statistical models of the protein families that are being assessed for overlap. Furthermore, the CAMEOS dynamic programming algorithm searches for optimal solutions that consider both short-range (local codon usage) and long-range (epistatic) interactions while minimizing amino acid changes of the encoded proteins. CAMEOS was capable of creating a synthetic amino acid biosynthetic gene containing two additional out-of-frame nested essential CDSs. Protein functionality was maintained in the encoded enzymes despite up to 50% non-conservative amino acid changes and runs of up to six consecutive amino acid changes. Assessments of mutational robustness in the first 30 codons of the new CDS overlaps showed that the recodings were able to prevent any sequence changes to the non-essential CDS over 150 generations of growth, whereas the control CDS without overlap mutated by generation 50. The method was also shown to have some promise in the biocontainment of engineered constructs by overlapping a toxin gene with a gene of interest. If the engineered CDS is transferred to another organism, the toxin CDS will either kill the host or there will be a mutation in the toxin CDS that also inactivates the engineered CDS, thereby ensuring that the enhanced bioengineered phenotype is not transferred into the environment[152].

The RiBoSor method[153] takes a distinct approach to create a synthetic CDS overlap to protect a CDS of interest from mutation. The algorithm searches for locations within a CDS to silently create an out-of-frame RBS and start codon (FIG. 6b). The objective is to create a CDS in a different reading frame, called a Riboverlap, that runs uninterrupted to the 3′ end of the CDS of interest. If stop codons occur that would interrupt the new synthetic overlapped CDS, the algorithm tries to silently change them. An essential CDS is then fused in-frame to the newly created CDS just 3′ of the stop codon of the CDS of interest. Theoretically, this method should be both computationally simpler and more flexible than CAMEOS but also potentially less effective at constraining mutational pressure on the CDS of interest.

Another engineered genetic architecture taking inspiration from natural genomes features a CDS of interest directly downstream and overlapping a short translated CDS. Importantly, within this short coding sequence is the RBS site of the gene of interest that leads to a stop–start overlapping codon junction, facilitating coupled translation. Originally implemented by placing the *trpE*/*trpD* translationally coupled stop–start sequence upstream of the human γ-interferon gene[154], a standardized bicistronic device architecture was recently created[155]. The bicistronic architecture results in robust and tunable protein expression regardless of the gene of interest (FIG. 6c). The success of this approach has been demonstrated in several studies[155–157]. Translational coupling in eukaryotes is currently less amenable to exploitation due to the mainly monocistronic mRNAs. However, there are increasing numbers of polycistronic transcripts being documented that suggest that this architecture may be useful in eukaryotes if correctly implemented[158].

Creating organisms with new genetic codes will have a profound effect on the exploitation of overlapping genes in both positive and negative ways. For example, removing synonymous coding capacity within a genome to free up codons for encoding unnatural amino acids[132,159,160] will make it difficult or impossible to retain existing CDS overlaps that rely on the degeneracy of the second and third codon positions. Conversely, engineering ribosomes to decode 4-nt codons[161,162] will also expand the potential for synonymous codons and overlapping CDSs. New six-letter and eight-letter genetic codes[163,164] could provide many additional synonymous codons for extensive overlapping CDS possibilities. However, substantial effort would be needed to create the multitude of tRNA–aminoacyl synthetase pairs[126] to make this a reality. A 256-codon genetic code would allow up to 12 synonymous codons per amino acid, greatly expanding CDS overlap opportunities.

Conclusions and future perspectives

In this Review, we sought to highlight gene overlaps from a wide variety of genomes across the diversity of biology. There has been a vigorous renewal of interest in overlapping genes that can be directly attributed to recent advances in bioinformatics, sequencing and allied proteogenomic technologies. Overlapping genes, transcripts and ORFs have been a part of genome biology from the first sequenced RNA and DNA-based genomes[2,165]; however, their abundance and ubiquity have only just come into focus for eukaryotic genomes with the advent of recent genome-scale measurement technologies. From past and present literature, it seems clear that the definitions and assessments of overlap topology between eukaryotic, prokaryotic and viral genomes have been disconnected. It is unclear how this discordance arose; however, differing genome

Bicistronic
A transcript (mRNA) that encodes two CDSs.

Monocistronic
An mRNA transcript that contains a single CDS.

Polycistronic
An mRNA transcript that contains two or more CDSs.

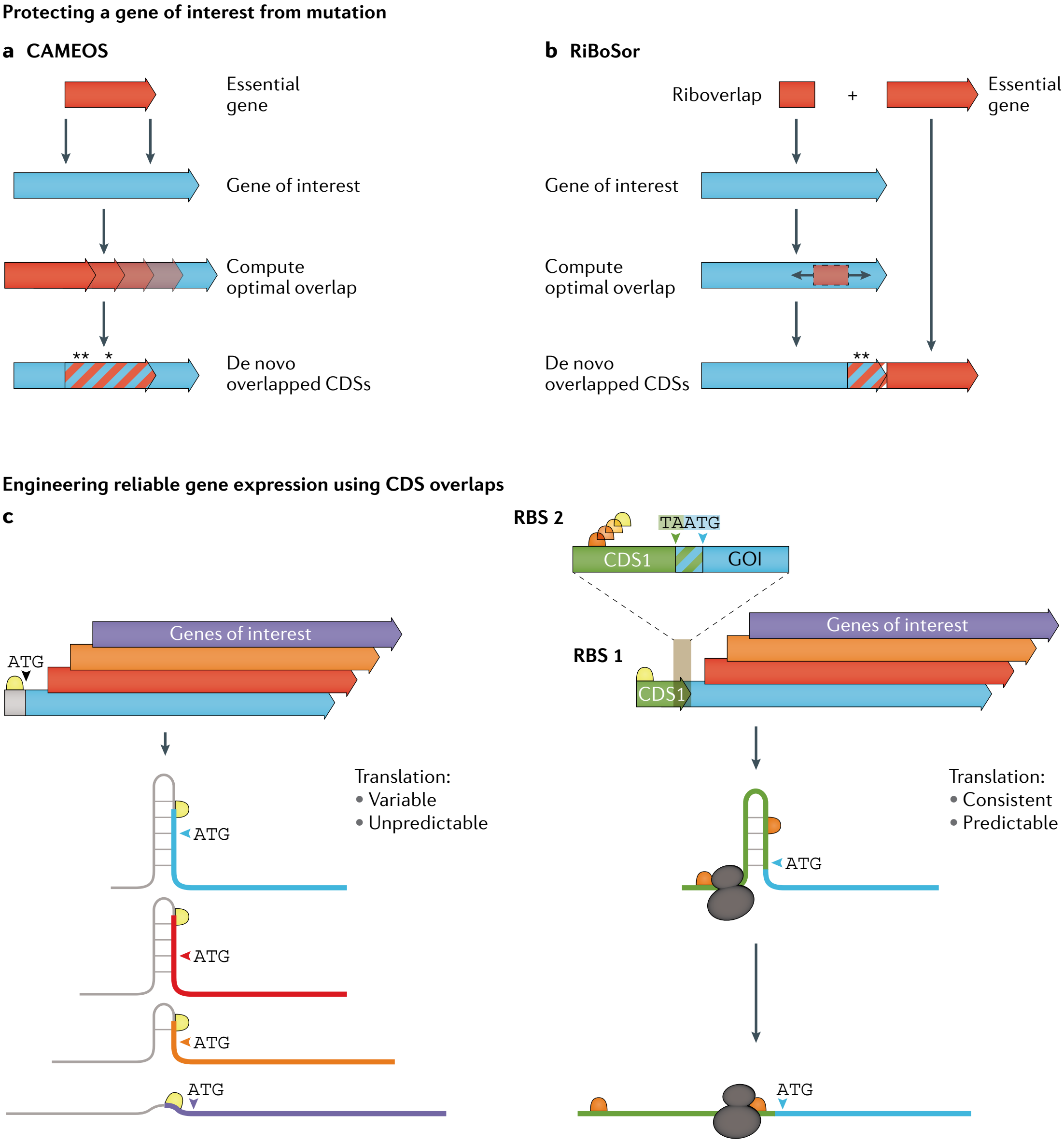

Fig. 6 | **Exploiting CDS overlap for applications in bioengineering. a** | The Constraining Adaptive Mutations using Engineered Overlapping Sequences (CAMEOS) method searches for available overlaps between an essential gene and a gene of interest to be shielded from mutation while minimizing sequence changes[152]. Asterisks identify amino acid modifications to accommodate coding sequence (CDS) overlap. **b** | The RiBoSor algorithm searches for places within a gene of interest to silently create a ribosome binding site followed by a start codon in a different reading frame than the existing CDS. This generates a CDS that extends to the 3′ end of the existing CDS. An essential gene is then fused in-frame to the newly created CDS just 3′ of the stop codon of the original CDS[153]. Asterisks identify amino acid modifications to accommodate overlap. **c** | Reliable and tunable expression of a gene of interest can be facilitated by the bicistronic device[155]. (Left) A single ribosome binding site upstream of a variety of different CDSs can result in different interactions with the RBS and the coding sequence, causing variable translation initation rates that are difficult to predict. (Right) In the bicistronic device, the binding of a ribosome to the upstream ribosome binding site 1 of CDS 1 and its translation towards the gene of interest will disrupt inhibitory sequence structures. The ribosome will recognize the ribosome binding site (RBS2) of the downstream gene of interest and re-initiate translation providing a platform for reliable expression of the gene of interest.

architecture, biology and researcher fields of interest are likely notable contributors. As new technologies, such as ribosome profiling, are showcasing, eukaryotes (and in particular humans) seem to encode an abundance of small and alternative overlapping ORFs[6,11,12,14,166] spurring excitement in this genome biology. Future work will show whether the majority are true functional overlaps between protein-coding ORFs, non-coding translational regulatory regions or a result of measurement biases.

Going forwards, it would be highly desirable to harmonize the definition of gene overlap between eukaryotes, prokaryotes and viruses. This would enable true comparisons of overlap topology prevalence, more robust evolutionary studies, and highlight any domain-specific mechanisms contributing to overlap birth and fixation. One likely reason for the differences in gene overlap definition is the transcript-centric gene definition currently dominant in eukaryotes, which has not yet been adopted in prokaryote biology. This is undoubtedly due to the technical difficulty in defining prokaryotic transcripts compared to eukaryotic transcripts as well as to its lower emphasis in the field (until recently[167,168]).

In addition to different definitions of gene overlap, the way overlapping CDSs in the same loci are treated in prokaryote and eukaryote genome annotation is distinctly different. For example, p16INK4a and tumour suppressor ARF in humans are considered splice variants of the *CDKN2A* gene despite not sharing any sequence identity whereas, if these were found in a prokaryotic genome, they would be annotated as different genes. For eukaryotes, this is possibly changing with new proposals to annotate these overlapping ORFs as different genes are currently proposed[14].

Lastly, the conventional idea of the monocistronic eukaryotic transcript is slowly being eroded[64,169] with the advent of new research demonstrating transcripts harbouring multiple CDSs[66,67]. Moving away from this out-of-date convention will encourage researchers to pursue new lines of enquiry, such as the biological significance of polycistronic arrangements (well-known as operons in prokaryotes), or expanded insights into translation initiation.

We also discussed instances where engineered systems can take inspiration from natural overlapped gene systems for a variety of applications. Synthetic biology and genetic refactoring methods are frequently testing the limits of modifying and reformatting gene architectures in heterologous and endogenous hosts. We also identified future bioengineering research in expanded genetic codes that could make engineered gene overlaps more accessible and exploitable.

The rapidly advancing area of synthetic genomics, where entire genomes are being constructed anew, often with radically different topologies and overlapping genes disrupted or removed entirely, will require a much deeper understanding of genotype–phenotype relationships than we currently enjoy. Alternative and expanded genetic codes and codon-decoding capacity will open up new exciting possibilities for the design of extensively overlapped genetic systems to resist evolutionary drift and add additional functionality to new biotechnological applications of engineered overlapping genes not yet envisioned.

Published online 5 October 2021

1. Barrell, B. G., Air, G. M. & Hutchison, C. A. 3rd Overlapping genes in bacteriophage phiX174. *Nature* **264**, 34–41 (1976).
2. Sanger, F. et al. Nucleotide sequence of bacteriophage φX174 DNA. *Nature* **265**, 687 (1977).
3. Linney, E. & Hayashi, M. Intragenic regulation of the synthesis of ΦX174 gene A proteins. *Nature* **249**, 345 (1974).
4. Roznowski, A. P., Doore, S. M., Kemp, S. Z. & Fane, B. A. Finally, a role befitting Astar: the strongly conserved, unessential microvirus A* proteins ensure the product fidelity of packaging reactions. *J. Virol.* **94**, e01593-19 (2020).
5. Schlub, T. E. & Holmes, E. C. Properties and abundance of overlapping genes in viruses. *Virus Evol.* **6**, veaa009 (2020).
This article provides an unparalleled look into the properties of gene overlaps amongst the NCBI virus reference genome database.
6. Nelson, C. W. et al. Dynamically evolving novel overlapping gene as a factor in the SARS-CoV-2 pandemic. *eLife* **9**, e59633 (2020).
7. Normark, S. et al. Overlapping genes. *Annu. Rev. Genet.* **17**, 499–525 (1983).
8. Sanna, C. R., Li, W.-H. & Zhang, L. Overlapping genes in the human and mouse genomes. *BMC Genomics* **9**, 169 (2008).
9. Veeramachaneni, V., Makalowski, W., Galdzicki, M., Sood, R. & Makalowska, I. Mammalian overlapping genes: the comparative perspective. *Genome Res.* **14**, 280–286 (2004).
10. Chen, C.-H., Pan, C.-Y. & Lin, W.-C. Overlapping protein-coding genes in human genome and their coincidental expression in tissues. *Sci. Rep.* **9**, 13377 (2019).
11. Chen, J. et al. Pervasive functional translation of noncanonical human open reading frames. *Science* **367**, 1140–1146 (2020).
This seminal study provides robust evidence for many new overlapping genes within human genomes combining proteogenomics and CRISPR functional screens.
12. Prensner, J. R. et al. Noncanonical open reading frames encode functional proteins essential for cancer cell survival. *Nat. Biotechnol.* **39**, 697–704 (2021).
13. Wang, B. et al. Identification and analysis of small proteins and short open reading frame encoded peptides in Hep3B cell. *J. Proteom.* **230**, 103965 (2021).
14. Mudge, J. M. et al. A community-driven roadmap to advance research on translated open reading frames detected by Ribo-seq. *bioRxiv* https://doi.org/10.1101/2021.06.10.447896 (2021).
15. Hinnebusch, A. G., Ivanov, I. P. & Sonenberg, N. Translational control by 5′-untranslated regions of eukaryotic mRNAs. *Science* **352**, 1413–1416 (2016).
16. Wu, Q. et al. Translation of small downstream ORFs enhances translation of canonical main open reading frames. *EMBO J.* **39**, e104763 (2020).
17. Johnson, Z. I. & Chisholm, S. W. Properties of overlapping genes are conserved across microbial genomes. *Genome Res.* **14**, 2268–2272 (2004).
18. Cock, P. J. A. & Whitworth, D. E. Evolution of relative reading frame bias in unidirectional prokaryotic gene overlaps. *Mol. Biol. Evol.* **27**, 753–756 (2009).
19. Lillo, F. & Krakauer, D. C. A statistical analysis of the three-fold evolution of genomic compression through frame overlaps in prokaryotes. *Biol. Direct* **2**, 22–22 (2007).
20. Assis, R., Kondrashov, A. S., Koonin, E. V. & Kondrashov, F. A. Nested genes and increasing organizational complexity of metazoan genomes. *Trends Genet.* **24**, 475–478 (2008).
This study analyses nested protein-coding genes within introns of select metazoan genomes and shows that increasing overlap complexity is a result of a neutral evolutionary process.
21. Yu, P., Ma, D. & Xu, M. Nested genes in the human genome. *Genomics* **86**, 414–422 (2005).
22. Van Oss, S. B. & Carvunis, A. R. De novo gene birth. *PLoS Genet.* **15**, e1008160 (2019).
23. Makalowska, I., Lin, C.-F. & Makalowski, W. Overlapping genes in vertebrate genomes. *Comput. Biol. Chem.* **29**, 1–12 (2005).
24. Wichmann, S. & Ardern, Z. Optimality in the standard genetic code is robust with respect to comparison code sets. *Biosystems* **185**, 104023 (2019).
25. Soldà, G. et al. Non-random retention of protein-coding overlapping genes in Metazoa. *BMC Genomics* **9**, 174 (2008).
26. Krakauer, D. C. Stability and evolution of overlapping genes. *Evol. Int. J. Org. Evol.* **54**, 731–739 (2000).
27. Rogozin, I. B. et al. Purifying and directional selection in overlapping prokaryotic genes. *Trends Genet.* **18**, 228–232 (2002).
This paper describes selective pressures on conserved regions of long (>15 bp) overlap within prokaryotes.
28. Hamoen, L. W., Eshuis, H., Jongbloed, J., Venema, G. & van Sinderen, D. A small gene, designated *comS*, located within the coding region of the fourth amino acid-activation domain of *srfA*, is required for competence development in *Bacillus subtilis*. *Mol. Microbiol.* **15**, 55–63 (1995).
29. Meydan, S. et al. Retapamulin-assisted ribosome profiling reveals the alternative bacterial proteome. *Mol. Cell* **74**, 481–493.e6 (2019).
This article describes the first use of retapamulin translation initiation inhibitor enabling Ribo-seq measurements of many new translation initiation sites within existing genes in *E. coli*.
30. Feltens, R., Gossringer, M., Willkomm, D. K., Urlaub, H. & Hartmann, R. K. An unusual mechanism of bacterial gene expression revealed for the RNase P protein of Thermus strains. *Proc. Natl Acad. Sci. USA* **100**, 5724–5729 (2003).
31. Jones, C. E., Fleming, T. M., Cowan, D. A., Littlechild, J. A. & Piper, P. W. The phosphoglycerate kinase and glyceraldehyde-3-phosphate dehydrogenase genes from the thermophilic archaeon *Sulfolobus solfataricus* overlap by 8-bp. Isolation, sequencing of the genes and expression in *Escherichia coli*. *Eur. J. Biochem.* **233**, 800–808 (1995).
32. Fukuda, Y., Nakayama, Y. & Tomita, M. On dynamics of overlapping genes in bacterial genomes. *Gene* **323**, 181–187 (2003).
33. Sakharkar, K. R., Sakharkar, M. K., Verma, C. & Chow, V. T. K. Comparative study of overlapping genes in bacteria, with special reference to *Rickettsia prowazekii* and *Rickettsia conorii*. *Int. J. Syst. Evol. Microbiol.* **55**, 1205–1209 (2005).
34. Fonseca, M. M., Harris, D. J. & Posada, D. Origin and length distribution of unidirectional prokaryotic overlapping genes. *G3* **4**, 19–27 (2013).
35. Palleja, A., Harrington, E. D. & Bork, P. Large gene overlaps in prokaryotic genomes: result of functional constraints or mispredictions? *BMC Genomics* **9**, 335 (2008).
36. Price, M. N., Arkin, A. P. & Alm, E. J. The life-cycle of operons. *PLoS Genet.* **2**, e96 (2006).
37. Delcher, A. L., Kingsford, C. & Salzberg, S. L. A unified model explaining the offsets of overlapping and near-overlapping prokaryotic genes. *Mol. Biol. Evol.* **24**, 2091–2098 (2007).
38. Huvet, M. & Stumpf, M. P. H. Overlapping genes: a window on gene evolvability. *BMC Genomics* **15**, 721 (2014).
39. Tatusova, T. et al. NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Res.* **44**, 6614–6624 (2016).
40. Willis, S. & Masel, J. Gene birth contributes to structural disorder encoded by overlapping genes. *Genetics* **210**, 303–313 (2018).

41. Pavesi, A. et al. Overlapping genes and the proteins they encode differ significantly in their sequence composition from non-overlapping genes. *PLoS One* **13**, e0202513 (2018).
42. Kreitmeier, M. et al. Shadow ORFs illuminated: long overlapping genes in *Pseudomonas aeruginosa* are translated and under purifying selection. *bioRxiv* https://doi.org/10.1101/2021.02.09.430400 (2021).
43. Zehentner, B., Ardern, Z., Kreitmeier, M., Scherer, S. & Neuhaus, K. Evidence for numerous embedded antisense overlapping genes in diverse *E. coli* strains. *bioRxiv* https://doi.org/10.1101/2020.11.18.388249 (2020).
44. Zehentner, B., Ardern, Z., Kreitmeier, M., Scherer, S. & Neuhaus, K. 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 (2020).
45. Sabath, N., Ferrada, E., Barve, A. & Wagner, A. Growth temperature and genome size in bacteria are negatively correlated, suggesting genomic streamlining during thermal adaptation. *Genome Biol. Evol.* **5**, 966–977 (2013).
46. Saha, D., Panda, A., Podder, S. & Ghosh, T. C. Overlapping genes: a new strategy of thermophilic stress tolerance in prokaryotes. *Extremophiles* **19**, 345–353 (2015).
47. Pradhan, P., Li, W. & Kaur, P. Translational coupling controls expression and function of the DrrAB drug efflux pump. *J. Mol. Biol.* **385**, 831–842 (2009).
48. Huber, M. et al. Translational coupling via termination-reinitiation in archaea and bacteria. *Nat. Commun.* **10**, 4006 (2019).
49. Das, A. & Yanofsky, C. Restoration of a translational stop-start overlap reinstates translational coupling in a mutant *trpB'-trpA* gene pair of the *Escherichia coli* tryptophan operon. *Nucleic acids Res.* **17**, 9333–9340 (1989).
50. Das, A. & Yanofsky, C. A ribosome binding site sequence is necessary for efficient expression of the distal gene of a translationally-coupled gene pair. *Nucleic acids Res.* **12**, 4757–4768 (1984).
51. Price, M. N., Huang, K. H., Arkin, A. P. & Alm, E. J. Operon formation is driven by co-regulation and not by horizontal gene transfer. *Genome Res.* **15**, 809–819 (2005).
52. Osbourn, A. E. & Field, B. Operons. *Cell. Mol. Life Sci.* **66**, 3755–3775 (2009).
53. Shieh, Y. W. et al. Operon structure and cotranslational subunit association direct protein assembly in bacteria. *Science* **350**, 678–680 (2015).
54. Meydan, S., Vázquez-Laslop, N. & Mankin, A. S. Genes within genes in bacterial genomes. *Microbiol Spectr.* https://doi.org/10.1128/microbiolspec.RWR-0020-2018 (2018).
55. Jeffares, D. C., Mourier, T. & Penny, D. The biology of intron gain and loss. *Trends Genet.* **22**, 16–22 (2006).
56. Williams, B. A., Slamovits, C. H., Patron, N. J., Fast, N. M. & Keeling, P. J. A high frequency of overlapping gene expression in compacted eukaryotic genomes. *Proc. Natl Acad. Sci. USA* **102**, 10936–10941 (2005).
57. Makałowska, I., Lin, C.-F. & Hernandez, K. Birth and death of gene overlaps in vertebrates. *BMC Evolut. Biol.* **7**, 193 (2007).
    This paper gives a detailed account of gene overlaps in model vertebrate genomes and quantifies the gain and loss of overlaps across evolutionary time, revealing that many overlaps are young and lineage specific.
58. Kumar, A. An overview of nested genes in eukaryotic genomes. *Eukaryot. Cell* **8**, 1321 (2009).
59. Lee, Y. C. G. & Chang, H.-H. The evolution and functional significance of nested gene structures in Drosophila melanogaster. *Genome Biol. Evol.* **5**, 1978–1985 (2013).
60. Heames, B., Schmitz, J. & Bornberg-Bauer, E. A continuum of evolving de novo genes drives protein-coding novelty in *Drosophila*. *J. Mol. Evol.* **88**, 382–398 (2020).
61. Rindfleisch, B. C., Brown, M. S., VandeBerg, J. L. & Munroe, S. H. Structure and expression of two nuclear receptor genes in marsupials: insights into the evolution of the antisense overlap between the α-thyroid hormone receptor and Rev-erbα. *BMC Mol. Biol.* **11**, 97 (2010).
62. Loughran, G. et al. Unusually efficient CUG initiation of an overlapping reading frame in POLG mRNA yields novel protein POLGARF. *Proc. Natl Acad. Sci. USA* **117**, 24936–24946 (2020).
63. Khan, Y. A. et al. Evidence for a novel overlapping coding sequence in POLG initiated at a CUG start codon. *BMC Genet.* **21**, 25 (2020).
64. Brunet, M. A., Levesque, S. A., Hunting, D. J., Cohen, A. A. & Roucou, X. Recognition of the polycistronic nature of human genes is critical to understanding the genotype-phenotype relationship. *Genome Res.* **28**, 609–624 (2018).
65. Cao, X. et al. Alt-RPL36 downregulates the PI3K-AKT-mTOR signaling pathway by interacting with TMEM24. *Nat. Commun.* **12**, 508 (2021).
    This paper describes the discovery and biology of an alternative overlapping reading frame within the human *RPL36* gene.
66. Samandi, S. et al. Deep transcriptome annotation enables the discovery and functional characterization of cryptic small proteins. *eLife* **6**, e27860 (2017).
67. Karginov, T. A., Pastor, D. P. H., Semler, B. L. & Gomez, C. M. Mammalian polycistronic mRNAs and disease. *Trends Genet.* **33**, 129–142 (2017).
68. Sherr, C. J. The INK4a/ARF network in tumour suppression. *Nat. Rev. Mol. Cell Biol.* **2**, 731–737 (2001).
69. Brunet, M. A. et al. The FUS gene is dual-coding with both proteins contributing to FUS-mediated toxicity. *EMBO Rep.* **22**, e50640 (2021).
70. Muñoz-Baena, L. & Poon, A. F. Y. Using networks to analyze and visualize the distribution of overlapping reading frames in virus genomes. *bioRxiv* https://doi.org/10.1101/2021.06.10.447953 (2021).
71. Simon-Loriere, E., Holmes, E. C. & Pagán, I. The effect of gene overlapping on the rate of RNA virus evolution. *Mol. Biol. Evol.* **30**, 1916–1928 (2013).
72. Krakauer, D. C. & Plotkin, J. B. Redundancy, antiredundancy, and the robustness of genomes. *Proc. Natl Acad. Sci. USA* **99**, 1405–1409 (2002).
73. Chirico, N., Vianelli, A. & Belshaw, R. Why genes overlap in viruses. *Proc. Biol. Sci.* **277**, 3809–3817 (2010).
74. Belshaw, R., Pybus, O. G. & Rambaut, A. The evolution of genome compression and genomic novelty in RNA viruses. *Genome Res.* **17**, 1496–1504 (2007).
75. Fernandes, J. D. et al. Functional segregation of overlapping genes in HIV. *Cell* **167**, 1762–1773.e12 (2016).
    Using mutational and statistical analyses on the overlapped *tat* and *rev* genes of HIV1, the authors reveal that overlaps have population fitness advantages.
76. Brandes, N. & Linial, M. Gene overlapping and size constraints in the viral world. *Biol. Direct* **11**, 26 (2016).
    This article performs a unique analysis of the relationship between virus genome sizes and associated capsids, coming to the surprising conclusion that most viruses likely have excess room within their capsids.
77. Pavesi, A. New insights into the evolutionary features of viral overlapping genes by discriminant analysis. *Virology* **546**, 51–66 (2020).
78. Feiss, M., Fisher, R. A., Crayton, M. A. & Egner, C. Packaging of the bacteriophage λ chromosome: Effect of chromosome length. *Virology* **77**, 281–293 (1977).
79. Aoyama, A. & Hayashi, M. Effects of genome size on bacteriophage phi X174 DNA packaging in vitro. *J. Biol. Chem.* **260**, 11033–11038 (1985).
80. Wu, Z., Yang, H. & Colosi, P. Effect of genome size on AAV vector packaging. *Mol. Ther.* **18**, 80–86 (2010).
81. Vaidyanathan, S. et al. Targeted replacement of full-length CFTR in human airway stem cells by CRISPR-Cas9 for pan-mutation correction in the endogenous locus. *Mol. Ther.* https://doi.org/10.1016/j.ymthe.2021.03.023 (2021).
82. Bartonek, L., Braun, D. & Zagrovic, B. Frameshifting preserves key physicochemical properties of proteins. *Proc. Natl Acad. Sci. USA* **117**, 5907 (2020).
    This article revealed that some key physicochemical properties are maintained in an altered protein sequence that is produced as a result of a +1 or −1 frameshift mutation, which therefore suggests a plausible explanation for the abundance of this overlap offset.
83. Li, H., Havens, W. M., Nibert, M. L. & Ghabrial, S. A. RNA sequence determinants of a coupled termination-reinitiation strategy for downstream open reading frame translation in Helminthosporium victoriae virus 190S and other victoriviruses (family Totiviridae). *J. Virol.* **85**, 7343–7352 (2011).
84. Toledo-Arana, A. & Lasa, I. Advances in bacterial transcriptome understanding: From overlapping transcription to the excludon concept. *Mol. Microbiol.* **113**, 593–602 (2020).
85. Gelsinger, D. R. & DiRuggiero, J. Transcriptional landscape and regulatory roles of small noncoding RNAs in the oxidative stress response of the Haloarchaeon *haloferax volcanii*. *J. Bacteriol.* **200**, e00779-17 (2018).
86. Choi, J. S., Park, H., Kim, W. & Lee, Y. Coordinate regulation of the expression of SdsR toxin and its downstream pphA gene by RyeA antitoxin in Escherichia coli. *Sci. Rep.* **9**, 9627 (2019).
87. Lee, E.-J. & Groisman, E. A. An antisense RNA that governs the expression kinetics of a multifunctional virulence gene. *Mol. Microbiol.* **76**, 1020–1033 (2010).
88. Sesto, N., Wurtzel, O., Archambaud, C., Sorek, R. & Cossart, P. The excludon: a new concept in bacterial antisense RNA-mediated gene regulation. *Nat. Rev. Microbiol.* **11**, 75–82 (2013).
89. Dar, D. & Sorek, R. Bacterial noncoding RNAs excised from within protein-coding transcripts. *mBio* **9**, e01730-18 (2018).
90. Adams, P. P. & Storz, G. Prevalence of small base-pairing RNAs derived from diverse genomic loci. *Biochim. Biophys. Acta Gene Regul. Mech.* **1863**, 194524 (2020).
91. Adams, P. P. et al. Regulatory roles of Escherichia coli 5′ UTR and ORF-internal RNAs detected by 3′ end mapping. *eLife* **10**, e62438 (2021).
92. Neuhaus, K. et al. 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 (2017).
93. Vanderpool, C. K., Balasubramanian, D. & Lloyd, C. R. Dual-function RNA regulators in bacteria. *Biochimie* **93**, 1943–1949 (2011).
94. Jin, H., Vacic, V., Girke, T., Lonardi, S. & Zhu, J.-K. Small RNAs and the regulation of cis-natural antisense transcripts in Arabidopsis. *BMC Mol. Biol.* **9**, 6 (2008).
95. Faghihi, M. A. & Wahlestedt, C. Regulatory roles of natural antisense transcripts. *Nat. Rev. Mol. Cell Biol.* **10**, 637–643 (2009).
96. Pelechano, V. & Steinmetz, L. M. Gene regulation by antisense transcription. *Nat. Rev. Genet.* **14**, 880–893 (2013).
97. Werner, A. Biological functions of natural antisense transcripts. *BMC Biol.* **11**, 31 (2013).
98. Matsuda, E. & Garfinkel, D. J. Posttranslational interference of Ty1 retrotransposition by antisense RNAs. *Proc. Natl Acad. Sci. USA* **106**, 15657–15662 (2009).
99. Chu, J. & Dolnick, B. J. Natural antisense (rTSα) RNA induces site-specific cleavage of thymidylate synthase mRNA. *Biochim. Biophys. Acta* **1587**, 183–193 (2002).
100. Morrissy, A. S., Griffith, M. & Marra, M. A. Extensive relationship between antisense transcription and alternative splicing in the human genome. *Genome Res.* **21**, 1203–1212 (2011).
101. Gong, C. & Maquat, L. E. lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3′ UTRs via Alu elements. *Nature* **470**, 284–288 (2011).
102. Su, W.-Y. et al. Bidirectional regulation between WDR83 and its natural antisense transcript DHPS in gastric cancer. *Cell Res.* **22**, 1374–1389 (2012).
103. Jeon, Y., Sarma, K. & Lee, J. T. New and Xisting regulatory mechanisms of X chromosome inactivation. *Curr. Opin. Genet. Dev.* **22**, 62–71 (2012).
104. Chu, C., Qu, K., Zhong, F. L., Artandi, S. E. & Chang, H. Y. Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions. *Mol. Cell* **44**, 667–678 (2011).
105. Chen, J., Sun, M., Hurst, L. D., Carmichael, G. G. & Rowley, J. D. Genome-wide analysis of coordinate expression and evolution of human cis-encoded sense-antisense transcripts. *Trends Genet.* **21**, 326–329 (2005).
106. Kapranov, P., Willingham, A. T. & Gingeras, T. R. Genome-wide transcription and the implications for genomic organization. *Nat. Rev. Genet.* **8**, 413–423 (2007).
107. Wu, P. et al. Emerging role of tumor-related functional peptides encoded by lncRNA and circRNA. *Mol. Cancer* **19**, 22 (2020).
108. Reis, E. M. et al. Antisense intronic non-coding RNA levels correlate to the degree of tumor differentiation in prostate cancer. *Oncogene* **23**, 6684–6692 (2004).

109. Yin, J. et al. UXT-AS1-induced alternative splicing of UXT is associated with tumor progression in colorectal cancer. *Am. J. Cancer Res.* **7**, 462–472 (2017).
110. Tu, Q. et al. CDKN2B deletion is essential for pancreatic cancer development instead of unmeaningful co-deletion due to juxtaposition to CDKN2A. *Oncogene* **37**, 128–138 (2018).
111. Yu, W. et al. Epigenetic silencing of tumour suppressor gene p15 by its antisense RNA. *Nature* **451**, 202–206 (2008).
112. Tufarelli, C. et al. Transcription of antisense RNA leading to gene silencing and methylation as a novel cause of human genetic disease. *Nat. Genet.* **34**, 157–165 (2003).
113. Jackson, R. et al. The translation of non-canonical open reading frames controls mucosal immunity. *Nature* **564**, 434–438 (2018).
114. Slavoff, S. A. et al. Peptidomic discovery of short open reading frame–encoded peptides in human cells. *Nat. Chem. Biol.* **9**, 59–64 (2013).
115. Ma, J. et al. Discovery of human sORF-encoded polypeptides (SEPs) in cell lines and tissue. *J. Proteome Res.* **13**, 1757–1765 (2014).
116. Bazzini, A. A. et al. Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO J.* **33**, 981–993 (2014).
117. Guo, B. et al. Micropeptide CIP2A-BP encoded by LINC00665 inhibits triple-negative breast cancer progression. *EMBO J.* **39**, e102190 (2020).
118. Lee, C. Q. E. et al. Coding and non-coding roles of MOCCI (C15ORF48) coordinate to regulate host inflammation and immunity. *Nat. Commun.* **12**, 2130 (2021).
119. Nam, J.-W., Choi, S.-W. & You, B.-H. Incredible RNA: dual functions of coding and noncoding. *Mol. Cell* **39**, 367–374 (2016).
120. Schlesinger, D. & Elsässer, S. J. Revisiting sORFs: overcoming challenges to identify and characterize functional microproteins. *FEBS J.* https://doi.org/10.1111/febs.15769 (2021).
121. Ruiz-Orera, J., Villanueva-Cañas, J. L. & Albà, M. M. Evolution of new proteins from translated sORFs in long non-coding RNAs. *Exp. Cell Res.* **391**, 111940 (2020).
122. Smanski, M. J. et al. Synthetic biology to access and expand nature's chemical diversity. *Nat. Rev. Microbiol.* **14**, 135–149 (2016).
123. Bayer, T. S. et al. Synthesis of methyl halides from biomass using engineered microbes. *J. Am. Chem. Soc.* **131**, 6508–6515 (2009).
124. Segall-Shapiro, T. H., Meyer, A. J., Ellington, A. D., Sontag, E. D. & Voigt, C. A. A 'resource allocator' for transcription based on a highly fragmented T7 RNA polymerase. *Mol. Syst. Biol.* **10**, 742 (2014).
125. Rhodius, V. A. et al. Design of orthogonal genetic switches based on a crosstalk map of σs, anti-σs, and promoters. *Mol. Syst. Biol.* **9**, 702 (2013).
126. Cervettini, D. et al. Rapid discovery and evolution of orthogonal aminoacyl-tRNA synthetase–tRNA pairs. *Nat. Biotechnol.* **38**, 989–999 (2020).
127. Aleksashin, N. A. et al. A fully orthogonal system for protein synthesis in bacterial cells. *Nat. Commun.* **11**, 1858 (2020).
128. Tang, T.-C. et al. Materials design by synthetic biology. *Nat. Rev. Mater.* **6**, 332–350 (2021).
129. Chen, Y. et al. Genetic circuit design automation for yeast. *Nat. Microbiol.* **5**, 1349–1360 (2020).
130. Robertson, W. E. et al. Creating custom synthetic genomes in *Escherichia coli* with REXER and GENESIS. *Nat. Protoc.* **16**, 2345–2380 (2021).
131. Pretorius, I. & Boeke, J. Yeast 2.0 — connecting the dots in the construction of the world's first functional synthetic eukaryotic genome. *FEMS Yeast Res.* **18**, foy032 (2018).
132. Fredens, J. et al. Total synthesis of *Escherichia coli* with a recoded genome. *Nature* **569**, 514–518 (2019).
The article outlines the design and synthesis of an *E. coli* genome with a 61-codon genome and describes, in detail, refactoring strategies for overlapped genes.
133. Chan, L. Y., Kosuri, S. & Endy, D. Refactoring bacteriophage T7. *Mol. Syst. Biol.* **1**. 2005.0018. (2005).
134. Jaschke, P. R., Lieberman, E. K., Rodriguez, J., Sierra, A. & Endy, D. A fully decompressed synthetic bacteriophage ΦX174 genome assembled and archived in yeast. *Virology* **434**, 278–284 (2012).
This article describes the first fully refactored viral genome with all gene overlaps removed.
135. Gimpel, J. A., Nour-Eldin, H. H., Scranton, M. A., Li, D. & Mayfield, S. P. Refactoring the six-gene photosystem II core in the chloroplast of the green algae *Chlamydomonas reinhardtii*. *ACS Synth. Biol.* **5**, 589–596 (2016).
136. Hutchison, C. A. III et al. Design and synthesis of a minimal bacterial genome. *Science* **351**, aad6253 (2016).
137. Venetz, J. E. et al. Chemical synthesis rewriting of a bacterial genome to achieve design flexibility and biological functionality. *Proc. Natl Acad. Sci. USA* **116**, 8070 (2019).
138. Springman, R., Molineux, I. J., Duong, C., Bull, R. J. & Bull, J. J. Evolutionary stability of a refactored phage genome. *ACS Synth. Biol.* **1**, 425–430 (2012).
139. Wright, B. W., Ruan, J., Molloy, M. P. & Jaschke, P. R. Genome modularization reveals overlapped gene topology is necessary for efficient viral reproduction. *ACS Synth. Biol.* **9**, 3079–3090 (2020).
140. Temme, K., Zhao, D. & Voigt, C. A. Refactoring the nitrogen fixation gene cluster from *Klebsiella oxytoca*. *Proc. Natl Acad. Sci. USA* **109**, 7085–7090 (2012).
This article describes the first fully refactored complex gene cluster, which required disrupting multiple overlapping genes.
141. Song, M. et al. Control of type III protein secretion using a minimal genetic system. *Nat. Commun.* **8**, 14737 (2017).
142. Li, G.-Q et al. Improvement of dibenzothiophene desulfurization activity by removing the gene overlap in the *dsz* operon. *Biosci. Biotechnol. Biochem.* **71**, 849–854 (2007).
143. Ghosh, D., Kohli, A. G., Moser, F., Endy, D. & Belcher, A. M. Refactored M13 bacteriophage as a platform for tumor cell imaging and drug delivery. *ACS Synth. Biol.* **1**, 576–582 (2012).
144. Richardson, S. M. et al. Design of a synthetic yeast genome. *Science* **355**, 1040 (2017).
145. Lajoie, M. J. et al. Genomically recoded organisms expand biological functions. *Science* **342**, 357–360 (2013).
146. Baba, T. et al. Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol. Syst. Biol.* **2**. 2006.0008. (2006).
147. Wichmann, S., Scherer, S. & Ardern, Z. Computational design of genes encoding completely overlapping protein domains: Influence of genetic code and taxonomic rank. *bioRxiv* https://doi.org/10.1101/2020.09.25.312959 (2020).
148. Opuu, V., Silvert, M. & Simonson, T. Computational design of fully overlapping coding schemes for protein pairs and triplets. *Sci. Rep.* **7**, 15873 (2017).
149. Inouye, M., Ishida, Y. & Inouye, K. Designing of a single gene encoding four functional proteins. *J. Theor. Biol.* **419**, 266–268 (2017).
150. Frénoy, A., Taddei, F. & Misevic, D. Genetic architecture promotes the evolution and maintenance of cooperation. *PLoS Comput. Biol.* **9**, e1003339 (2013).
151. Bull, J. J. & Barrick, J. E. Arresting evolution. *Trends Genet.* **33**, 910–920 (2017).
152. Blazejewski, T., Ho, H.-I. & Wang, H. H. Synthetic sequence entanglement augments stability and containment of genetic information in cells. *Science* **365**, 595 (2019).
This article outlines a powerful new method to protect a gene of interest from mutation by embedding an essential gene within it.
153. Decrulle, A. L. et al. Engineering gene overlaps to sustain genetic constructs in vivo. *bioRxiv* https://doi.org/10.1101/659243 (2019).
154. Makoff, A. J. & Smallwood, A. E. The use of two-cistron constructions in improving the expression of a heterologous gene in *E.coli*. *Nucleic Acids Res.* **18**, 1711–1718 (1990).
155. Mutalik, V. K. et al. Precise and reliable gene expression via standard transcription and translation initiation elements. *Nat. Methods* **10**, 354–360 (2013).
This article describes a means through which to control translation in *E. coli* with the use of engineered translational coupling using stop–start codon overlaps.
156. Roy, V. et al. A bicistronic vector with destabilized mRNA secondary structure yields scalable higher titer expression of human neurturin in *E. coli*. *Biotechnol. Bioeng.* **114**, 1753–1761 (2017).
157. Claassens, N. J. et al. Bicistronic design-based continuous and high-level membrane protein production in *Escherichia coli*. *ACS Synth. Biol.* **8**, 1685–1690 (2019).
158. Stallmeyer, B., Drugeon, G., Reiss, J., Haenni, A. L. & Mendel, R. R. Human molybdopterin synthase gene: identification of a bicistronic transcript with overlapping reading frames. *Am. J. Hum. Genet.* **64**, 698–705 (1999).
159. Ostrov, N. et al. Design, synthesis, and testing toward a 57-codon genome. *Science* **353**, 819 (2016).
160. Calles, J., Justice, I., Brinkley, D., Garcia, A. & Endy, D. Fail-safe genetic codes designed to intrinsically contain engineered organisms. *Nucleic Acids Res.* **47**, 10439–10451 (2019).
161. Anderson, J. C. et al. An expanded genetic code with a functional quadruplet codon. *Proc. Natl Acad. Sci. USA* **101**, 7566 (2004).
162. Wang, K., Schmied, W. H. & Chin, J. W. Reprogramming the genetic code: from triplet to quadruplet codes. *Angew. Chem. Int. Ed.* **51**, 2288–2297 (2012).
163. Malyshev, D. A. et al. Efficient and sequence-independent replication of DNA containing a third base pair establishes a functional six-letter genetic alphabet. *Proc. Natl Acad. Sci. USA* **109**, 12005–12010 (2012).
164. Hoshika, S. et al. Hachimoji DNA and RNA: a genetic system with eight building blocks. *Science* **363**, 884–887 (2019).
165. Fiers, W. et al. Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature* **260**, 500–507 (1976).
166. van Heesch, S. et al. The translational landscape of the human heart. *Cell* **178**, 242–260.e29 (2019).
167. James, K., Cockell, S. J. & Zenkin, N. Deep sequencing approaches for the analysis of prokaryotic transcriptional boundaries and dynamics. *Methods* **120**, 76–84 (2017).
168. Güell, M., Yus, E., Lluch-Senar, M. & Serrano, L. Bacterial transcriptomics: what is beyond the RNA horiz-ome? *Nat. Rev. Microbiol.* **9**, 658–669 (2011).
169. Mouilleron, H., Delcourt, V. & Roucou, X. Death of a dogma: eukaryotic mRNAs can code for more than one protein. *Nucleic Acids Res.* **44**, 14–23 (2016).
170. Ho, M.-R., Tsai, K.-W. & Lin, W.-C. A unified framework of overlapping genes: towards the origination and endogenic regulation. *Genomics* **100**, 231–239 (2012).
171. Majic, P. & Payne, J. L. Enhancers facilitate the birth of de novo genes and gene integration into regulatory networks. *Mol. Biol. Evol.* **37**, 1165–1178 (2020).
172. Cordaux, R. & Batzer, M. A. The impact of retrotransposons on human genome evolution. *Nat. Rev. Genet.* **10**, 691–703 (2009).
173. Kaessmann, H., Vinckenbosch, N. & Long, M. RNA-based gene duplication: mechanistic and evolutionary insights. *Nat. Rev. Genet.* **10**, 19–31 (2009).
174. Brophy, J. A. N. & Voigt, C. A. Antisense transcription as a tool to tune gene expression. *Mol. Syst. Biol.* **12**, 854 (2016).
175. Shearwin, K. E., Callen, B. P. & Egan, J. B. Transcriptional interference–a crash course. *Trends Genet.* **21**, 339–345 (2005).
176. Pavesi, A., Magiorkinis, G. & Karlin, D. G. Viral proteins originated de novo by overprinting can be identified by codon usage: application to the "gene nursery" of Deltaretroviruses. *PLoS Comput. Biol.* **9**, e1003162 (2013).
177. Aziz, R. K. et al. The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* **9**, 75 (2008).
178. Overbeek, R. et al. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Res.* **42**, D206–D214 (2014).
179. National Center for Biotechnology Information. *NCBI Prokaryotic Genome Annotation Pipeline* https://www.ncbi.nlm.nih.gov/genome/annotation_prok/standards/ (2021).
180. Nelson, C. W., Ardern, Z. & Wei, X. OLGenie: estimating natural selection to predict functional overlapping genes. *Mol. Biol. Evol.* **37**, 2440–2449 (2020).
181. McCauley, S. & Hein, J. Using hidden Markov models and observed evolution to annotate viral genomes. *Bioinformatics* **22**, 1308–1316 (2006).
182. Pareja-Tobes, P., Manrique, M., Pareja-Tobes, E., Pareja, E. & Tobes, R. BG7: a new approach for bacterial genome annotation designed for next generation sequencing data. *PLoS One* **7**, e49239 (2012).
183. Delcher, A. L., Bratke, K. A., Powers, E. C. & Salzberg, S. L. Identifying bacterial genes and endosymbiont DNA with glimmer. *Bioinformatics* **23**, 673–679 (2007).

184. Brunet, M. A. et al. OpenProt: a more comprehensive guide to explore eukaryotic coding potential and proteomes. *Nucleic Acids Res.* **47**, D403–D410 (2019).
185. Brunet, M. A. et al. OpenProt 2021: deeper functional annotation of the coding potential of eukaryotic genomes. *Nucleic Acids Res.* **49**, D380–D388 (2020).
186. Nesvizhskii, A. I. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J. Proteom.* **73**, 2092–2123 (2010).
187. Firnberg, E., Labonte, J. W., Gray, J. J. & Ostermeier, M. A comprehensive, high-resolution map of a gene's fitness landscape. *Mol. Biol. Evol.* **31**, 1581–1592 (2014).
188. Hecht, A. et al. Measurements of translation initiation from all 64 codons in *E. coli*. *Nucleic Acids Res.* **45**, 3615–3626 (2017).
189. Berry, I. J., Steele, J. R., Padula, M. P. & Djordjevic, S. P. The application of terminomics for the identification of protein start sites and proteoforms in bacteria. *Proteomics* **16**, 257–272 (2016).
190. Willems, P. et al. N-terminal proteomics assisted profiling of the unexplored translation initiation landscape in *Arabidopsis thaliana*. *MCP* **16**, 1064–1080 (2017).
191. Vanderperre, B. et al. Direct detection of alternative open reading frames translation products in human significantly expands the proteome. *PLoS One* **8**, e70698 (2013).
192. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. & Weissman, J. S. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324**, 218–223 (2009).
This article describes the first use of Ribo-seq, which is a method to monitor translation in vivo using next-generation sequencing and has been used frequently to discover new overlapping genes.
193. Finkel, Y. et al. Comprehensive annotations of human herpesvirus 6A and 6B genomes reveal novel and conserved genomic features. *eLife* **9**, e50960 (2020).
194. Machkovech, H. M., Bloom, J. D. & Subramaniam, A. R. Comprehensive profiling of translation initiation in influenza virus infected cells. *PLoS Pathog.* **15**, e1007518 (2019).
195. Liu, X., Jiang, H., Gu, Z. & Roberts, J. W. High-resolution view of bacteriophage lambda gene expression by ribosome profiling. *Proc. Natl Acad. Sci. USA* **110**, 11928–11933 (2013).
196. Sharma, C. M. et al. The primary transcriptome of the major human pathogen *Helicobacter pylori*. *Nature* **464**, 250–255 (2010).
197. Jaschke, P. R., Dotson, G. A., Hung, K. S., Liu, D. & Endy, D. Definitive demonstration by synthesis of genome annotation completeness. *Proc. Natl Acad. Sci. USA* **116**, 24206–24213 (2019).

Acknowledgements

The authors thank R. Walker, H. Kroukamp, H. Goold, T. Williams, R. Willows, M. Maselko (Macquarie University), E. Holmes (University of Sydney), Y. Chen (UCLA), J. T. Beatty (UBC), V. Mutalik (Lawrence Berkeley National Lab), J. Chen (UT Southwestern Medical Centre), K. Y. Wei (Amgen), A. S. Khalil (Boston University), Y. Shen (BGI-Shenzhen), and J. Calles (Stanford University) for helpful discussions. P. R. J. thanks D. Endy for introducing him to gene overlaps of terrestrial origin and possibly beyond. The authors apologize to those whose work was not cited due to space limitations. The authors acknowledge funding support from Macquarie Research Excellence PhD Scholarship (BWW). P. R. J. is supported by NHMRC Ideas Grant APP1185399. M. P. M. acknowledges funding support from Bowel Cancer Research Foundation Australia.

Author contributions

B. W. W. and P. R. J. researched the literature and wrote the article. All authors provided substantial contributions to discussions of the content, and reviewed and/or edited the manuscript before submission.

Competing interests

The authors declare no competing interests.

Peer review information

Nature Reviews Genetics thanks K. Neuhaus and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.