genes

MDPI

Review

Origin, Evolution and Stability of Overlapping Genes in Viruses: A Systematic Review

Angelo Pavesi

Department of Chemistry, Life Sciences and Environmental Sustainability, University of Parma, Parco Area delle Scienze 23/A, I-43124 Parma, Italy; angelo.pavesi@unipr.it; Tel.: +39-0521905647

Abstract: During their long evolutionary history viruses generated many proteins *de novo* by a mechanism called "overprinting". Overprinting is a process in which critical nucleotide substitutions in a pre-existing gene can induce the expression of a novel protein by translation of an alternative open reading frame (ORF). Overlapping genes represent an intriguing example of adaptive conflict, because they simultaneously encode two proteins whose freedom to change is constrained by each other. However, overlapping genes are also a source of genetic novelties, as the constraints under which alternative ORFs evolve can give rise to proteins with unusual sequence properties, most importantly the potential for novel functions. Starting with the discovery of overlapping genes in phages infecting *Escherichia coli*, this review covers a range of studies dealing with detection of overlapping genes in small eukaryotic viruses (genomic length below 30 kb) and recognition of their critical role in the evolution of pathogenicity. Origin of overlapping genes, what factors favor their birth and retention, and how they manage their inherent adaptive conflict are extensively reviewed. Special attention is paid to the assembly of overlapping genes into ad hoc databases, suitable for future studies, and to the development of statistical methods for exploring viral genome sequences in search of undiscovered overlaps.

Keywords: asymmetric evolution; codon usage; *de novo* protein creation; modular evolution; multivariate statistics; negative selection: phylogenetic distribution; positive selection; prediction methods; sequence-composition features; symmetric evolution; virus evolution

check for updates

Citation: Pavesi, A. Origin, Evolution and Stability of Overlapping Genes in Viruses: A Systematic Review. *Genes* **2021**, *12*, 809. https://doi.org/10.3390/genes12060809

Academic Editors: Luigi Viggiano and Renè Massimiliano Marsano

Received: 6 May 2021
Accepted: 24 May 2021
Published: 26 May 2021

1. Introduction

Modification of existing genes, such as duplication followed by functional divergence, fusion (two adjacent genes fuse into a single gene), fission (a single gene splits into two genes), exon shuffling (rearrangement of protein modules), or horizontal gene transfer (gene exchange between unrelated species), is a common mechanism by which new genes arose during the evolution of living organisms [1–4]. However, genes can also originate *de novo* by taking place in non-coding regions, such as intergenic regions or introns [5,6].

During their long evolutionary history viruses generated many proteins *de novo* by a mechanism called "overprinting" [7]. Overprinting is a process in which critical nucleotide substitutions in a pre-existing gene can induce the expression of a novel protein by translation of an alternative open reading frame (ORF), while preserving the function of pre-existing gene [8]. It is thought that most overlapping genes evolve by this mechanism, and that consequently each overlap contains one ancestral frame and one originated *de novo* [9]. It is also believed that overprinting is a source of genetic novelties, because the *de novo* proteins, unlike the ancestral ones, usually lack any remote homologs in databases [10].

Most of new genes originated by overprinting are expressed by the sense strand. They are classified as same-strand, or parallel, overlapping genes because of transcription from the same strand of DNA. They are usually denoted as +1 overlapping genes, when the *de novo* frame is shifted one nucleotide 3′ with respect to the ancestral frame (Figure 1A), or as +2 overlapping genes when the *de novo* frame is shifted two nucleotides

3′ (Figure 1B). According to genetic code, 71.6% of substitutions in the third codon position are synonymous, compared to only 0 and 4.6% of substitutions in the second and first codon positions respectively. In overlapping genes, therefore, a nucleotide substitution that is synonymous in one frame is highly likely to be non-synonymous in the alternative frame.

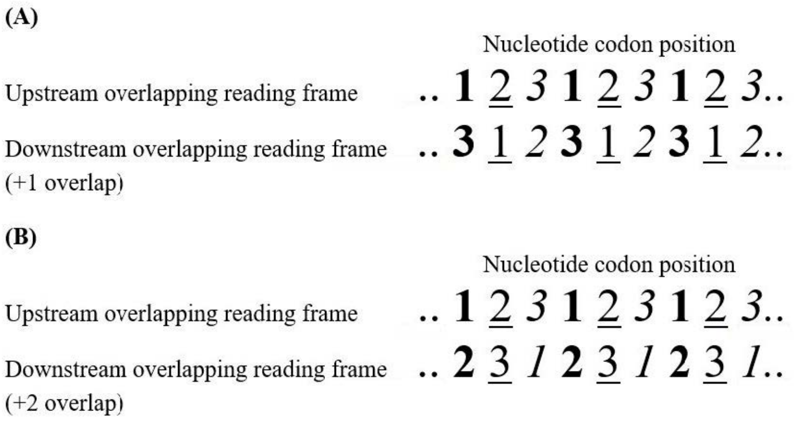
(A)
Nucleotide codon position
Upstream overlapping reading frame
.. 1 2 3 1 2 3 1 2 3..
Downstream overlapping reading frame (+1 overlap)
.. 3 1 2 3 1 2 3 1 2..
(B)
Nucleotide codon position
Upstream overlapping reading frame
.. 1 2 3 1 2 3 1 2 3..
Downstream overlapping reading frame (+2 overlap)
.. 2 3 1 2 3 1 2 3 1..

Figure 1. Orientation of same-strand overlapping genes. (**A**) Overlapping gene with the downstream frame shifted one nucleotide 3′ with respect to the upstream frame. There are 3 types of codon position (cp): cp13 (bold character), in which the first codon position of the upstream frame overlaps the third codon position of the downstream frame; cp21 (underlined character), in which the second codon position of the upstream frame overlaps the first codon position of the downstream frame; cp32 (italic character), in which the third codon position of the upstream frame overlaps the second codon position of the downstream frame. (**B**) Overlapping gene with the downstream frame shifted two nucleotides 3′ with respect to the upstream frame. There are 3 types of codon position (cp): cp12 (bold character), in which the first codon position of the upstream frame overlaps the second codon position of the downstream frame; cp23 (underlined character), in which the second codon position of the upstream frame overlaps the third codon position of the downstream frame; cp31 (italic character), in which the third codon position of the upstream frame overlaps the first codon position of the downstream frame. According to the genetic code, a nucleotide substitution at first codon position causes an amino acid change in 95.4% of cases, at second position in 100% of cases, and at third position in 28.4% of cases.

One of the reasons why overlapping genes have long attracted attention of researchers is that they represent an intriguing example of adaptive conflict. Indeed, they simultaneously encode two proteins whose freedom to change is constrained by each other, which would be expected to severely reduce the ability of the virus to adapt. On the other hand, the unusual constraints under which alternative ORFs evolve can give rise to proteins with unusual sequence properties, most importantly the potential for novel structural folds and mechanisms of action.

This review deals with the origin of overlapping genes, what factors favor their birth and retention, how they influence the evolution of viral genome, and how they manage their inherent adaptive conflict. The review is focused on overlapping genes from small viruses (genomic length below 30 kb), in which both members of the pair are known to be expressed during infection. Special attention is paid to the genealogy of the overlap, that is inferring which frame is ancestral and which one is de novo. Special attention is paid to the assembly of overlapping genes into ad hoc databases, suitable for future studies, and to the development of statistical methods for exploring viral genome sequences in search of undiscovered overlapping coding regions.

2. Discovery of Overlapping Genes and Evolutionary Implications

Overlapping genes, also called "dual-coding genes", were first discovered by Barrell et al. [11] in the genome of ΦX174, a small single-stranded DNA virus (5386 nt) that infects *Escherichia coli*. Analysis of the fully sequenced genome revealed that it contains, thanks to overprinting, 15% more coding ability than a co-linear relationship between nucleotide and protein sequences would suggest [12].

Genome sequence analysis of ΦX174 showed that there are two types of overlaps: in "internal overlaps" one overlapping gene is contained entirely within the other gene (e.g., gene E is nested within gene D) whereas "terminal overlaps" involve only the 3′ terminal region of one gene and the 5′ start region of another (e.g., the 3′ end of gene A overlaps the 5′ end of gene K) [12]. The strength of selection pressure acting on the gene overlap was estimated by a mathematical model, which pointed out a strong reduction of amino acid changes (at most 40 or 50%) in the overlapping genes B and D of ΦX174 [13].

By sequence analysis of the paired overlapping genes D and E, Fiddes and Godson proposed a simple method to predict the genealogy of the overlap [14]. Genealogy means to recognize which frame is ancestral and which frame is *de novo*. The authors first found that the genome of ΦX174 is rich in T nucleotides (31%) and these tend to occur at third codon position. They then found that in the region of D overlapping E (279 nt) the high incidence of T-ending codons is a feature of the frame D (39%) rather than the frame E (14%). Based on this finding, D was predicted as the ancestral gene and E as the *de novo* gene.

In addition, displacement of the high T content from the third codon position in frame D to the second codon position in frame E yields a high incidence of codons that specify leucine, one of the most hydrophobic amino acids, in frame E. The high content of leucine in protein E is mainly localized within a transmembrane domain, which induces lysis of the cell host *Escherichia coli* [15] by inhibiting biosynthesis of cell wall [16]. This finding suggests that de novo protein creation can be a significant factor in the evolution of pathogenicity.

The Fiddes's method to predict the genealogy of the overlap [14] was improved by means of a correlation analysis of the codon usage [17]. It was based on the assumption that the ancestral gene, which has co-evolved with the other viral genes over a long period of time, has a distribution of synonymous codons closer to that of the viral genome than the *de novo* gene. The codon-usage correlation analysis of ΦX174 demonstrated that E and K are *de novo* overlapping proteins and that the C-terminal region of protein A is a *de novo* overlapping extension [17].

When applied to ΦX174, α3 and G4 (the three evolutionary clades of the genus *Microvirus*, family *Microviridae*), the codon-usage correlation analysis predicted a gradual increase in the genome information content due to overprinting [18]. It predicted an ancestral genome having only single-coding genes, whose coding capacity increased over time due to the birth of novel overlapping coding regions (Figure 2). This fine evolutionary process led to the present genome, which contains two *de novo* overlapping genes (K and E) and two *de novo* overlapping extensions of genes A and C.

As said in introduction, an intriguing paradox of overlapping genes is that the biological information in the encoded proteins is strongly interdependent, yet each of the two proteins has evolved to its own well-defined function. Sander and Schultz [19] developed a mathematical model and applied it to the overlapping proteins A and B of ΦX174. The model postulated that the paradox can be explained by assuming sufficiently large degeneracy of the information content of amino acid sequences with respect to function.

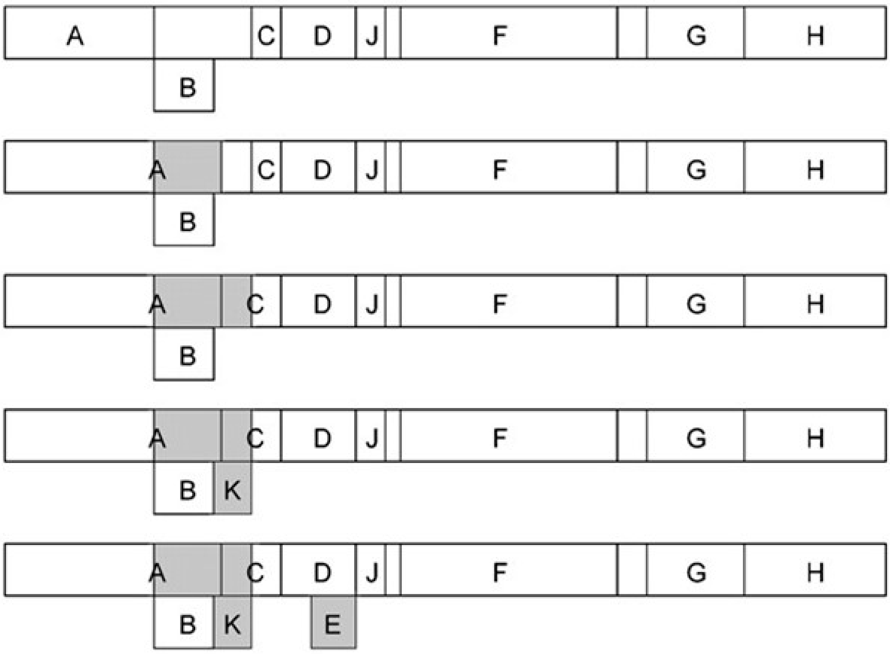

Figure 2. Increase in the genome information content during the evolution of microviruses (family *Microviridae*). The nomenclature of genes, from A to J, is that originally proposed in ΦX174 [11–13]. Empty boxes indicate ancestral pre-existing genes, while grey boxes indicate the new genes (or gene regions) that originated by overprinting. Figure reproduced from [18] with the permission of the Microbiology Society.

3. *De Novo* Overlapping Genes Show a Restricted Phylogenetic Distribution and Encode Accessory Proteins

In 1992, Keese and Gibbs published a seminal paper [9] in which the birth of new genes by overprinting was described as a continuous, and significant, evolutionary process. They proposed a new method to predict the genealogy of overlapping genes. It is based on the assumption that the protein with the most restricted phylogenetic distribution is encoded by the *de novo* frame, while that with the widest distribution is encoded by the ancestral frame.

As an example to explain the phylogenetic method, the genome of tymoviruses contains a large dual-coding region in which the 5′ one-third of replicase, encoding a methyltransferase domain, overlaps an ORF that encodes a movement protein necessary for viral spread [20]. While the methyltransferase domain has a wide phylogenetic distribution, including the closely related sister groups of potexviruses and closteroviruses or outgroups such as tricornaviruses and furoviruses, the movement protein is unique to tymoviruses (Figure 3).

Based on this finding, Keese and Gibbs inferred that replicase is the ancestral gene and that the overlapping ORF arose later, de novo, after the evolutionary divergence between tymoviruses and potexviruses. It is unlikely, indeed, that this ORF was present earlier but was subsequently lost in all virus groups with the exception of tymoviruses. It follows that the genome region of potexviruses homologous to the gene overlap unique to tymoviruses should have sequence-composition features typical of a "pre-overlapping" coding region.

Using the phylogenetic method, Rancurel et al. [21] were able to recognize the ancestral and the *de novo* frame for 17 pairs of overlapping genes, covering a wide evolutionary range of RNA viruses. Almost all de novo frames resulted to encode accessory proteins, rather than proteins central to viral replication or to the structure of capsid. "Accessory" does not mean that they are dispensable in vivo, because most novel proteins play an important role in viral pathogenicity or spread. Indeed, six de novo proteins promote a systemic diffusion of infection in plants [20,22–25], for example by binding viral RNA and forming protective ribonucleoprotein complexes [26]. Two de novo proteins contribute to evade or counteract the innate host defense, acting as inhibitor of interferon response [27] or suppressor of RNA silencing [28].

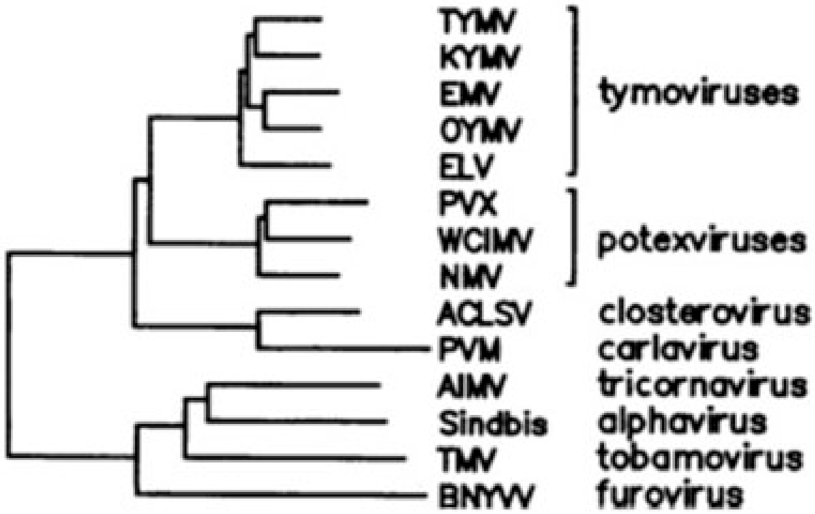

Figure 3. Dendrogram of the methyltransferase-like domain of replicase from Turnip yellow mosaic virus (TYMV), Kennedya yellow mosaic tymovirus (KYMV), Eggplant mosaic tymovirus (EMV), Ononis yellow mosaic tymovirus (OYMV), Erysimum latent tymovirus (ELV), Potato X potexvirus (PVX), White clover mosaic potexvirus (WClMV), Narcissus mosaic potexvirus (NMV), Apple chlorotic leaf spot closterovirus (ACLSV), Potato M carlavirus (PVM), Alfalfa mosaic alfamovirus (AlMV), Tobacco mosaic tobamovirus (TMV), and Beet necrotic yellow vein furovirus (BNYVV). The overlapping ORF encoding a movement protein (entirely nested within replicase) is a genetic novelty unique to tymoviruses. Figure reproduced from [9] with the permission of the authors.

The same study [21] demonstrated that most de novo proteins have a sequence composition globally biased toward disorder-promoting amino acids and that overlapping proteins are predicted to contain significantly more structural disorder than non-overlapping proteins (the term disorder applies to proteins which lack a stable secondary and tertiary structure, at least in the absence of a binding partner) (Figure 4). Based on the notion that disordered proteins are generally subjected to less structural constraint than ordered ones [29], Rancurel et al. proposed that presence of disorder in one or both overlapping proteins could relieve the evolutionary constraints imposed by the overlap.

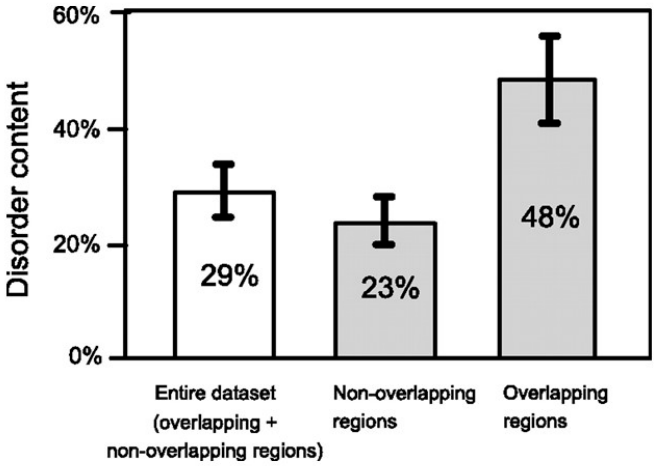

Figure 4. Predicted disorder content of proteins encoded by overlapping genes. The error bars correspond to a 95% confidence interval. Figure reproduced from [21] with the permission of the American Society of Microbiology.

This feature was further investigated by Willis and Masel [30], who analyzed a dataset of 92 overlapping genes spanning 33 viral families, 47 of them with a predicted ancestral and *de novo* frame. In accordance to [21], the authors found that the mean predicted value of the intrinsic structural disorder (ISD) in overlapping proteins is significantly higher than that in non-overlapping proteins. In addition, they found that the *de novo* proteins have a higher ISD than the ancestral ones, but this feature is specific to overlapping genes with a *de novo* frame shifted two nucleotides 3′ (+2 overlap) with respect to the ancestral frame.

The Willis study also demonstrated that the majority of overlapping genes (75%) shows a de novo frame shifted one nucleotide 3′ (+1 overlap) with respect to the ancestral frame. This feature was stronger for internal overlaps, in which one gene is completely contained within its overlapping partner, and was not found for terminal overlaps, in which the 3′ end of the upstream gene overlaps with the 5′ end of the downstream member of the pair. The prevalence of +1 gene births, despite the advantage of higher ISD in +2 gene births, was explained by the mutation bias. By sequence analysis of a control set of non-overlapping genes, Willis and Masel found that +1 frameshifts are evolutionary advantaged, because they yield significantly more ATG start codons (1 per 27 codons) than +2 frameshifts (1 per 111) and slightly fewer termination codons (1 per 14 codons) than +2 frameshifts (1 per 11).

4. Advanced Evolutionary Studies and Creation of a Curated Dataset of Overlapping Genes with Known Expression

As reported in the previous paragraph, identifying which frame of a gene overlap is ancestral and which one is de novo can be done by assessing their phylogenetic distribution (the frame phylogenetically most restricted is assumed to be the de novo one). This approach is simple and reliable but is not applicable to cases where the two frames have an identical phylogenetic distribution.

To overcome this drawback, Pavesi et al. [31] developed a new method to identify the de novo proteins. Like the previous ones [14,17,18], the method relied on the codon usage but was statistically more robust (the method assumes that the novel frame has a codon usage significantly less related to that of viral genome than the ancestral frame). It used as benchmark a reference dataset of 27 overlapping genes whose genealogy was predicted using the phylogenetic criterion. For each overlap, the method calculated: (i) the correlation coefficient (r_1) between the codon usage of the ancestral frame and that of the viral genome; (ii) the correlation coefficient (r_2) between the codon usage of the novel frame and that of the viral genome. Using the t-Hotelling test, the method evaluated the significance of the difference between r_1 and r_2, and predicted the genealogy of the overlap only in the case of r_2 significantly lower (and not simply lower) than r_1.

The method was applied to seven cases of overlap in which both frames have the same phylogenetic distribution, making the phylogenetic criterion not applicable. It demonstrated that the codon usage of overlapping frames was significantly different (or very close to significance) in only three cases: the overlap Tax protein/Rex protein of *Deltaretrovirus* and the overlap replicase/protein B2 of *Alphanodavirus* and *Betanodavirus*. Indeed, Tax and replicases had a codon usage significantly closer to that of the viral genome than the alternative frames, suggesting that they are the ancestral frames. Therefore, the *de novo* frames are those encoding the Rex protein, a post-transcription regulatory factor [32], and the protein B2, a suppressor of RNA silencing [33]. In the four other overlaps, both frames had a comparable codon usage, preventing prediction of genealogy.

The discrepancy between overlapping genes in which the novel frame has a codon usage significantly different from that of the ancestral frame and overlapping genes in which there is no significant difference was investigated by Sabath et al. [34]. They analyzed the evolution of 12 viral genes that arose de novo by overprinting and estimated their relative ages. They found that young de novo genes have a different codon usage from the rest of the genome and that evolve rapidly, under positive or weak purifying selection. In contrast, older de novo genes have a codon usage that is similar to the rest of the genome. They evolve slowly and are under strong purifying selection. Therefore, de novo genes can evolve very rapidly shortly after their origin. As they age, they tend to experience increasingly severe selective constraints, and their codon usage tends to approach that of the ancestral gene from which they originate [34].

To provide a benchmark for systematic studies, Pavesi et al. [35] assembled a high-quality dataset of 80 overlapping genes experimentally proven. They were selected from small or medium-sized eukaryotic viruses with a genome shorter than 30 kb, including single-stranded and double-stranded DNA viruses and single-stranded and double-stranded RNA viruses. The authors found that the overall nucleotide and amino acid composition of overlapping genes is significantly different from that of non-overlapping genes for several composition features. In particular, the proteins they encode show an enrichment in amino acids with high codon degeneracy (the 6-fold degenerate amino acids L, R, and S) and a depletion in amino acids with low codon degeneracy (the 2- and 1-fold degenerate amino acids C, D, E, F, H, K, N, Q, Y, M, and W), a feature that could have been selected because it mitigates the constraints under which the two frames evolve. Using a multivariate statistical method, that is the principal component analysis [36], the study demonstrated that the vast majority of overlapping genes (75 out of 80) follow a similar composition bias, despite their heterogeneity in length and function [35].

A valuable feature of the dataset is that it contains detailed biological information for each pair of overlapping genes (type of experimental evidence for expression, mechanism of translation, function of the two gene products, phenotypic effects upon mutation, and bibliography). By examining this information, Pavesi et al. [35] identified 11 overlaps in which the two encoded proteins take part in the same pathway and interact directly each other. This interaction is critical for viral assembly [37], viral replication [38], relocation of viral genome from nucleus to cytoplasm [39], and viral entry in the host cell [40].

The same study [35] pointed out that the most common mechanisms to express overlapping genes occur at the level of translation. Indeed, more than two thirds of overlapping genes with a known or suspected mechanism of expression (54 out of 71 cases) are expressed by translational processes, such as the use of an alternative start codon [41], ribosomal frameshifting [42], and internal ribosome entry site [43]. The remaining third of overlapping genes is expressed by transcriptional mechanisms, such as the use of sub-genomic RNAs [44] and transcriptional slippage [45].

5. Symmetric and Asymmetric Evolution in Viral Overlapping Genes

As first proposed by Miyata and Yasunaga [13], we would expect, in principle, that overlapping genes evolve under strong constraints, because a single nucleotide substitution can simultaneously impair two proteins (e.g., codon position 12 in Figure 1B). An example of "constrained evolution" is that observed in hepatitis B virus (HBV), a small double-stranded DNA virus (3.2 kb) with a high content of overlapping genes. Mizokami et al. [46] found that the mean number of synonymous nucleotide substitutions per site in the five overlapping coding regions of HBV is significantly lower (0.234) than that in non-overlapping regions (0.508).

However, dual-coding genes can also show a less constrained pattern of change, as a consequence of a high rate of non-synonymous substitution in one frame (positive adaptive selection) with concurrent dominance of synonymous substitution in the other (negative purifying selection). In simian immunodeficiency virus, Hughes et al. [47] found that the region of protein Tat under strongest positive selection is encoded by a frame which overlaps, for a length of 150 nt, the frame encoding protein Vpr. Another case is the overlapping gene protein p19/protein p22 (549 nt) of tombusviruses. Allison et al. [48] demonstrated that p19, a suppressor of the host RNA interference mechanism in response to viral infection [49], is under positive selection, whereas p22, a membrane-bound protein essential for cell-to-cell movement of virus [50], is under purifying selection.

These studies suggest that the evolution of overlapping genes can be summarized in accordance to two different models. The first claims that the two proteins encoded by the overlap can evolve under similar selection pressures. In the case of strong selection against amino acid change, both proteins (or protein regions) are highly conserved. For example, comparative analysis of 27 strains of HBV showed that the RNase domain of polymerase and the N-terminal half of protein X have both a percentage of conserved amino acids higher than 90% [46]. In the case of weak selection against amino acid change, both proteins can vary considerably. For example, the same study [46] showed that the spacer domain of polymerase and the pre-S1 region of surface protein show a percentage of conserved amino acids of 30 and 40%, respectively. This model was named "symmetric evolution", because the number of amino acid substitutions of one protein is expected to be not significantly different from that of the other [51]. It corresponds to the "shared model" described by Fernandes et al. [52].

The other model claims that the two proteins encoded by the overlap can evolve under significantly different selection pressures. Support for this model, which implies positive selection on one frame and negative selection on the other, was provided by a number of studies. In addition to those mentioned previously [47,48], they concern the overlapping gene P/C of Sendai virus [53], the overlapping genes ORF0/ORF1 and ORF3/ORF4 of potato leafroll virus [54], and the overlapping gene VP1/VP2 of human parvovirus B19 [55]. Interestingly, an accordance to this model was also found in the overlapping gene p16INK4a/p19ARF of mammals [56]. This model was named "asymmetric evolution", because the number of amino acid substitutions of one protein is expected to be significant different from that of the other [51]. It corresponds to the "segregated model" described by Fernandes et al. [52].

As most individual overlapping genes examined in [35] have at least one homolog, I assembled a dataset of 75 pairs of homologous overlaps and analyzed it to determine which of the two evolutionary models is the prevailing one [51]. The study demonstrated that half of overlaps (38 out of 75) evolve in accordance with the asymmetric model. A clear example was the overlapping gene of apple stem grooving virus (ASGV) that encodes a movement protein and a linker-region connecting the RdRp (RNA-dependent RNA polymerase) domain to the coat-protein domain. In detail, the percent amino acid diversity between the linker-region of ASGV and the homolog from citrus tatter leaf virus (39%; 125 differences and 195 identities) resulted to be ten-fold higher than that between the movement protein and the homolog (4%; 13 differences and 307 identities).

The same study [51] pointed out that in all overlapping genes evolving asymmetrically and with known genealogy (23 cases) the most variable protein is that encoded by the de novo frame. Despite the small number of cases, this finding suggests that de novo proteins are the preferred target of selection. As shown in Table 1, most of de novo proteins (14 out of 23) are known to play a role in viral pathogenicity: six act as suppressor of interferon response, four as suppressor of RNA silencing, two as suppressor of interferon response and apoptosis factor, one as apoptosis factor, and one has the ability to selectively degrade the host RNA-polymerase II transcripts. Very interesting is the notion that two *de novo* proteins are known to exert functions that are not virus-specific. They are the apoptin of *Chicken anemia virus*, which induces cell death in a broad range of human tumour cell lines but not in normal cells [57,58], and the protein X of Borna disease virus, which shows protective properties against neurodegeneration in vitro and in vivo [59,60].

Table 1. List of 14 overlapping genes evolving asymmetrically and with a known function of the *de novo* protein.

Virus Species and Genome Ac. Number	Overlapping Gene (Protein Products)	Predicted *De Novo* Protein (Prediction Criterion)	Most Variable Protein (Length of Overlap)	Function [Bibliographic Reference]
Theiler's murine encephalomyelitis virus (NC_001366)	Polyprotein region encoding the leader and VP4 capsid proteins/protein L*	Protein L* (phylogeny and codon usage)	Protein L* (156 aa)	Suppressor of interferon response [61]
Hepatitis C virus (NC_004102)	Polyprotein region encoding the core protein/protein F	Protein F (codon usage)	Protein F (151 aa)	Suppressor of interferon response [62]
Puumala virus (NC_005224)	Nucleocapsid protein/non-structural protein NSs	Non-structural protein NSs (codon usage)	Non-structural protein NSs (90 aa)	Suppressor of interferon response [63]
Infectious pancreatic necrosis virus (NC_001915)	Protein VP5/polyprotein region encoding the N-half of capsid protein VP2	Protein VP5 (phylogeny and codon usage)	Protein VP5 (131 aa)	Suppressor of interferon response [64]
Borna disease virus (NC_001607)	Protein X/phosphoprotein (P)	Protein X (codon usage)	Protein X (71 aa)	Suppressor of interferon response [65]
Infectious salmon anemia virus (NC_006497)	Protein p6/protein p7	Protein p6 (codon usage)	Protein p6 (183 aa)	Suppressor of interferon response [66]
Apple chlorotic leaf spot virus (NC_001409)	Protein p50/capsid protein	Protein p50 (phylogeny)	Protein p50 (105 aa)	Suppressor of RNA silencing [23]
Tomato bushy stunt virus (NC_001554)	Protein p19/protein p22	Protein p19 (phylogeny)	Protein p19 (172 aa)	Suppressor of RNA silencing [67]
Turnip yellow mosaic virus (NC_004063)	Protein p69/replicase (methyltransferase domain and downstream region)	Protein p69 (phylogeny and codon usage)	Protein p69 (626 aa)	Suppressor of RNA silencing [68]
East African cassava mosaic virus (NC_004674)	Protein AC1/protein AC4	Protein AC4 (phylogeny)	Protein AC4 (77 aa)	Suppressor of RNA silencing [69]
Murine norovirus (NC_008311)	Capsid protein VP1/virulence factor VF1	Virulence factor VF1 (phylogeny and codon usage)	Virulence factor VF1 (213 aa)	Suppressor of interferon response and apoptosis factor [70]
Influenza A virus (NC_002021)	Subunit PB1 of RdRp/protein PB1-F2	Protein PB1-F2 (phylogeny and codon usage)	Protein PB1-F2 (87 aa)	Suppressor of interferon response and apoptosis factor [71,72]
Chicken anemia virus (NC_001427)	Capsid protein VP4/apoptin	Apoptin (phylogeny)	Apoptin (119 aa)	Apoptosis factor [73]
Influenza A virus (NC_002022)	Subunit PA of RdRp/protein PA-X	Protein PA-X (codon usage)	Protein PA-X (61 aa)	Degradation of the host RNA-polymerase II transcripts [74]

Symmetric evolution (similar selection pressures on the two proteins) was found in the remaining 37 overlaps of the dataset [51]. A strong selection against amino acid change was found in the overlapping gene protein 3a/protein 3b of human severe acute respiratory syndrome-related coronavirus (SARS-CoV): the amino acid diversity between protein 3a of human SARS-CoV and the homolog from bat SARS-CoV was rather low (5.3%), as well as that between protein 3b and the homolog (8.8%). A weak selection against amino acid change was found in the overlapping gene of spinach latent virus (SLV) encoding the zinc-finger domain of polymerase and protein 2b: the amino acid diversity between the zinc-finger domain of SLV and the homolog from elm mottle virus was high (47%), as well as that between protein 2b and the homolog (44%).

6. Overlapping Genes Show a Peculiar Pattern of Nucleotide and Amino Acid Composition

Overlapping genes represent an unusual pattern of the genetic language [75,76], as two, or exceptionally three, reading frames may lie inside a single nucleotide sequence. The first attempts to detect composition features peculiar to the overlap were carried out using the information theory indices [77]. They are D_1, the divergence from a random nucleotide composition, and D_2, the divergence from a random dinucleotide composition [78,79]. The assumption is that the smallness of D_1, which implies a frequency of each nucleotide around 25%, represents the richness of vocabulary, while the largeness of D_2 represents the clarity of grammatical rules, that is the constraints against a random dinucleotide composition [80]. Thus, information theory predicts that dual-coding genes should have a lower D_1 value and a higher D_2 value when compared to single-coding genes, as hallmarks of a greater information content.

However, comparative analysis of overlapping and non-overlapping genes in the genome of three microviruses (ΦX174, α3 and G4), two avian hepadnaviruses, three strains of HIV-1, two plant luteoviruses, and two plant tymoviruses showed that the pattern predicted by information theory is valid for the first three groups of viruses, but weak for luteoviruses and inconsistent for tymoviruses [17].

In the following years, comparative analyses of overlapping and non-overlapping genes were limited to individual virus species, such as *Infectious bursal disease virus* [81], to virus families such as *Papillomaviridae* [82], or to a small dataset of RNA viruses [21]. Only recently, it was possible to perform a wide-scale analysis using the curated dataset assembled in [35]. It contains, indeed, not only the nucleotide sequence of 80 overlapping genes but also that of the entire complement of non-overlapping genes in the virus genome.

Pavesi et al. [35] found that overlapping genes differ significantly from non-overlapping genes for 20 composition features (Figure 5). Some of them are clearly linked. For example, the enrichment in C of overlapping genes is linked to that in dinucleotide CC, codons CCC and CCG, and proline. The depletion in A and T of overlapping genes is linked to that in amino acids with a low codon degeneracy, because they are encoded by codons rich in A and T. Depletion in T, A, and TA of overlapping genes reduces the probability of occurrence of stop codons (TGA, TAG and TAA) and thereby increases that of occurrence of long overlapping frames.

The dataset in [35] was also a valuable start point to assemble a much larger one [83]. For each overlapping gene, it included all the homologs gathered from the NCBI Viral Genome Database [84]. The size of the sample increased from 80 to 319 overlaps, coming from 244 virus species (the number of virus species is lower than that of overlaps because some viruses contain more than one overlap). Consider for example the overlapping gene replicase/movement protein of tymoviruses. The dataset in [35] contains only the overlap of turnip yellow mosaic virus (TYMV), the dataset in [51] contains the overlap of TYMV and the homolog of watercress white vein virus (nucleotide diversity of 28%), while the dataset in [83] contains as many as 20 homologous overlaps, covering a nucleotide diversity from 28 to 50%.

By comparative analysis of overlapping and non-overlapping genes (319 overlaps and 244 non-overlaps), I detected a total of 37 significantly different composition features [83]. Principal component analysis, aimed to evaluate whether the observed differences were homogeneously distributed in individual overlapping genes, revealed the presence of only four outliers (Figure 6). This finding confirmed that overlapping genes follow a common pattern of composition bias, despite their different length and function.

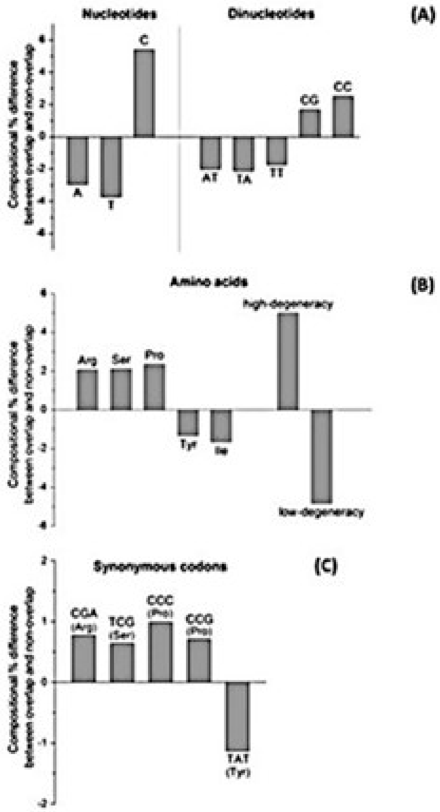

Figure 5. Difference between the pooled sets of overlapping and non-overlapping genes for the 20 most critical composition features. (**A**) Nucleotides and dinucleotides. (**B**) Amino acids and amino acids grouped in accordance to codon degeneracy. (**C**) Synonymous codons. The figure, made by A. Vianelli, was reproduced from [35].

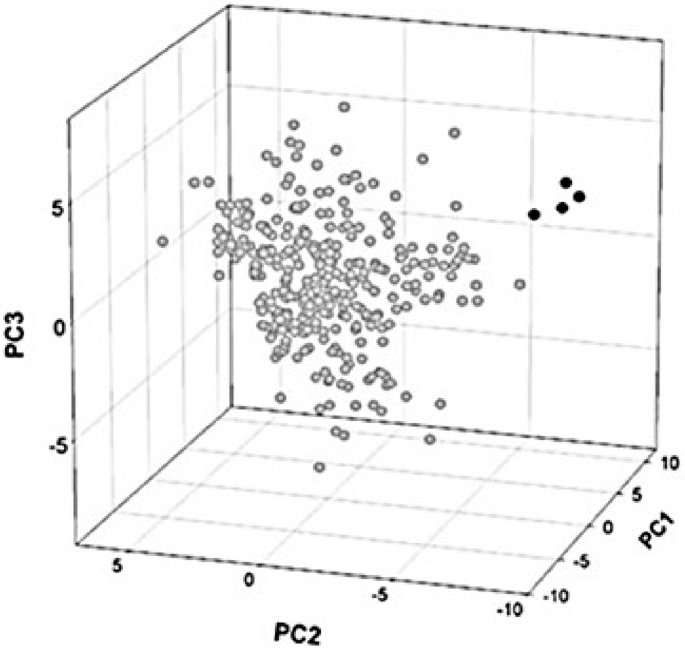

Figure 6. Principal component analysis (PCA) of a sample set of 319 overlapping genes. The three-dimensional map was obtained using the first (PC1), second (PC2), and third (PC3) principal component. Black circles indicate the 4 homologs of the overlapping gene polymerase/protein X of Hepatitis B virus. They were classified as outlier because of a highly atypical sequence composition. Figure reproduced from [83] with the permission of Elsevier.

With the aim to distinguish overlapping from non-overlapping genes with the best accuracy, I compared the sample set of 319 overlaps to the control-set of 244 non-overlaps using multivariate statistics [83]. The methods were the Fisher's linear discriminant analysis (LDA) [85,86] and the partial least squares-discriminant analysis (PLS-DA) [87,88].

The best performance of LDA was given by a linear function of 21 coefficients, corresponding to 21 significantly different composition features between overlap and non-overlap (two from nucleotides, four from dinucleotides, eight from amino acids, and seven from synonymous codons). As shown in Figure 7, the strong discriminant power of the function is highlighted by the different distribution of the LDA score in overlapping genes (grey columns) compared to that in non-overlapping genes (black columns).

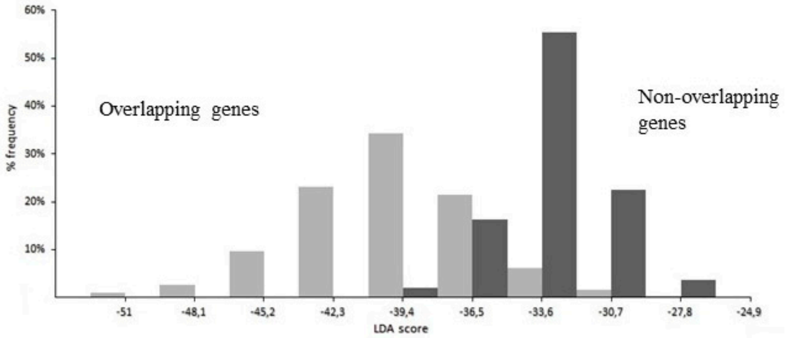

Figure 7. Histogram of the distribution of LDA score in overlapping genes (grey columns) and in non-overlapping genes (black columns). With a discriminant score of −35.31, a high percentage (96.5%) of overlapping genes were correctly classified as overlap (score below −35.31) and a high percentage (97.1%) of non-overlapping genes were correctly classified as non-overlap (score above −35.31). Figure was reproduced from [83] with the permission of Elsevier.

The best performance of PLS-DA was given by a linear regression function of 23 regression coefficients, corresponding to 23 significantly different composition features between overlap and non-overlap (one from nucleotides, six from dinucleotides, seven from amino acids, and nine from synonymous codons). The strong discriminant power of the function is evident in Figure 8, which shows the distribution of the PLS-DA score in overlapping (grey columns) and non-overlapping genes (black columns).

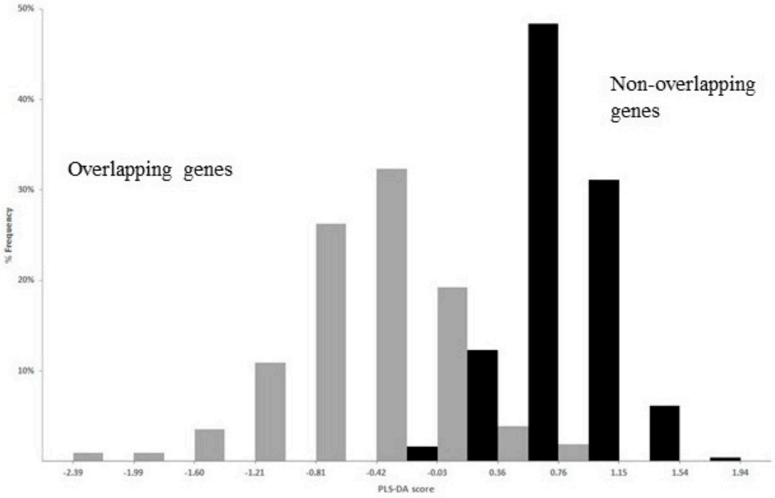

Figure 8. Histogram of the distribution of PLS-DA score in overlapping genes (grey columns) and in non-overlapping genes (black columns). With a discriminant score of 0, a high percentage of overlapping genes (94.9%) were correctly classified as overlap (score below 0) and a high percentage (98.4) of non-overlapping genes were correctly classified as non-overlap (score above 0). Figure reproduced from [83] with the permission of Elsevier.

7. Birth of Overlapping Genes in Viruses: Gene Compression or Gene Novelty?

The abundance of overlapping genes in viruses [89] was explained by two, not mutually exclusive, theories. The gene-compression theory states that the gene overlap is a valuable strategy to maximize the coding ability of small genomes [13,17,90–92], as a consequence of biophysical constraints on the size of the capsid structure [93] or of a high mutation rate such that occurring in RNA viruses [94]. As most mutations are deleterious, the high mutation rate will limit the genome size, and thus new genes must come from overprinting [95]. The gene-novelty theory claims that the birth of novel proteins by overprinting is driven by selection pressures providing the virus with a fitness advantage that lead to their fixation [9,21,96].

Using as benchmark the dataset of overlapping genes assembled in [83], I could determine which of the two theories is the most plausible one. Using the phylogenetic and codon-usage criteria, I first predicted the genealogy of 46 overlapping genes. By extending the inferred genealogy to the homologs, I then obtained a dataset of 194 overlapping genes with a known ancestral and de novo frame: 126 overlaps with a +1 de novo frame and 68 overlaps with a +2 de novo frame. Analysis of amino acid and synonymous codon composition revealed that the +1 and +2 de novo frames differ significantly from the respective ancestral frames for 25 and 23 composition features, respectively [83].

On the basis of these differences in composition, the linear discriminant analysis clearly separated the ancestral frames from the +1 de novo frames (Figure 9A), as well as the ancestral frames from the +2 de novo frames (Figure 9B). When compared to the respective ancestral proteins, the +1 de novo proteins were found enriched in hydrophobic residues and depleted in acidic residues, while the +2 de novo proteins were found enriched in basic residues and cysteine and depleted in hydrophobic residues [83]. Although one theory does not entirely exclude the other, the different amino acid composition of de novo proteins vs. the ancestral ones should better support gene-novelty than gene-compression.

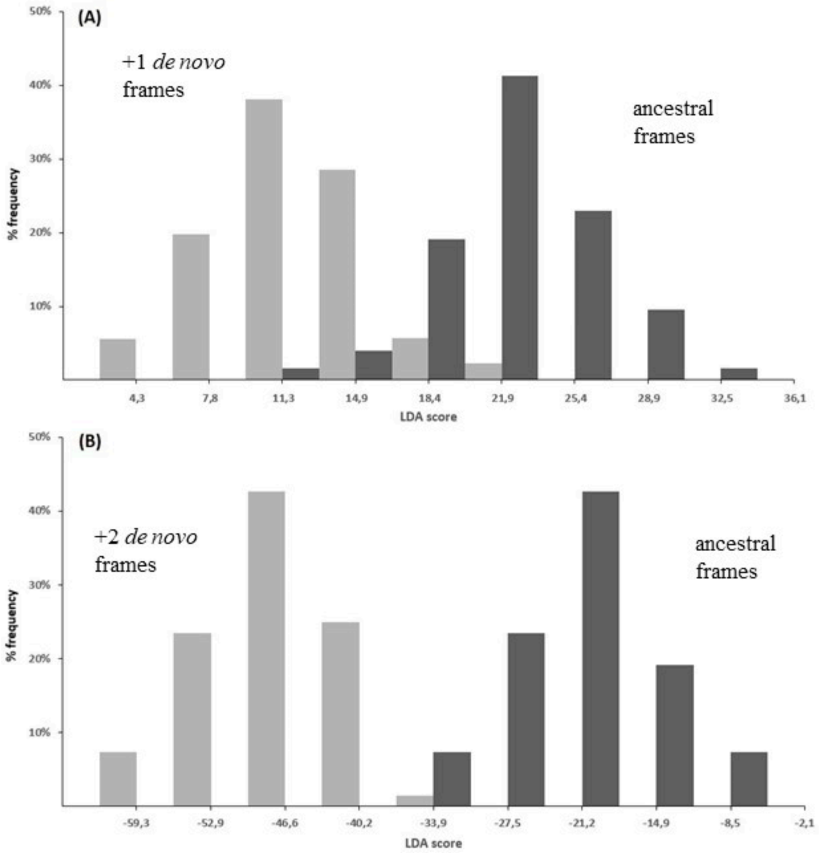

Figure 9. (**A**) Histogram of the distribution of the LDA score in 126 ancestral frames (black columns)

and in the respective +1 de novo frames (grey columns). With a discriminant score of 17.20, a high percentage (96.8%) of ancestral frames were correctly classified as ancestral (score above 17.20) and a high percentage (97.6%) of +1 de novo frames were correctly classified as de novo (score below 17.20). (**B**) Histogram of the distribution of the LDA score in 68 ancestral frames (black columns) and in the respective +2 de novo frames (grey columns). With a discriminant score of −34.98, all ancestral frames and all +2 de novo frames were correctly classified as ancestral and de novo, respectively. Figure reproduced from [83] with the permission of Elsevier.

In the same study [83], I examined the 244 virus species in the dataset to determine whether there is a negative relationship between the length of their genomes and that of their overlapping genes, a feature in accordance to the gene-compression theory. Using the Spearman rank correlation coefficient, I found a significant negative correlation of −0.31, too weak however for supporting the gene-compression theory. A similar study demonstrated that gene overlap is not a significant factor in the compression of viral genomes [96].

8. Modular Evolution in Overlapping Genes: The Case of Hepatitis B Virus

The theory of modular evolution for viruses predicts that various coding sequences are used as functional modules during recombination events [97]. This is thought to speed up virus evolution by utilizing various combinations of functional modules to gain novel genes [98,99]. However, viruses can also evolve through a mechanism in which the gain of novel modules depends on overprinting. Two studies showed that modular evolution played a critical role in the genesis of the overlapping gene polymerase/surface protein of hepadnaviruses [100,101].

Hepatitis B virus (HBV), a member of the familiy *Hepadnaviridae*, is a DNA reverse-transcribing virus with a circular genome of 3.2 kb. About 50% of the genome contains overlapping coding regions, due to the large overlap between the gene for polymerase (P) and the genes for capsid (C), X, and surface (S) proteins (Figure 10).

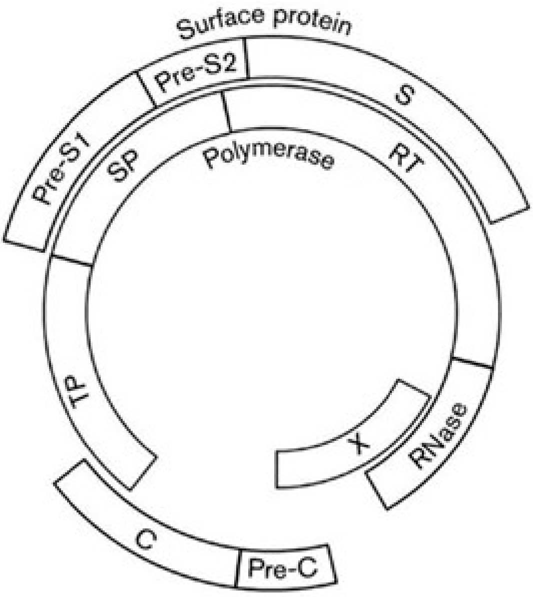

Figure 10. Map of the genome of HBV with overlapping and non-overlapping coding regions. Pre-S1, Pre-S2, and S are the domains of surface protein. TP, SP, RT, and RNase are the domains of polymerase. TP, terminal protein domain; SP, spacer domain; RT, reverse transcriptase domain; RNase, ribonuclease domain; C, capsid. Figure reproduced from [100] with the permission of the Microbiology Society.

Several studies were carried out to investigate the role of gene overlap in the evolution of HBV [46,102–105]. The genetic diversity of the overlapping proteins P and S was also related to virus survival in response to antiviral drugs [106], to virus escape from neutralizing antibodies [107], and to the clinical significance of mutations induced by selection [108].

Using the phylogenetic method, the genealogy of the overlap between the RNase domain of polymerase and the N-terminal half of protein X was clearly elucidated. The method predicts that protein X arose de novo, because of its presence in *Orthohepadnavirus* but not in the sister genus *Avihepadnavirus* [21,109]. In contrast, the genealogy of the overlap between the surface protein and the spacer (SP) and reverse-transcriptase (RT) domains of polymerase was difficult to predict. In this case, the phylogenetic criterion was not applicable because the homologs of both frames show an identical phylogenetic distribution, making possible only the codon-usage approach.

By a sliding-window analysis of the codon usage along the entire overlapping coding region (1200 nt), I found that the overlap P/S can be subdivided into two regions, each with its own pattern of codon usage [100]. By predicting the ancestral and the de novo frame in each region, I hypothesized a primordial genome with a short gene S placed between the gene encoding the terminal protein (TP) and the gene encoding the RT and RNase domains of polymerase (Figure 11A). A first increase in coding density was due to the birth, within gene S, of a de novo frame encoding the spacer (SP) domain (Figure 11B). Acting as linker, it led to creation of a multi-domain polymerase (TP, SP, RT, and RNase domains).

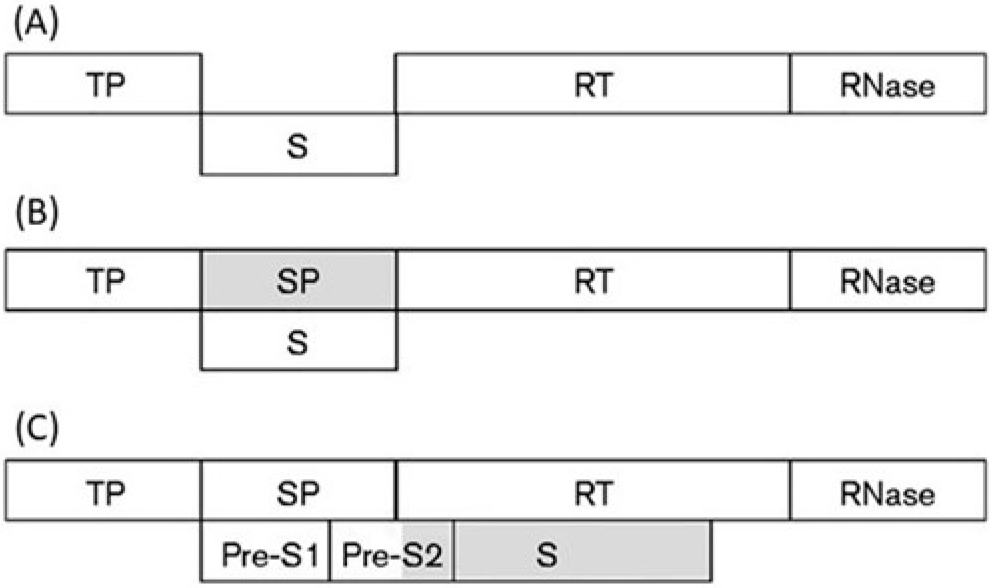

Figure 11. Modular evolution in the genesis of the overlapping gene polymerase/surface protein of hepadnaviruses. (**A**) Putative primordial genome of HBV. (**B**) Birth of a novel frame encoding the SP domain of polymerase (shaded box). (**C**) Birth of a novel frame encoding the C-terminal region of the Pre-S2 domain and the S domain of surface protein (shaded box). Figure reproduced from [100] with the permission of the Microbiology Society.

A further increase in coding density was due to a long overlapping extension of gene S. In addition to a full-length Pre-S2 domain, it led to a de novo creation of the S domain of surface protein (Figure 11C). As a result, this overlapping extension generated a surface gene consisting of three in-phase ORFs, whose co-translation yields the large surface protein. Taken together, these evolutionary inferences suggest that the overlapping gene polymerase/surface protein attained its present complexity through modular evolution [100].

The hypothesis that the Pre-S/S ORF is an innovation unique to the hepadnaviral lineage was confirmed by Lauber et al. [101]. In addition, they dated the de novo emergence of Pre-S/S about 400 million years ago. This date corresponds to the inferred separation time between hepadnaviruses (enveloped viruses with a surface-protein gene) and nackednaviruses (non-enveloped fish viruses lacking a surface-protein gene). Both studies [100,101] pointed out that overprinting is a source not only of de novo accessory proteins with regulatory function [20–28], but also of de novo essential structural proteins, such as the large surface protein of hepadnaviruses.

9. Estimation of Selection Intensities in Overlapping Genes by "*Ad Hoc*" Methods

The strength of selection pressure in protein-coding genes is usually inferred by comparing the number of non-synonymous nucleotide substitutions per site (d_n) with that of synonymous nucleotide substitutions per site (d_s), with $d_n/d_s > 1$ indicative of positive selection and $d_n/d_s < 1$ of negative selection [110,111]. Extending this standard approach to overlapping genes is inappropriate, because a nucleotide substitution that is synonymous in one frame is highly likely to be non-synonymous in the alternative frame. It follows that the constraints against synonymous substitutions in a frame significantly lowers its d_s value, causing an artifactual increase of d_n/d_s and a wrong inference of positive selection if $d_n/d_s > 1$.

To overcome this problem, several researchers have developed methods for correctly estimating the strength of selection intensities in overlapping genes. The maximum-likelihood model by Hein and Støvlbæk [112] was an extension of the notion of degeneracy class of a site [111] to that of a combination of two degeneracy classes (one for each frame to which a site belongs). De Groot et al. integrated this model into a statistical alignment framework and estimated selection in the overlapping genes of HBV and HIV-2 [113]. McCauley et al. developed a Hidden Markov Model (HMM) capable of accounting for varying levels of selection along the viral genome, including those acting on overlapping ORFs [114]. When applied to a multiple alignment of HIV-2 sequences, HMM was able to make truly statistically significant statements about the nature of selection on dual-coding regions. The Markov-chain Monte Carlo model by Pedersen and Jensen [115] incorporated the constraints imposed by both of the overlapping genetic codes in an exact manner. This model, indeed, included parameters representing the degrees of selection constraints operating in the different frames.

Sabath et al. proposed a non-stationary method, similar to that of Pedersen and Jensen but with the advantage to avoid the need for computationally-expensive procedure [116]. The method was tested on the overlapping genes PB1-F2 and NS1 of influenza A virus, because they were previously reported to exhibit values of d_n/d_s remarkably higher than 1 (9.4 for PB1-F2 and 1.9 for NS1) and thus indicative of strong positive selection [117,118]. The method demonstrated that PB1-F2 and NS1 appear to be under weak negative selection, because of a d_n/d_s value of 0.50 and 0.70 respectively. Therefore, the previous estimates of selection on PB1-F2 and NS1 were wrong, because they were calculated ignoring the interdependence with the respective overlapping frames PB1 and NS2. A limitation of the Sabath's method is that it restricts the analysis to homologous overlapping genes in which the two encoded proteins have both an amino acid diversity smaller than 50% or greater than 5%.

The method developed by Wei and Zhang [119] was an extension of the standard method for protein-coding genes originally proposed by Nei and Gojobori [111]. The method first classifies each site in the reference overlapping gene into four categories (NN, NS, SN, and SS, where N stands for non-synonymous and S for synonymous), depending on the impacts of potential mutations on the two overlapping ORFs (ORF1 and ORF2). The method then classifies all nucleotide differences between the reference overlapping gene and its homolog into four categories (NN, NS, SN, and SS) and counts their numbers (M_{NN}, M_{NS}, M_{SN}, and M_{SS}, respectively). Finally, the method estimates the strength of natural selection acting on ORF1 by $\omega_1 = d_{NN}/d_{SN}$ and that acting on ORF2 by $\omega_2 = d_{NN}/d_{NS}$.

10. Computational Methods to Predict Overlapping Genes in Viruses

To identify overlapping genes by sequence analysis, several groups have developed methods that detect the atypical pattern of nucleotide substitution induced by the overlap. Firth and Brown developed a method called Maximum-Likelihood Overlapping Gene Detector (MLOGD), which was designed to detect the mutation signature of overlapping coding sequences in pairwise alignments of two sequences, under a double-coding model [120]. The same authors presented an improved version of MLOGD, whose ability to estimate the magnitude of constraints on the gene overlap yielded a sensitivity of 90% in the detection of known overlapping genes [121].

A further improvement was provided by the computational tool Synplot2 [122]. It analyzed alignments of protein-coding virus sequences to identify regions where there is a statistically significant reduction in the degree of variability at synonymous sites, a characteristic signature of overlapping functional elements such an overlapping gene or a conserved RNA structure. The same approach was followed by Sealfon et al., who developed a phylogenetic codon-model based method (FRESCo, that is Finding Regions of Excess Synonymous Constraints) for detecting virus regions with a significantly reduced synonymous variability [123]. When applied to a multiple alignment of over 2000 whole-genome sequences of HBV, FRESCo detected strong synonymous constraint elements within known regions of overlapping function (overlapping ORFs or regulatory elements).

By modifying the method in [119], Nelson et al. [124] developed a computational tool named OLGenie, where OLG means OverLapping Gene. It estimated signs of strong purifying (negative) selection in aligned sequences as hallmark of functional overlapping genes. Assessment with simulations and controls from viral genomes (58 OLGs and 176 non-OLGs) demonstrated low false-positive rates and good ability in differentiating true OLGs from non-OLGs.

Although powerful, these computational methods are necessarily constrained by the requirement for multiple sequences of sufficient diversity to reliably detect overlapping genes. Therefore, these methods are not applicable in the case of a single nucleotide sequence or sequences with a low nucleotide diversity. To overcome this drawback, Schlub et al. developed a statistical method that relies on only a single gene, or genome, nucleotide sequence [125]. The method detects candidate overlapping genes in viruses by selecting overlapping ORFs that are significantly longer than expected by chance. It consists of a codon-permutation test and a synonymous-mutation test. The limit of the method is that the sensitivity was high (90% for codon-permutation test and 95% for synonymous-mutation test) for overlapping genes longer than 300 nt, but rather low for those longer than 100 nt (65% for codon-permutation test and 71% for synonymous-mutation test).

Another prediction method that relies on single nucleotide sequences was the combined use of linear discriminant analysis (LDA) and partial least squares-discriminant analysis (PLS-DA) [83]. Taken individually, LDA correctly classified 96.5% of overlapping genes and 97.1% of non-overlapping genes (Figure 7) and PLS-DA 94.9% of overlapping genes and 98.4% of non-overlapping genes (Figure 8). The performance of the combined use of LDA and PLS-DA is summarized in Figure 12. Grey circles in part A indicate the overlaps correctly classified by both methods (94.2% of the total). Black circles in part C indicate the non-overlaps correctly classified by both methods (97.1% of the total). Application of the method to the genome sequence of SARS-CoV-2 (isolate Wuhan-Hu-1), the etiological agent of current pandemic [126], led to detection of two new potential overlapping ORFs (asterisks in part A of the figure).

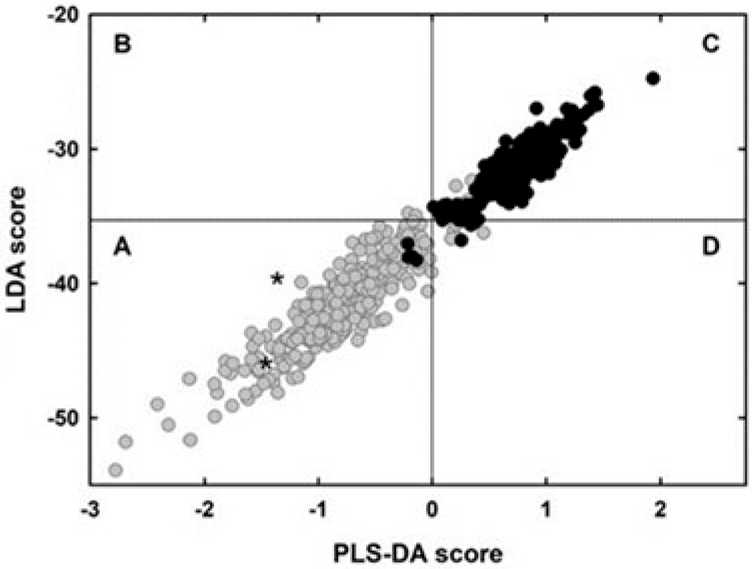

Figure 12. Map of overlapping genes (grey circles) and non-overlapping genes (black circles), in which the PLS-DA score is plotted against the respective LDA score. Grey circles in part (**A**) indicate overlaps correctly classified by both methods (94.2% of the total). Black circles in part C indicate non-overlaps correctly classified by both methods (97.1% of the total). Gray circles in part (**B**–**D**) indicate overlaps misclassified by one or both methods (5.8% of the total). Black circles in part (**A**) and (**D**) indicate non-overlaps misclassified by one or both methods (2.9% of the total). Asterisks in part (**A**) indicate two new potential overlapping genes detected in the genome of SARS-CoV-2 (isolate Wuhan-Hu-1). Figure reproduced from [83] with the permission of Elsevier.

Another method analyzing single, or closely related, genome sequences was GOFIX [127]. It detects overlapping ORFs on the basis of a significant enrichment in the X motif (a set of 20 codons over-represented in viral genes).

11. Brief Note on the Presence of Overlapping Genes in Prokaryotes and Eukaryotes

Although the present review is focused on viral overlapping genes, it is important to note that experimental and computational reports suggest that the birth of new genes by overprinting is not confined to viruses. It is a much wider phenomenon than previously thought, both in prokaryotic [128,129] and eukaryotic genomes [130–135]. Thus, the expression of two proteins from the same mRNA has changed the traditional view that a mature eukaryotic mRNA is a mono-cistronic molecule with a single translated ORF [136,137]. Interestingly, it has also been found that some human cancer-specific antigens, silent in normal tissues, are translated from alternative open reading frames (AltORFs) [138–142]. These neoantigens are promising targets for the development of anti-tumour immunotherapies with a potentially broader coverage of patients [143].

12. Brief Note on the Presence of Anti-Sense Overlapping Genes in Viruses

Overlapping genes can be classified broadly into two types: (1) same-strand overlapping genes, which are transcribed from the same strand of DNA (also known as sense-overlap); (2) different-strand overlapping genes, which are transcribed from two opposite strands of DNA (also known as anti-sense overlap).

As the great majority of known overlapping genes are of same-strand type, they were the primary focus this review. However, I would briefly report two cases of anti-sense overlap experimentally validated. The first was found in the pX region of Human T-lymphotropic virus 1 (HTLV-1). The sense strand encodes p30, a protein playing a role in viral replication, host immunity, and cellular proliferation [144]. The anti-sense strand encodes HBZ, a transcription factor playing a critical role in HTLV-1 associated diseases [145,146]. Because the pX region of HTLV-1 also contains the sense-overlap Tax protein/Rex protein, it constitutes a hotspot of gene origination, or gene "nursery" [147]. Its complex pattern of origin and evolution is accurately presented in [31].

The other anti-sense coding sequence, termed ASP and overlapping the gene Env, was predicted in HIV-1 by Cassan et al. [148]. Using computer simulations, they showed that conservation of ASP in HIV-1 (specifically in the group M) could not be due to chance but to selection pressure conserving the start codon and avoiding stop codons. Affram et al. demonstrated the presence of the ASP protein on the surfaces of both infected cells and viral particles, yielding evidence that this accessory protein is a new structural component of HIV-1 [149].

13. Concluding Remarks and Future Directions

Over four decades after the discovery of overlapping genes [11,12], we have an accurate knowledge of their origin and evolution. This review highlights that *de novo* protein creation by overprinting is a significant factor in viral evolution, in particular in the evolution of pathogenicity. At the same time, it is a valuable start point for future studies.

For example, factors affecting the birth of overlapping genes can be further investigated by a sequence-composition analysis of "pre-overlapping coding regions", that is the genome regions homologous to a gene overlap but lacking it. This analysis could assess if the composition bias is a contributing factor (i.e., a cause) to the existence of overlapping genes or a consequence of selection acting on overlapping genes after they are born.

The accuracy of multivariate statistics (LDA and PLS-DA) in determining whether a candidate overlapping ORF is coding or non-coding can be improved by comparing the sample set of overlapping genes to a control set of spurious overlapping genes, rather than of non-overlapping genes (a spurious overlapping gene is a protein-coding region that overlaps purely by chance an ORF not interrupted by stop codon).

Having found that a small set of mammalian overlapping genes follows a composition bias similar to viral one [35], a few prediction methods could be used to detect overlapping genes in eukaryotic genome sequences. They probably contain numerous undetected overlapping genes, as suggested by increasing experimental evidence [136]. Because stop codons (TGA, TAG, and TAA) are GC-poor, overlapping genes are expected to occur less frequently by chance in eukaryotic GC-rich sequences [150]. Theoretical studies focused on constraints (and their combinatorics) acting on the amino acid composition of paired overlapping proteins may form the basis for a quick and simple method to detect overlapping regions within proteins [151–153].

The computational methods reported in Section 10 are also a valuable tool to detect new potential overlapping genes in the NCBI Viral Genome Database (e.g., in large DNA viruses), to include in database proven overlaps overlooked during genome annotation, or to exclude hypothetical overlaps that may be artefacts of genome annotation.

The wide collection of proven overlapping genes and their homologs [35,83] can be used by others as reference datasets for further studies. They could expand our knowledge about their relative age, thus increasing the number of known cases of oldest and youngest *de novo* overlapping genes. They could test the occurrence of symmetric/asymmetric evolution in different regions of the same overlapping gene, as done for example in the overlap Tat protein/Rev protein of HIV-1 [52]. The relationship between gene overlap and evolutionary rate, investigated in RNA viruses [154], could be extended to DNA viruses.

A web server, called Coevolution in OVerlapped sequences by Tree analysis (COVTree), has been developed recently by Teppa et al. [155]. COVTree analyzes the effect of mutations in one protein over the other and detects coevolution signals in "mirrored" positions. It could be applied to the large dataset of homologous overlapping genes assembled in [83].

As viral protein synthesis is completely dependent upon the translational machinery of the eukaryotic host cell, studying overlapping genes has greatly improved our knowledge of gene expression. Indeed, non-canonical translational strategies such as leaky scanning, ribosomal frameshifting and alternative initiation are essential for expression of overlapping genes [41–45,156]. Therefore, detection of overlapping genes in eukaryotes may further improve our knowledge of gene expression by translational recoding [157].

Finally, the finding that a few de novo proteins have previously unknown 3D structural folds [158,159] and mechanisms of action [160] suggests that overlapping genes provide powerful model systems to test ideas about protein folding and evolution.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: The author is grateful to Alberto Vianelli (University of Insubria) for helpful suggestions. This work has benefited from the equipment and framework of the COMP-HUB initiative, funded by the 'Departments of Excellence' program of the Italian Ministry for Education, University and Research (MIUR, 2018–2022).

Conflicts of Interest: The author declares no conflict of interest.

References

1. Taylor, J.S.; Raes, J. Duplication and divergence: The evolution of new genes and old ideas. *Annu. Rev. Genet.* **2004**, *38*, 615–643. [CrossRef]
2. Long, M.; Betran, E.; Thornton, K.; Wang, W. The origin of new genes: Glimpses from the young and old. *Nat. Rev. Genet.* **2003**, *4*, 865–875. [CrossRef] [PubMed]
3. Patthy, L. Genome evolution and the evolution of exon-shuffling—A review. *Gene* **1999**, *238*, 103–114. [CrossRef]
4. Treangen, T.J.; Rocha, E.P.C. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLoS Genet.* **2011**, *7*, e1001284. [CrossRef] [PubMed]
5. Li, C.Y.; Zhang, Y.; Wang, Z.; Cao, C.; Zhang, P.W.; Lu, S.J.; Li, X.M.; Yu, Q.; Zheng, Y.; Du, Q.; et al. A human-specific de novo protein-coding gene associated with human brain functions. *PLoS Comput. Biol.* **2010**, *6*, e1000734. [CrossRef] [PubMed]
6. Sorek, R. The birth of new exons: Mechanisms and evolutionary consequences. *RNA* **2007**, *13*, 1603–1608. [CrossRef]
7. Grassé, P.P. *Evolution of Living Organisms*; Academic Press: New York, NY, USA, 1977; p. 297.
8. Normark, S.; Bergström, S.; Edlund, T.; Grundström, T.; Jaurin, B.; Lindberg, F.P.; Olsson, O. Overlapping genes. *Annu. Rev. Genet.* **1983**, *17*, 499–525. [CrossRef] [PubMed]
9. Keese, P.K.; Gibbs, A. Origin of genes: "big bang" or continuous creation? *Proc. Natl. Acad. Sci. USA* **1992**, *89*, 9489–9493. [CrossRef]
10. Gibbs, A.; Keese, P.K. *Molecular Basis of Virus Evolution*; Cambridge University Press: Cambridge, UK, 1995; pp. 76–90.
11. Barrell, B.G.; Air, G.M.; Hutchison, C.A. Overlapping genes in bacteriophage phiX174. *Nature* **1976**, *264*, 34–41. [CrossRef]
12. Sanger, F.; Air, G.M.; Barrell, B.G.; Brown, N.L.; Coulson, A.R.; Fiddes, C.A.; Hutchinson, C.A.; Slocombe, P.M.; Smith, M. Nucleotide sequence of bacteriophage phi X174. *Nature* **1977**, *265*, 687–695. [CrossRef]
13. Miyata, T.; Yasunaga, T. Evolution of overlapping genes. *Nature* **1978**, *272*, 532–535. [CrossRef]
14. Fiddes, J.C.; Godson, G.N. Evolution of the three overlapping gene systems in G4 and phi X174. *J. Mol. Biol.* **1979**, *133*, 19–43. [CrossRef]
15. Buckley, K.J.; Hayashi, M. Lytic activity localized to membrane spanning region of ΦX174 E protein. *Mol. Gen. Genet.* **1986**, *204*, 120–125. [CrossRef]
16. Bernhardt, T.G.; Roof, W.D.; Young, R. Genetic evidence that the bacteriophage phi X174 lysis protein inhibits cell wall synthesis. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 4297–4302. [CrossRef]
17. Pavesi, A.; De Iaco, B.; Granero, M.I.; Porati, A. On the informational content of overlapping genes in prokaryotic and eukaryotic viruses. *J. Mol. Evol.* **1997**, *44*, 625–631. [CrossRef]
18. Pavesi, A. Origin and evolution of overlapping genes in the family Microviridae. *J. Gen. Virol.* **2006**, *87*, 1013–1017. [CrossRef] [PubMed]
19. Sander, C.; Schultz, G.E. Degeneracy of the information contained in amino acid sequences: Evidence from overlaid genes. *J. Mol. Evol.* **1979**, *13*, 245–252. [CrossRef] [PubMed]
20. Bozarth, C.S.; Weiland, J.J.; Dreher, T.W. Expression of ORF-69 of turnip yellow mosaic virus is necessary for viral spread in plants. *Virology* **1992**, *187*, 124–130. [CrossRef]
21. Rancurel, C.; Khosravi, M.; Dunker, K.A.; Romero, P.R.; Karlin, D. Overlapping genes produce proteins with unusual sequence properties and offer insight into de novo protein creation. *J. Virol.* **2009**, *83*, 10719–10736. [CrossRef]
22. Krishnamurthy, K.; Heppler, M.; Mitra, R.; Blancaflor, E.; Payton, M.; Nelson, R.S.; Verchot-Lubicz, J. The potato virus X TGBp3 protein associates with the ER network for virus cell-to-cell movement. *Virology* **2003**, *309*, 135–151. [CrossRef]

23. Yaegashi, H.; Tamura, A.; Isogai, M.; Yoshikawa, N. Inhibition of long-distance movement of RNA silencing signals in Nicotiana benthamiana by Apple chlorotic leaf spot virus 50 kDa movement protein. *Virology* **2008**, *382*, 199–206. [CrossRef]
24. Zhou, T.; Fan, Z.F.; Li, H.F.; Wong, S.M. Hibiscus chlorotic ringspot virus p27 and its isoforms affect symptom expression and potentiate virus movement in kenaf (*Hibiscus cannabinus* L.). *Mol. Plant. Microbe Interact.* **2006**, *19*, 948–957. [CrossRef]
25. Samuilova, O.; Santala, J.; Valkonen, J.P.T. Tyrosine phosphorylation of the triple gene block protein 3 regulates cell-to-cell movement and protein interactions of Potato mop-top virus. *J. Virol.* **2013**, *87*, 4313–4321. [CrossRef]
26. Taliansky, M.; Roberts, I.M.; Kalinina, N.; Ryabov, E.V.; Raj, S.K.; Robinson, D.J.; Oparka, K.J. An umbraviral protein, involved in long-distance RNA movement, binds viral RNA and forms unique, protective ribonucleoprotein complexes. *J. Virol.* **2003**, *77*, 3031–3040. [CrossRef]
27. Skjesol, A.; Aamo, T.; Hegseth, M.N.; Robertsen, B.; Jørgensen, J.B. The interplay between infectious pancreatic necrosis virus (IPNV) and the IFN system: IFN signaling is inhibited by IPNV infection. *Virus Res.* **2009**, *143*, 53–60. [CrossRef]
28. Vargason, J.M.; Szittya, G.; Burgyan, J.; Hall, T.M. Size selective recognition of siRNA by an RNA silencing suppressor. *Cell* **2003**, *115*, 799–811. [CrossRef]
29. Brown, C.J.; Takayama, S.; Campen, A.M.; Vise, P.; Marshall, T.W.; Oldfield, C.J.; Williams, C.J.; Dunker, A.K. Evolutionary rate heterogeneity in proteins with long disordered regions. *J. Mol. Evol.* **2002**, *55*, 104–110. [CrossRef]
30. Willis, S.; Masel, J. Gene birth contributes to structural disorder encoded by overlapping genes. *Genetics* **2018**, *210*, 303–313. [CrossRef]
31. 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.* **2013**, *9*, e10031632. [CrossRef] [PubMed]
32. Hidaka, M.; Inoue, J.; Yoshida, M.; Seiki, M. Post-transcriptional regulator (rex) of HTLV-1 initiates expression of viral structural proteins but suppresses expression of regulatory proteins. *EMBO J.* **1988**, *7*, 519–523. [CrossRef]
33. Iwamoto, T.; Mise, K.; Takeda, A.; Okinaka, Y.; Mori, K.I.; Arimoto, M.; Okuno, T.; Nakai, T.J. Characterization of Striped jack nervous necrosis virus subgenomic RNA3 and biological activities of its encoded protein B2. *J. Gen. Virol.* **2005**, *86*, 2807–2816. [CrossRef]
34. Sabath, N.; Wagner, A.; Karlin, D. Evolution of viral proteins originated de novo by overprinting. *Mol. Biol. Evol.* **2012**, *29*, 3767–3780. [CrossRef]
35. Pavesi, A.; Vianelli, A.; Chirico, N.; Bao, Y.; Blinkova, O.; Belshaw, R.; Firth, A.; Karlin, D. Overlapping genes and the proteins they encode differ significantly in their sequence composition from non-overlapping genes. *PLoS ONE* **2018**, *13*, e0202513. [CrossRef]
36. Ringnér, M. What is principal component analysis? *Nat. Biotechnol.* **2008**, *26*, 303–304. [CrossRef]
37. Sun, F.; Pan, W.; Gao, H.; Qi, X.; Qin, L.; Wang, Y.; Gao, Y.; Wang, X. Identification of the interaction and interaction domains of chicken anemia virus VP2 and VP3 proteins. *Virology* **2017**, *513*, 188–194. [CrossRef]
38. Mazur, I.; Anhlan, D.; Mitzner, D.; Wixler, L.; Schubert, U.; Ludwig, S. The proapoptotic influenza A virus protein PB1-F2 regulates viral polymerase activity by interaction with the PB1 protein. *Cell Microbiol.* **2008**, *10*, 1140–1152. [CrossRef]
39. Davy, C.; McIntosh, P.; Jackson, D.J.; Sorathia, R.; Miell, M.; Wang, Q.; Khan, J.; Soneji, Y.; Doorbar, J. A novel interaction between the human papillomavirus type 16 E2 and E1-E4 proteins leads to stabilization of E2. *Virology* **2009**, *394*, 266–275. [CrossRef]
40. Wieringa, R.; de Vries, A.A.; Rottier, P.J. Formation of disulfide-linked complexes between the three minor envelope glycoproteins (GP2b, GP3, and GP4) of equine arteritis virus. *J. Virol.* **2003**, *77*, 6216–6226. [CrossRef]
41. Kobayashi, T.; Watanabe, M.; Kamitani, W.; Tomonaga, K.; Ikuta, K. Translation initiation of a bicistronic mRNA of Borna disease virus: A 16-kDa phosphoprotein is initiated at an internal start codon. *Virology* **2000**, *277*, 296–305. [CrossRef]
42. Loughran, G.; Firth, A.E.; Atkins, J.F. Ribosomal frameshifting into an overlapping gene in the 2B-encoding region of the cardiovirus genome. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, E1111-9. [CrossRef]
43. Ren, Q.; Wang, S.Q.; Firth, A.E.; Chan, M.M.Y.; Gouw, J.W.; Guarna, M.M.; Foster, L.J.; Atkins, J.F.; Jan, E. Alternative reading frame selection mediated by a tRNA-like domain of an internal ribosome entry site. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, E630-9. [CrossRef] [PubMed]
44. Ding, S.W.; Anderson, B.J.; Haase, H.R.; Svmons, R.H. New overlapping gene encoded by the cucumber mosaic virus genome. *Virology* **1994**, *198*, 593–601. [CrossRef]
45. Olspert, A.; Carr, J.P.; Firth, A.E. Mutational analysis of the Potyviridae transcriptional slippage site utilized for expression of the P3N-PIPO and P1N-PISPO proteins. *Nucleic Acids Res.* **2016**, *44*, 7618–7629. [CrossRef] [PubMed]
46. Mizokami, M.; Orito, E.; Ohba, K.; Ikeo, K.; Lau, J.Y.; Gojobori, T. Constrained evolution with respect to gene overlap of hepatitis B virus. *J. Mol. Evol.* **1997**, *44*, S83–S90. [CrossRef]
47. Hughes, A.L.; Westover, K.; da Silva, J.; O'Connor, D.H.; Watkins, D.I. Simultaneously positive and purifying selection on overlapping reading frames of the tat and vpr genes of simian immunodeficiency virus. *J. Virol.* **2001**, *75*, 7966–7972. [CrossRef]
48. Allison, J.R.; Lechner, M.; Hoeppner, M.P.; Poole, A.M. Positive selection or free to vary? Assessing the functional significance of sequence change using molecular dynamics. *PLoS ONE* **2016**, *11*, e0147619. [CrossRef]
49. Scholthof, H.B. The Tombusvirus-encoded P19: From irrelevance to elegance. *Nat. Rev. Microbiol.* **2006**, *4*, 405–411. [CrossRef]
50. Scholthof, H.B.; Scholthof, K.B.; Kikkert, M.; Jackson, A.O. Tomato bushy stunt virus spread is regulated by two nested genes that function in cell-to-cell movement and host-dependent systemic invasion. *Virology* **1995**, *213*, 425–438. [CrossRef] [PubMed]
51. Pavesi, A. Asymmetric evolution in viral overlapping genes is a source of selective protein adaptation. *Virology* **2019**, *532*, 39–47. [CrossRef]

52. Fernandes, J.D.; Faust, T.B.; Strauli, N.B.; Smith, C.; Crosby, D.C.; Nakamura, R.L.; Hernandez, R.D.; Frankel, A.D. Functional segregation of overlapping genes. *Cell* **2016**, *167*, 1762–1773. [CrossRef]
53. Fujii, Y.; Kiyotani, K.; Yoshida, T.; Sakaguchi, T. Conserved and non-conserved regions in the Sendai virus genome: Evolution of a gene possessing overlapping reading frames. *Virus Genes* **2001**, *22*, 47–52. [CrossRef] [PubMed]
54. Guyader, S.; Ducray, D.G. Sequence analysis of Potato leafroll virus isolates reveals genetic stability, major evolutionary events and differential selection pressure between overlapping reading frame products. *J. Gen. Virol.* **2002**, *83*, 1799–1807. [CrossRef]
55. Stamenković, G.G.; Ćirković, V.S.; Šiljić, M.M.; Blagojević, J.V.; Knežević, A.M.; Joksić, I.D.; Stanojević, M.P. Substitution rate and natural selection in parvovirus B19. *Sci. Rep.* **2016**, *6*, 35759. [CrossRef] [PubMed]
56. Szklarczyk, R.; Heringa, J.; Pond, S.K.; Nekrutenko, A. Rapid asymmetric evolution of a dual-coding tumor suppressor INK4a/ARF locus contradicts its function. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 12807–12812. [CrossRef]
57. Danen-Van Oorschot, A.A.; Fischer, D.F.; Grimbergen, J.M.; Klein, B.; Zhuang, S.; Falkenburg, J.H.; Backendorf, C.; Quax, P.H.; Van der Eb, A.J.; Noteborn, M.H. Apoptin induces apoptosis in human transformed and malignant cells but not in normal cells. *Proc. Natl. Acad. Sci. USA* **1997**, *94*, 5843–5847. [CrossRef]
58. Malla, W.A.; Arora, R.; Khan, R.I.N.; Mahajan, S.; Tiwari, A.K. Apoptin as a tumor-specific therapeutic agent: Current perspective on mechanism of action and delivery systems. *Front. Cell Dev. Biol.* **2020**, *8*, 524. [CrossRef]
59. Szelechowski, M.; Bétourné, A.; Monnet, Y.; Ferré, C.A.; Thouard, A.; Foret, C.; Peyrin, J.M.; Hunot, S.; Gonzalez-Dunia, D. A viral peptide that targets mitochondria protects against neuronal degeneration in models of Parkinson's disease. *Nat. Commun.* **2014**, *5*, 5181. [CrossRef]
60. Ferré, C.A.; Davezac, N.; Thouard, A.; Peyrin, J.M.; Belenguer, P.; Miquel, M.C.; Gonzalez-Dunia, D.; Szelechowski, M. Manipulation of the N-terminal sequence of the Borna disease virus X protein improves its mitochondrial targeting and neuroprotective potential. *FASEB J.* **2016**, *30*, 1523–1533. [CrossRef] [PubMed]
61. Sorgeloos, F.; Jha, B.K.; Silverman, R.H.; Michiels, T. Evasion of antiviral innate immunity by Theiler's virus L* protein through direct inhibition of RNase L. *PLoS Pathog.* **2013**, *9*, e1003474. [CrossRef]
62. Park, S.B.; Seronello, S.; Mayer, W.; Ojcius, D.M. Hepatitis C virus frameshift/alternate reading frame protein suppresses interferon responses mediated by pattern recognition receptor retinoic-acid-inducible gene-I. *PLoS ONE* **2016**, *11*, e0158419. [CrossRef]
63. Jääskeläinen, K.M.; Kaukinen, P.; Minskaya, E.S.; Plyusnina, A.; Vapalahti, O.; Elliott, R.M.; Weber, F.; Vaheri, A.; Plyusnin, A. Tula and Puumala hantavirus NSs ORFs are functional and the products inhibit activation of the interferon-beta promoter. *J. Med. Virol.* **2007**, *79*, 1527–1536. [CrossRef]
64. Lauksund, S.; Greiner-Tollersrud, L.; Chang, C.J.; Robertsen, B. Infectious pancreatic necrosis virus proteins VP2, VP3, VP4 and VP5 antagonize IFNa1 promoter activation while VP1 induces IFNa1. *Virus Res.* **2015**, *196*, 113–121. [CrossRef]
65. Wensman, J.J.; Munir, M.; Thaduri, S.; Hörnaeus, K.; Rizwan, M.; Blomström, A.L.; Briese, T.; Lipkin, W.I.; Berg, M. The X proteins of bornaviruses interfere with type I interferon signaling. *J. Gen. Virol.* **2013**, *94*, 263–269. [CrossRef]
66. García-Rosado, E.; Markussen, T.; Kileng, O.; Baekkevold, E.S.; Robertsen, B.; Mjaaland, S.; Rimstad, E. Molecular and functional characterization of two infectious salmon anaemia virus (ISAV) proteins with type I interferon antagonizing activity. *Virus Res.* **2008**, *133*, 228–238. [CrossRef] [PubMed]
67. Silhavy, D.; Molnár, A.; Lucigli, A.; Szittya, G.; Hornyik, C.; Tavazza, M.; Burgyán, J. A viral protein suppresses RNA silencing and binds silencing-generated, 21- to 25-nucleotide double-stranded RNAs. *EMBO J.* **2002**, *21*, 3070–3080. [CrossRef]
68. Chen, J.; Li, W.X.; Xie, D.; Peng, J.R.; Ding, S.W. Viral virulence protein suppresses RNA silencing-mediated defense but upregulates the role of microma in host gene expression. *Plant. Cell* **2004**, *16*, 1302–1313. [CrossRef] [PubMed]
69. Chellappan, P.; Vanitharani, R.; Fauquet, C.M. MicroRNA-binding viral protein interferes with Arabidopsis development. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 10381–10386. [CrossRef]
70. McFadden, N.; Bailey, D.; Carrara, G.; Benson, A.; Chaudhry, Y.; Shortland, A.; Heeney, J.; Yarovinsky, F.; Simmonds, P.; Macdonald, A.; et al. Norovirus regulation of the innate immune response and apoptosis occurs via the product of the alternative open reading frame 4. *PLoS Pathog.* **2011**, *7*, e1002413. [CrossRef]
71. Varga, Z.T.; Ramos, I.; Hai, R.; Schmolke, M.; García-Sastre, A.; Fernandez-Sesma, A.; Palese, P. The influenza virus protein PB1-F2 inhibits the induction of type I interferon at the level of the MAVS adaptor protein. *PLoS Pathog.* **2011**, *7*, e1002067. [CrossRef] [PubMed]
72. Chen, W.; Calvo, P.A.; Malide, D.; Gibbs, J.; Schubert, U.; Bacik, I.; Basta, S.; O'Neill, R.; Schickli, J.; Palese, P.; et al. A novel influenza A virus mitochondrial protein that induces cell death. *Nat. Med.* **2001**, *7*, 1306–1312. [CrossRef]
73. Noteborn, M.H.; Todd, D.; Verschueren, C.A.; de Gauw, H.W.; Curran, W.L.; Veldkamp, S.; Douglas, A.J.; McNulty, M.S.; van der EB, A.J.; Koch, G. A single chicken anemia virus protein induces apoptosis. *J. Virol.* **1994**, *68*, 346–351. [CrossRef]
74. Khaperskyy, D.A.; Schmaling, S.; Larkins-Ford, J.; McCormick, C.; Gaglia, M.M. Selective degradation of host RNA polymerase II transcripts by influenza A virus PA-X host shutoff protein. *PLoS Pathog.* **2016**, *12*, e1005427. [CrossRef]
75. Trifonov, E.N. Searching for Codes in the Sequences. In *Biomolecular Data. A Resource in Transition*; Oxford University Press: Oxford, UK, 1989; p. 199.
76. Smith, T.F. Semantic and Syntactic Patterns in the Genetic Language. In *Biomolecular Data. A Resource in Transition*; Oxford University Press: Oxford, UK, 1989; p. 211.
77. Shannon, C.E.; Weaver, W. *The Mathematical Theory of Communication*; University of Illinois Press: Urbana, IL, USA, 1949.

78. Granero-Porati, M.I.; Porati, A.; Zani, L. Informational parameters of an exact DNA base sequence. *J. Theor. Biol.* **1980**, *86*, 401–403. [CrossRef]
79. Smith, T.F.; Waterman, M.S. Overlapping genes and information theory. *J. Theor. Biol.* **1981**, *91*, 379–380. [CrossRef]
80. Luo, L.F.; Tsai, L.; Zhou, Y.M. Informational parameters of nucleic acid and molecular evolution. *J. Theor. Biol.* **1988**, *130*, 351–361. [CrossRef]
81. Tan, D.Y.; Bejo, M.H.; Aini, I.; Omar, A.R.; Goh, Y.M. Base usage and dinucleotide frequency of infectious bursal disease virus. *Virus Genes* **2004**, *28*, 41–53. [CrossRef]
82. Hughes, A.L.; Hughes, M.A.K. Patterns of nucleotide difference in overlapping and non-overlapping reading frames of papillomavirus genomes. *Virus Res.* **2005**, *113*, 81–88. [CrossRef] [PubMed]
83. Pavesi, A. New insights into the evolutionary features of viral overlapping genes by discriminant analysis. *Virology* **2020**, *546*, 51–66. [CrossRef] [PubMed]
84. Brister, J.R.; Ako-Adjei, D.; Bao, Y.; Blinkova, O. NCBI viral genomes resource. *Nucleic Acids Res.* **2015**, *43*, D571–D577. [CrossRef] [PubMed]
85. Fisher, R.A. The use of multiple measurements in taxonomic problems. *Ann. Eugen.* **1936**, *8*, 376–386. [CrossRef]
86. Lachenbruch, P.A.; Goldstein, M. Discriminant analysis. *Biometrics* **1979**, *35*, 69–85. [CrossRef]
87. Brereton, R.G.; Lloyd, G.R. Partial least squares discriminant analysis: Taking the magic away. *Chemometrics* **2014**, *28*, 213–225. [CrossRef]
88. Lee, L.C.; Liong, C.Y.; Jemain, A.A. Partial least squares-discriminant analysis (PLS-DA) for classification of high-dimensional (HD) data: A review of contemporary practice strategy and knowledge gaps. *Analyst* **2014**, *143*, 3526–3539. [CrossRef]
89. Schlub, T.E.; Holmes, E.C. Properties and abundance of overlapping genes in viruses. *Virus Evol.* **2020**, *6*, veaa009. [CrossRef]
90. Lamb, R.A.; Orvath, C.M. Diversity of coding strategies in influenza viruses. *Trends Genet.* **1991**, *7*, 261–266. [CrossRef]
91. Krakauer, D.C. Stability and evolution of overlapping genes. *Evolution* **2000**, *54*, 731–739. [CrossRef]
92. Peleg, O.; Kirzhner, V.; Trifonov, E.; Bolshoy, A. Overlapping messages and survivability. *J. Mol. Evol.* **2004**, *59*, 520–527. [CrossRef]
93. Chirico, N.; Vianelli, A.; Belshaw, R. Why genes overlap in viruses. *Proc. Biol. Sci.* **2010**, *277*, 3809–3817. [CrossRef] [PubMed]
94. Belshaw, R.; Pybus, O.G.; Rambaut, A. The evolution of genome compression and genomic novelty in RNA viruses. *Genome Res.* **2007**, *17*, 1496–1504. [CrossRef]
95. Holmes, E.C. *The Evolution and Emergence of RNA Viruses*; Oxford University Press: Oxford, UK, 2009.
96. Brandes, N.; Linial, M. Gene overlapping and size constraints in the viral world. *Biol. Direct.* **2016**, *11*, 26. [CrossRef]
97. Botstein, D. A theory of modular evolution for bacteriophages. *Ann. N.Y. Acad. Sci.* **1980**, *354*, 484–491. [CrossRef]
98. Gibbs, A. Molecular evolution of viruses; 'trees', 'clocks' and 'modules'. *J. Cell Sci.* **1987**, *1987* (Suppl. 7), 319–337. [CrossRef]
99. Lucchini, S.; Desiere, F.; Brüssow, H. Comparative genomics of Streptococcus thermophilus phage species supports a modular evolution theory. *J. Virol.* **1999**, *73*, 8647–8656. [CrossRef] [PubMed]
100. Pavesi, A. Different patterns of codon usage in the overlapping polymerase and surface genes of hepatitis B virus suggest a de novo origin by modular evolution. *J. Gen. Virol.* **2015**, *96*, 3577–3586. [CrossRef] [PubMed]
101. Lauber, C.; Seitz, S.; Mattei, S.; Suh, A.; Beck, J.; Herstein, J.; Börold, J.; Salzburger, W.; Kaderali, L.; Briggs, J.A.G.; et al. Deciphering the origin and evolution of hepatitis B viruses by means of a family of non-enveloped fish viruses. *Cell Host Microbe* **2017**, *22*, 387–399. [CrossRef]
102. Zaaijer, H.L.; van Hemert, F.J.; Koppelman, M.H.; Lukashov, V.V. Independent evolution of overlapping polymerase and surface protein genes of hepatitis B virus. *J. Gen. Virol.* **2007**, *88*, 2137–2143. [CrossRef] [PubMed]
103. Zhang, D.; Chen, J.; Deng, L.; Mao, Q.; Zheng, J.; Wu, J.; Zeng, C.; Li, Y. Evolutionary selection associated with the multi-function of overlapping genes in the hepatitis B virus. *Infect. Genet. Evol.* **2010**, *10*, 84–88. [CrossRef]
104. Campo, D.S.; Dimitrova, Z.; Lara, J.; Purdy, M.; Thai, H.; Ramachandran, S.; Ganova-Raeva, L.; Zhai, X.; Forbi, J.C.; Teo, C.G.; et al. Coordinate evolution of the hepatitis B virus polymerase. *Silico Biol.* **2011**, *11*, 175–182. [CrossRef]
105. Torres, C.; Fernández, M.D.; Flichman, D.M.; Campos, R.H.; Mbayed, V.A. Influence of overlapping genes on the evolution of human hepatitis B virus. *Virology* **2013**, *441*, 40–48. [CrossRef] [PubMed]
106. Moskovitz, D.N.; Osiowy, C.; Giles, E.; Tomlinson, G.; Heathcote, E.J. Response to long-term lamivudine treatment (up to 5 years) in patients with severe chronic hepatitis B, role of genotype and drug resistance. *J. Viral. Hepat.* **2005**, *12*, 398–404. [CrossRef]
107. Cooreman, M.P.; Leroux-Roels, G.; Paulij, W.P. Vaccine- and hepatitis B immune globulin-induced escape mutations of hepatitis B virus surface antigen. *J. Biomed. Sci.* **2001**, *8*, 237–247. [CrossRef]
108. Torresi, J. The virological and clinical significance of mutations in the overlapping envelope and polymerase genes of hepatitis B virus. *J. Clin. Virol.* **2002**, *25*, 97–106. [CrossRef]
109. Suh, A.; Weber, C.C.; Kehlmaier, C.; Braun, E.L.; Green, R.E.; Fritz, U.; Ray, D.A.; Ellegren, H. Early mesozoic coexistence of amniotes and hepadnaviridae. *PLoS Genet.* **2014**, *10*, e1004559. [CrossRef]
110. Gojobori, T.; Li, W.H.; Graur, D. Patterns of nucleotide substitution in pseudogenes and functional genes. *J. Mol. Evol.* **1982**, *18*, 360–369. [CrossRef]
111. Nei, M.; Gojobori, T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **1986**, *3*, 418–426. [CrossRef] [PubMed]
112. Hein, J.; Støvlbæk, J. A maximum-likelihood approach to analyzing non-overlapping and overlapping reading frames. *J. Mol. Evol.* **1995**, *40*, 181–189. [CrossRef] [PubMed]

113. de Groot, S.; Mailund, T.; Lunter, G.; Hein, J. Investigating selection on viruses: A statistical alignment approach. *BMC Bioinform.* **2008**, *9*, 304. [CrossRef] [PubMed]
114. McCauley, S.; de Groot, S.; Mailund, T.; Hein, J. Annotation of selection strengths in viral genomes. *Bioinformatics* **2007**, *23*, 2978–2986. [CrossRef] [PubMed]
115. Pedersen, A.M.; Jensen, J.L. A dependent-rates model and an MCMC-based methodology for the maximum-likelihood analysis of sequences with overlapping reading frames. *Mol. Biol. Evol.* **2001**, *18*, 763–776. [CrossRef] [PubMed]
116. Sabath, N.; Landan, G.; Graur, D. A method for the simultaneous estimation of selection intensities in overlapping genes. *PLoS ONE* **2008**, *3*, e3996. [CrossRef] [PubMed]
117. Obenauer, J.C.; Denson, J.; Mehta, P.K.; Su, X.; Mukatira, S.; Finkelstein, D.B.; Xu, X.; Wang, J.; Ma, J.; Fan, Y.; et al. Large-scale sequence analysis of avian influenza isolates. *Science* **2006**, *311*, 1576–1580. [CrossRef]
118. Pavesi, A. Pattern of nucleotide substitutions in the overlapping nonstructural genes of influenza A virus and implication for the genetic diversity of the H5N1 subtype. *Gene* **2007**, *402*, 28–34. [CrossRef]
119. Wei, X.; Zhang, J. A simple method for estimating the strength of natural selection on overlapping genes. *Genome Biol. Evol.* **2014**, *7*, 381–390. [CrossRef]
120. Firth, A.E.; Brown, C.M. Detecting overlapping coding sequences with pairwise alignments. *Bioinformatics* **2005**, *21*, 282–292. [CrossRef] [PubMed]
121. Firth, A.E.; Brown, C.M. Detecting overlapping coding sequences in virus genomes. *BMC Bioinform.* **2006**, *7*, 75. [CrossRef] [PubMed]
122. Firth, A.E. Mapping overlapping functional elements embedded within the protein-coding regions of RNA viruses. *Nucleic Acids Res.* **2014**, *42*, 12425–12439. [CrossRef] [PubMed]
123. Sealfon, R.S.; Lin, M.F.; Jungreis, I.; Wolf, M.Y.; Kellis, M.; Sabeti, P.C. FRESCo: Finding regions of excess synonymous constraint in diverse viruses. *Genome Biol.* **2015**, *16*, 38. [CrossRef] [PubMed]
124. Nelson, C.W.; Ardern, Z.; Wei, X. OLGenie: Estimating Natural Selection to predict functional overlapping genes. *Mol. Biol. Evol.* **2020**, *37*, 2440–2449. [CrossRef]
125. Schlub, T.E.; Buchmann, J.P.; Holmes, E.C. A simple method to detect candidate overlapping genes using single genome sequences. *Mol. Biol. Evol.* **2018**, *35*, 2572–2581. [CrossRef] [PubMed]
126. Zhou, P.; Yang, X.L.; Wang, X.G.; Hu, B.; Zhang, L.; Zhang, W.; Si, H.R.; Zhu, Y.; Li, B.; Huang, C.L.; et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **2020**, *579*, 265–269. [CrossRef]
127. Michel, C.J.; Mayer, C.; Poch, O.; Thompson, J.D. Characterization of accessory genes in coronavirus genomes. *Virology J.* **2020**, *17*, 131. [CrossRef]
128. Delaye, L.; Deluna, A.; Lazcano, A.; Becerra, A. The origin of a novel gene through overprinting in Escherichia coli. *BMC Evol. Biol.* **2008**, *8*, 31. [CrossRef] [PubMed]
129. Fellner, L.; Simon, S.; Scherling, C.; Witting, M.; Schober, S.; Polte, C.; Schmitt-Kopplin, P.; Keim, D.A.; Scherer, S.; Neuhaus, K. Evidence for the recent origin of a bacterial protein-coding, overlapping orphan gene by evolutionary overprinting. *BMC Evol. Biol.* **2015**, *15*, 283. [CrossRef]
130. Chung, W.Y.; Wadhawan, S.; Szklarczyk, R.; Pond, S.K.; Nekrutenko, A. A first look at ARFome: Dual-coding genes in mammalian genomes. *PLoS Comput. Biol.* **2007**, *3*, e91. [CrossRef]
131. Ribrioux, S.; Brungger, A.; Baumgarten, B.; Seuwen, K.; John, M.R. Bioinformatics prediction of overlapping frameshifted translation products in mammalian transcripts. *BMC Genom.* **2008**, *9*, 122. [CrossRef] [PubMed]
132. Michel, A.M.; Choudhury, K.R.; Firth, A.E.; Ingolia, N.T.; Atkins, J.F.; Baranov, P.V. Observation of dually decoded regions of the human genome using ribosome profiling data. *Genome Res.* **2012**, *22*, 2219–2229. [CrossRef]
133. Vanderperre, B.; Lucier, J.F.; Roucou, X. HAltORF: A database of predicted out-of-frame alternative open reading frames in human. *Database* **2012**, *2012*, bas025. [CrossRef] [PubMed]
134. Bergeron, D.; Lapointe, C.; Bissonnette, C.; Tremblay, G.; Motard, J.; Roucou, X. An out-of-frame overlapping reading frame in the ataxin-1 coding sequence encodes a novel ataxin-1 interacting protein. *J. Biol. Chem.* **2013**, *288*, 21824–21835. [CrossRef]
135. Vanderperre, B.; Lucier, J.F.; Bissonnette, C.; Motard, J.; Tremblay, G.; Vanderperre, S.; Wisztorski, M.; Salzet, M.; Boisvert, F.M.; Roucou, X. Direct detection of alternative open reading frames translation products in human significantly expands the proteome. *PLoS ONE* **2013**, *8*, e70698. [CrossRef] [PubMed]
136. Mouilleron, H.; Delcourt, V.; Roucou, X. Death of a dogma: Eukaryotic mRNAs can code for more than one protein. *Nucleic Acids Res.* **2016**, *44*, 14–23. [CrossRef]
137. Brunet, M.A.; Levesque, S.A.; Hunting, D.J.; Cohen, A.A. Recognition of the polycistronic nature of human genes is critical to understanding the genotype-phenotype relationship. *Genome Res.* **2018**, *28*, 609–624. [CrossRef] [PubMed]
138. Wang, R.F.; Parkhurst, M.R.; Kawakami, Y.; Robbins, P.F.; Rosenberg, S.A. Utilization of an alternative open reading frame of a normal gene in generating a novel human cancer antigen. *J. Exp. Med.* **1996**, *183*, 1131–1140. [CrossRef] [PubMed]
139. Wang, R.F.; Johnson, S.L.; Zeng, G.; Topalian, S.L.; Schwartzentruber, D.J.; Rosenberg, S.A. A breast and melanoma-shared tumor antigen: T cell responses to antigenic peptides translated from different open reading frames. *J. Immunol.* **1998**, *161*, 3598–3606.
140. Rosenberg, S.A.; Tong-On, P.; Li, Y.; Riley, J.P.; El-Gamil, L.; Parkhurst, M.R.; Robbins, P.F. Identification of BING-4 cancer antigen from an alternative open reading frame of a gene in the extended MHC class II region using lymphocytes from a patient with a durable complete regression following immunotherapy. *J. Immunol.* **2002**, *168*, 2402–2407. [CrossRef] [PubMed]

141. Mandic, M.; Almunia, C.; Vicel, S.; Gillet, D.; Janjic, B.; Coval, K.; Maillere, B.; Kirkwood, J.M.; Zarour, H.M. The alternative open reading frame of LAGE-1 gives rise to multiple promiscuous HLA-DR-restricted epitopes recognized by T-helper 1-type tumor-reactive CD4+ T cells. *Cancer Res.* **2003**, *63*, 6506–6515.
142. Slager, E.H.; Borghi, M.; van der Minne, C.E.; Aarnoudse, C.A.; Havenga, M.J.E.; Schrier, P.I.; Osanto, S.; Griffioen, M. CD4+ Th2 cell recognition of HLA-DR-restricted epitopes derived from CAMEL: A tumor antigen translated in an alternative open reading frame. *J. Immunol.* **2003**, *170*, 1490–1497. [CrossRef]
143. Smith, C.C.; Selitsky, S.R.; Chai, S.; Armistead, P.M.; Vincent, B.G.; Serody, J.S. Alternative tumour-specific antigens. *Nat. Rev. Canc.* **2019**, *8*, 465–478. [CrossRef]
144. Moles, R.; Sarkis, S.; Galli, V.; Omsland, M.; Purcell, D.F.J.; Yurick, D.; Khoury, G.; Pise-Masison, C.A.; Franchini, G. p30 protein: A critical regulator of HTLV-1 viral latency and host immunity. *Retrovirology* **2019**, *16*, 42. [CrossRef] [PubMed]
145. Gaudray, G.; Gachon, F.; Basbous, J.; Biard-Piechaczyk, M.; Devaux, C.; Mesnard, J.M. The complementary strand of the human T-cell leukemia virus type 1 RNA genome encodes a bZIP transcription factor that down-regulates viral transcription. *J. Virol.* **2002**, *76*, 12813–12822. [CrossRef]
146. Baratella, M.; Forlani, G.; Accolla, R.S. HTLV-1 HBZ viral protein: A key player in HTLV-1 mediated diseases. *Front. Microbiol.* **2017**, *8*, 2615. [CrossRef] [PubMed]
147. Nahon, J.L. Birth of 'human-specific' genes during primate evolution. *Genetica* **2003**, *118*, 193–208. [CrossRef]
148. Cassan, E.; Arigon-Chifolleau, A.M.; Mesnard, J.M.; Gross, A.; Gascuel, O. Concomitant emergence of the antisense protein gene of HIV-1 and of the pandemic. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 11537–11542. [CrossRef] [PubMed]
149. Affram, Y.; Zapata, J.C.; Gholizadeh, Z.; Tolbert, W.D.; Zhou, W.; Iglesias-Ussel, M.D.; Pazgier, M.; Ray, K.; Latinovic, O.S.; Romerio, F. The HIV-1 antisense protein ASP is a transmembrane protein of the cell surface and an integral protein of the viral envelope. *J. Virol.* **2019**, *93*, e00574-19. [CrossRef]
150. Oliver, J.L.; Marin, A.A. A relationship between GC content and coding-sequence length. *J. Mol. Evol.* **1996**, *43*, 216–223. [CrossRef]
151. Smith, T.F.; Waterman, M.S. Protein constraints induced by multiframe encoding. *Math. Biosci.* **1980**, *49*, 17–26. [CrossRef]
152. Pavesi, A. Detection of signature sequences in overlapping genes and prediction of a novel overlapping gene in hepatitis G virus. *J. Mol. Evol.* **2000**, *50*, 284–295. [CrossRef]
153. Lèbre, S.; Gascuel, O. The combinatorics of overlapping genes. *J. Theor. Biol.* **2017**, *415*, 90–101. [CrossRef]
154. Simon-Loriere, E.; Holmes, E.C.; Pagán, I. The effect of gene overlapping on the rate of RNA virus evolution. *Mol. Biol. Evol.* **2013**, *30*, 1916–1928. [CrossRef] [PubMed]
155. Teppa, E.; Zea, D.J.; Oteri, F.; Carbone, A. COVTree: Coevolution in Overlapped sequences by Tree analysis server. *Nucleic Acids Res.* **2020**, *48*, W558–W565. [CrossRef]
156. Firth, A.E.; Brierley, I. Non-canonical translation in RNA viruses. *J. Gen. Virol.* **2012**, *93*, 1385–1409. [CrossRef] [PubMed]
157. Dinman, J.D. Translational recoding signals: Expanding the synthetic biology toolbox. *J. Biol. Chem.* **2019**, *294*, 7537–7545. [CrossRef]
158. Meier, C.; Aricescu, A.R.; Assenberg, R.; Aplin, R.T.; Gilbert, R.J.; Grimes, J.M.; Stuart, D.I. The crystal structure of ORF-9b, a lipid binding protein from the SARS coronavirus. *Structure* **2006**, *14*, 1157–1165. [CrossRef] [PubMed]
159. Baulcombe, D.C.; Molnar, A. Crystal structure of p19—A universal suppressor of RNA silencing. *Trends Biochem. Sci.* **2004**, *29*, 279–281. [CrossRef] [PubMed]
160. Lingel, A.; Simon, B.; Izaurralde, E.; Sattler, M. The structure of the flock house virus B2 protein, a viral suppressor of RNA interference, shows a novel mode of double-stranded RNA recognition. *EMBO Rep.* **2005**, *6*, 1149–1155. [CrossRef] [PubMed]