Overlapping genes in natural and engineered genomes

Bradley W. Wright1,2, Mark P. Molloy3 and Paul R. Jaschke4✉

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.

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. 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. The compressed nature of these viral genes led to the conclusion that hidden within the genome could be other undiscovered sites of polypeptide synthesis. 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. 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, including within the SARS-CoV-2 genome. However, until recently, their true abundance and importance was overlooked outside of the realm of viral genomics 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, encompassing 26% of all protein-coding genes. 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.

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. 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 on the same or opposite strands.

**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.

![Check for updates]

**Fig. 1a**

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. **Box 1**

**Reviews**

https://doi.org/10.1038/s41576-021-00477-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. For example, the RAST pipeline uses both ab initio (GLIMMER) and sequence homology steps (SEED genome database) to annotate genomes, but markedly penalizes overlaps between predicted open reading frames (ORFs), which potentially misses vital features. Furthermore, genome annotation standards are biased against feature overlaps, especially genes “completely contained in another gene”. The solution may be custom algorithms tailored for overlap mapping that have been created specifically for viral genome annotations (for example, OLGenie) and annotation pipelines based on hidden Markov models trained on databases of experimentally confirmed overlapping genes. Some tools, such as GLImmer3 and BCG7, are more tolerant of overlapping ORFs by retaining candidate ORFs even if they overlap other predicted ORFs. New annotation databases, such as OpenProt, are being created in response to the growing realization that eukaryotic gene models need to include polycistronic transcripts with non-AUG initiation sites. 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. In addition, recent studies have shown unexpectedly strong non-AUG translation initiation, 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 use, 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.

Complementary to mass spectrometry proteomics, ribosome profiling (Ribo-Seq) is a method that involves capturing ribosomes as they decode mRNAs and sequencing the section of the transcript bound by the ribosome. 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, eukaryotic and viral genomes. 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, and a transcription start site profiling study in Helicobacter pylori identified pervasive transcription on the opposite strand of canonical genes (that is, antisense transcription).

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, as well as an attempt at proof-by-synthesis to establish the absence of any undiscovered overlapping genes.

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.

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. 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, re-arrangement of existing genes, or de novo gene and ORF creation within an existing gene.

Three directional topologies are possible (Fig. 1b). Unidirectional overlaps occur when 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, whereas the divergent and convergent overlaps are more frequent in eukaryote genomes. The way the two genes interact can be described as either overlapped, with only part of each sequence occupying the same genomic region, or nested, 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’ or ‘mother–daughter’ genes.

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, whereas in eukaryotic literature overlaps are more often considered between the primary transcript boundaries (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 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 and archaea and, on average, 27% of CDSs in these groups are involved in at least one instance of overlap. Across prokaryotes, the frequency of CDS overlap within a genome seems to be constant regardless of genome size, 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.

In prokaryotic genomes, 84% of CDS overlaps are unidirectional and produced through start codon or stop codon loss, resulting in one member of a
Genes and ORFs on the same strand. Divergent (also called head-to-head) overlaps occur at the 3′ sequence of one partner falls within the boundaries of the other. 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.

**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 in orange). (Bottom) In contrast, prokaryote and virus genes are only considered to overlap if their coding sequences overlap. 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. 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 extending 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. Start codon loss of the downstream partner and de novo start codon creation within an existing CDS (Fig. 2c) also generate unidirectional overlaps. 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])19-21. 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 partners20.

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 lengths35. 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, CTG, TCA) in any other configuration37. Divergent overlaps (Fig. 1b) do not have strong phase biases but are substantially rarer than convergent overlaps38, 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 contentious35 and genome annotation pipelines are biased against their existence39. The unusual sequence characteristics of long overlapping CDSs may have also contributed to the difficulty of their discovery, resulting in undercounting44,46. 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 overlaps60,61 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 afterwards62. 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 discovered42.

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 expression46 (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 temperatures45,46. Mechanistically, overlaps between start and stop codons of adjacent unidirectional CDSs provide additional benefits for translational coupling47-50 and ribosome re-initiation51,52 (Fig. 3a,b), in addition to benefits already provided by operons42,53. 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. Correct protein complex assembly has been revealed to be spatially regulated at the translation level for genes linked in operons, which may explain the DrrA/DrrB finding. Overlap functions such as this are likely to be prevalent for overlapped CDSs given the functional assortment of genes involved in overlap.

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 but the 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.

Most overlapping genes in eukaryotes are classified as such because their 5‘ or 3‘ UTRs overlap. 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. True exon–exon overlaps make up the minority of transcript overlap in eukaryotes, but new technologies suggest that they may be more common than currently appreciated.

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, internal gene with a single exon residing within the intron of an external gene (for example, H2BFS within HSF2BP in humans) to multiple layered ‘Russian doll-like’ nestings in Drosophila melanogaster.

Eukaryotic overlapping protein–coding genes are implicated in lineage-specific groups. For example,
the majority of vertebrate genes with overlapping transcripts are not conserved across species likely because overlapping genes tend to be young and frequently lost during evolutionary time. 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. 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.

Although rare, eukaryotes contain genes with CDS overlaps as well as overlaps that span exon–intron boundaries. A community-driven roadmap on translated ORFs has proposed that these overlapping

---

**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.

* c | Overlapping sequence regions cause mutations to affect more than one ORF, increasing fitness cost and preserving overlapped sequences under mutational pressure.

* 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.

* e | Sense–antisense gene and ORF overlap is frequently involved with gene expression regulation, including non-coding RNA and long non-coding RNA.

* f | Transcriptional tuning from convergent overlapping genes and ORFs as a result of interactions between RNA polymerase collisions (transcriptional interference).
CDSs be annotated as novel genes despite the shared locus. The recently described alt-RPL36 (bottom) overlaps the human ribosomal protein gene RPL36 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. 

**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. 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. Segmented viruses, those with the genome split into separate pieces and packaged either all
in the same capsid or in separate capsids, are more likely to contain an overlap than non-segmented viruses. The retention of overlapping CDSs in viruses has been attributed to enabling evolutionary rate reduction and increasing mutational robustness as well as being a result of capsid size limitations.

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. Studies have supported the notion that CDS overlap increases hypersensitivity to mutation (as a mutation on average would affect more than one CDS) but that genome (or population) mutational robustness is increased overall (Fig. 3c). This has been eloquently demonstrated with the overlapping rev and tat genes of the RNA virus HIV-1 (Ref. 79). 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.

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? 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 (Fig. 3d). By contrast, in RNA viruses, overlaps may primarily be a means for maintaining mutational robustness in the face of higher mutational rates as exemplified with the population fitness advantage conferred by the rev and tat overlap of HIV-1 (Ref. 79).

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. For instance, increasing the single-stranded DNA genome length of ΦX174 by >1% results in almost complete loss of infectivity. 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 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, necessitating the use of multiple vectors to deliver large human genes such as CFTR. Studies have shown a strong prevalence of overlapping CDS births in the +2 frame over the +3 frame, which is likely due to two factors: mutational bias, whereby start codons are more prevalent in the +2 reading frame relative to known CDSs, and recent evidence suggesting that the sequence of known CDSs in the +2/−2 reading frames preserves key physicochemical properties of the original sequence.

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. Furthermore, viruses are highly biased towards short overlaps, with the vast majority less than 50 nt (Ref. 84) in length, overall negatively correlated with genome length, with absolute nucleotide overlap summed across the genome rarely exceeding 1,500 nt (Ref. 75). 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, whereas a 276-nt overlap in phase 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.

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, with cis-encoded antisense RNA regulation being a major player in physiological responses. Examples of these pairings have demonstrated tight-knit regulation of expression of the protein-coding gene such as in type I toxin/antitoxin systems and in Mg2+ tolerance and virulence. 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’. Overlapping regulatory RNAs embedded within the coding sequence of bacterial genes can act in diverse regulatory roles. Evidence is also emerging that ncRNAs in prokaryotes can contain protein-coding ORFs. For more information on prokaryotic overlapping ncRNAs, we refer readers to another review.

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. Cis-NATs have regulatory functions at the RNA level and the most frequent combination is one protein-coding transcript paired with an antisense non-protein-coding transcript enabling enhanced transcriptional and post-transcriptional gene regulation. The regulatory roles of cis-NATs span major biological functions but can be generalized into protein expression regulation, splice site masking, double-stranded RNA-dependent mechanisms and chromatin remodelling. Furthermore, due to the cis-acting mechanism and shared genetic loci, the evolutionary trajectories of both genes are closely entwined. As such, interesting questions surround their evolution and acquisition.
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 promotor or by other mechanisms (FIG. 2). Recently, some overlapping ncRNA antisense transcripts have been found to also encode proteins110, further increasing the complexity and constraints of these overlapping interactions.

Many cis-NATs have been associated with human disease, including cancer progression108-111. 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 cells102. In a subpopulation of patients with α-thalassemia, 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 methylation115.

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 transcripts113. In humans, 30% of sORF-derived proteins (also called microproteins) identified by mass spectrometry were mapped internally to annotated genes114. Subsequent studies have expanded this number using a variety of methods115,116, 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 ORFs117. Investigations into the functionality of the overlapping sORFs have implicated many in human disease pathology109,110. 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 classifications118,119. More information on this developing area can be found in recent reviews120,121, including the in-motion and recent community-driven initiative to comprehensively define and catalogue these classes of non-canonical ORFs in major databases118.

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 pathways122, enzymatic activities123, orthogonal transcription124,125, and translation initiation systems126,127, and complex genetic devices128,129,130. 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. coli130 or the yeast Saccharomyces cerevisiae131), with important choices to be made during the design stage for how to deal with overlapping sequences132.

Refactoring overlapping genes. Genome refactoring is a process of reorganizing gene architecture by reformatting the underlying sequences while maintaining functionality133. 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 134] 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 host135.

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 product135-137. For example, removing CDS overlaps in the bacteriophage T7 resulted in infectious virus yet significantly reduced fitness132. 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 levels138. Similarly, a project to ‘decompress’ bacteriophage φX174 had the explicit aim to test the essentiality of CDS overlaps133. 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 cycle139.

The first complete refactoring of a complex biosynthetic cluster involving overlapping CDSs involved moving the nitrogen fixation cluster of Klebsiella oxytoca into E. coli140. 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 system141. 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 CDS141. 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 (RBs) and transcriptional regulation142,143.

Other smaller-scale refactoring projects have targeted overlapping CDSs specifically to remove engineering limitations. For example, the gene overlaps in the dlt 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.
Through rational design targeting the rate-limiting enzyme of this operon (DsxB), 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\textsuperscript{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\textsuperscript{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–start coding sequence (CDS) overlaps and two fully nested overlapping CDSs\textsuperscript{144}. Refactoring the nitrogen fixation cluster from Klebsiella oxytoca disrupted four stop–start CDS overlaps and CDS overlaps varying from 1–14 bp (Ref.\textsuperscript{140}).

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\textsuperscript{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\textsuperscript{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\(^\text{86}\). 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\(^\text{86}\).

**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\(^\text{147,148}\), and even artificial proteins\(^\text{49}\), often with the stated aim to protect the overlapping CDS from genetic drift\(^\text{146}\) in similar ways to that found in viruses\(^\text{73,151}\).

Recently, two methods for generating artificial CDS overlaps between a gene of interest and an essential gene have been described and empirically tested\(^\text{152,153}\). The Constraining Adaptive Mutations using Engineered Overlapping Sequences (CAMEOS) method\(^\text{154}\) searches for available overlaps between the CDSs of an essential gene and a gene of interest to be shielded from mutation\(^\text{152}\) (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 overlapped 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\(^\text{152}\).

The RiBoSor method\(^\text{155}\) 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\(^\text{154}\), a standardized bicistronic device architecture was recently created\(^\text{156}\). 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\(^\text{157–159}\). 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\(^\text{159}\).

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\(^\text{160,161,162}\) 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-n tRNAs\(^\text{163,164}\) will also expand the potential for synonymous codons and overlapping CDSs. New six-letter and eight-letter genetic codes\(^\text{165,166}\) 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\(^\text{126}\) 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\(^\text{167}\); 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
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\textsuperscript{6,11,12,14,16} 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\(^\text{167,168}\)).

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 as current proposals\(^\text{14}\).

Lastly, the conventional idea of the monocistrionic eukaryotic transcript is slowly being eroded\(^\text{14,169}\) with the advent of new research demonstrating transcripts harbouring multiple CDSs\(^\text{167}\). 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 overlapping gene systems for a variety of applications. Synthetic biology and genetic reactoring 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 overlapping 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
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, e0202515 (2018).

42. Krementz, M. & A. Shadow ORFs illuminated: long overlapping genes in Pseudomonas aeruginosa are translated and under purifying selection. bioRxiv 2021:10.1101/2021.02.09.456020 (2021).

43. Zehentner, B., Arzden, Z., Krementz, M., Scherer, S. & Neuhaus, K. Evidence for numerous embedded antisense genes in diverse E. coli strains. bioRxiv 2020:10.1101/2020.11.38.885249 (2020).

44. Zehentner, B., Arzden, Z., Krementz, M., Scherer, S. & Neuhaus, K. A novel pH-regulated, unusual 605 bp overlapping protein coding gene pop is encoded antisense to omph in Escherichia coli O157:H7 (EHEC). Microbiol. 11, 577 (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., Ponida, A., Podder, S. & Ghosh, T. C. Overlapping genes: a new strategy of thermophilic stress tolerance in prokaryotes. Extremophiles 19, 545–554 (2015).

47. Pradhan, P. L., W. Kaur, P. Translational coupling controls expression and function of the DrAB drug efflux pump. J. Mol. Biol. 385, 831–842 (2009).

48. Hula, F., M. Long terminal control region 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, 9353–9360 (1989).

50. Das, A. & Yanofsky, C. A ribosome binding site sequence is necessary for efficient expression of the distal gene of a transcriptionally coupled gene pair. Nucleic acids Res. 12, 4757–4768 (1984).

51. Price, M. N., Huang, K. H., Arkin, A. P. & Alm, E. J. Transcriptome understanding: From overlapping genes to co-regulated small RNA networks. PLoS One 22, 6684–6692 (2010).

52. Budnev, M. et al. RNAseq analysis of small RNA populations reveals common transcriptional strategies in the viral world. BioRxiv 2016:10.1101/10.1101/2016.06.14.647505 (2016).

53. Simon-Loriere, S., F. Huard, and N. Cernadas, et al. The effect of gene overlapping on the rate of virus RNA evolution. Mol. Biol. Evol. 30, 1916–1928 (2013).

54. Kravchenko, V. G. et al. Antisense RNA: A new paradigm for the evolution of gene expression and regulation. Proc. Natl Acad. Sci. USA 99, 1405–1409 (2002).

55. Schar, C. The造RNA network in tumour suppression. Nat. Rev. Mol. Cell Biol. 2, 731–737 (2001).

56. Brunet, M. A. et al. The FUS gene is dual-coding with an antisense RNA pop. EMBO Rep. 22, e0202513 (2020).

57. Makałowska, I., Lin, C.-F. & Hernandez, K. Birth and death of gene overlaps in vertebrates. EMBO Rep. 22, 382–388 (2020).

58. Li, H., Havens, W. M., Nibert, M. L. & Ghabrial, S. A. The FUS gene is dual-coding with an antisense RNA pop. EMBO Rep. 22, e0202513 (2020).

59. Wu, Z., Yang, H. & Colosi, P. Effect of genome size on the distribution of overlapping genes by discriminant analysis. Virology 364, 508–517 (2007).

60. Williams, B. A., Slamovits, C. H., Patron, N. J., Price, M. N., Huang, K. H., Arkin, A. P. & Alm, E. J. The birth and death of gene overlaps in vertebrates. EMBO Rep. 22, 382–388 (2020).

61. Wu, P. et al. Emerging role of tumor-related functional peptides encoded by IncRNA and circRNA. Mol. Cancer 19, 22 (2020).

62. Reis, E. M. et al. Antisense intrinsic non-coding RNA levels correlate to the degree of tumor differentiation in prostate cancer. Oncogene 23, 6688–6692 (2004).
109. Yin, J. et al. UXTAS1- 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 meaningfull co-deletion due to juxtaposition to CDKN2A. Oncogene 37, 128–158 (2018).
111. Yu, W. & Yang. Epigenetic targeting of tumour suppressor gene p515 by its antisense RNA. Nature 451, 202–206 (2008).
112. Tubbs, C. J. 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. Jacobsen, E. S. The translation of non-canonical open reading frames controls mucosal immunity. Nature 564, 435–438 (2018).
114. Slama, S. A. et al. Epistatic 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. Bazzoni, 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, e101920 (2020).
118. Lee, C. et al. A non-coding and non-coding-roles of MOCC1 (C150RF48) coordinate to regulate host inflammation and immunity. Nat. Commun. 12, 21350 (2021).
119. Nam, J.-W., Choi, S.-W. & You, B.-H. Incredibly RNA dual functions of coding and non-coding. Mol. Cell 39, 365–376 (2016).
120. Schlesinger, D. & Elsässer, S. J. Revisiting sORFs: functional microproteins. EMBO J. 35, 1848–1889 (2016).
121. Brody, T. J. Principles of fully overlapping reading frames. Each frame contains an expanded genetic code. RNA 22, 2288–2297 (2012).
122. Malysh, D. A. et al. Efficient and sequence-independent replication of overlapping frames in E. coli. ACS Synth. Biol. 3, 250–257 (2014).
123. Hoshika, S. et al. Hachmioc DJ and RNA: a genetic system with eight building blocks. Science 353, 884–887 (2016).
124. Fiers, W. et al. Complete nucleotide sequence of bacteriophage M25 RNA: primary and secondary structure of the replicase gene. Nature 260, 500–507 (1976).
125. van Heesch, S. et al. The translational landscape of the human heart. Cell 178, 242–260 e29 (2019).
126. James, K. et al. Docking S. J. A fully decompressed synthetic eukaryotic genome. Nat. Rev. Mater. 5, 7085–7090 (2022).
127. This article describes the first fully refactored complex gene cluster, which required disrupting multiple overlapping genes. Song, M. et al. Control of type III protein secretion using a minimal genetic system. Nat. Commun. 8, 14737 (2017).
128. Li, G.-Q et al. Improvement of dibenzothiophene structure of the replicase gene. Nature 569, 849–854 (2020).
129. Ghosh, D., Kohl, A. G., Moser, F., Endy, D. & Bélanger, A. M. RNA polymerase as a platform for tumor cell imaging and drug delivery. ACS Synth. Biol. 1, 576–582 (2012).
130. Fiers, W. et al. Design of fully overlapping coding schemes for protein pairs and triplets. Sci. Rep. 7, 15875 (2017).
131. Inouye, M., Ishida, Y. & Inouye, K. Designing of a single gene encoding multiple functional proteins. J. Theor. Biol. 419, 266–268 (2017).
132. Frénoy, A., Taddí, F. & Mislevic, D. Genetic architecture and maintenance of cooperation. PLoS Comput. Biol. 9, e1003539 (2013).
133. Bull, J. R. & Barrick, J. E. Arresting evolution. Trends Genet. 33, 910–920 (2017).
134. Blazewycz, T., Hoi, H.-I. & Wang, H. H. Synthetic sequence entropy and the capacity of genetic information in cells. Science 365, 595 (2019).
135. This article outlines the design and synthesis of an E. coli genome with a 61-codon genome and describes, in detail, refactoring strategies for overlapping genes.
136. Chen, Y. et al. Genetic circuit design automation for yeast. Nat. Microbiol. 5, 1549–1560 (2020).
137. Robertson, W. E. et al. Creating synthetic custom genomes in Escherichia coli with REXER and GENESIS. Nat. Protoc. 16, 2545–2580 (2021).
138. Proctorius 2.0 — connecting the dots in the construction of the world’s first functional synthetic eukaryotic genome. FEBS Lett. 18, 898–999 (2020).
139. Altvater, N. et al. Fully functional orthogonal systems for protein synthesis in bacterial cells. Nat. Commun. 11, 1858 (2020).
140. Tang, T.-C. et al. Materials design by synthetic biology. Nat. Rev. Mater. 6, 332–350 (2021).
141. Chen, Y. et al. Circuit design automation for yeast. Nat. Microbiol. 5, 1549–1560 (2020).
142. Robertson, W. E. et al. Creating synthetic custom genomes in Escherichia coli with REXER and GENESIS. Nat. Protoc. 16, 2545–2580 (2021).
143. Proctorius 2.0 — connecting the dots in the construction of the world’s first functional synthetic eukaryotic genome. FEBS Lett. 18, 898–999 (2020).
144. Altvater, N. et al. Fully functional orthogonal systems for protein synthesis in bacterial cells. Nat. Commun. 11, 1858 (2020).
145. Gimpel, J. A., Nour-Eldin, H. H., Schmied, W. H. & Chin, J. W. Functional microproteins. Nat. Rev. Mater. 5, 7085–7090 (2022).
146. This article describes the first fully refactored viral genome, which contains an expanded genetic code.
147. Fiers, W. et al. Complete nucleotide sequence of bacteriophage M25 RNA: primary and secondary structure of the replicase gene. Nature 260, 500–507 (1976).
148. van Heesch, S. et al. The translational landscape of the human heart. Cell 178, 242–260 e29 (2019).
149. James, K. et al. Docking S. J. A fully decompressed synthetic eukaryotic genome. Nat. Rev. Mater. 5, 7085–7090 (2022).
150. This article describes the first fully refactored viral genome, which contains an expanded genetic code.
151. Fiers, W. et al. Complete nucleotide sequence of bacteriophage M25 RNA: primary and secondary structure of the replicase gene. Nature 260, 500–507 (1976).
152. van Heesch, S. et al. The translational landscape of the human heart. Cell 178, 242–260 e29 (2019).
153. James, K. et al. Docking S. J. A fully decompressed synthetic eukaryotic genome. Nat. Rev. Mater. 5, 7085–7090 (2022).
154. This article describes the first fully refactored viral genome, which contains an expanded genetic code.
184. Brunet, M. A. et al. OpenProt: a more comprehensive guide to explore eukaryotic coding potential and proteomes. *Nucleic Acids Res.* 47, D405–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 APP1185599. 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.

© Springer Nature Limited 2021