

Supporting information captions

S1 Fig. Principal component analysis (PCA) of overlapping and non-overlapping genes.

We carried out PCA on a matrix of 160 rows (the 80 overlapping genes of our dataset and the 80 corresponding non-overlapping genes in the virus genome) and 20 columns (the 20 critical composition features). Black circles indicate the 80 overlapping genes and red circles the 80 non-overlapping genes. PC1, PC2, and PC3 account for 54.8, 18.1, and 9.7% of the total amount of variation in the source data matrix, respectively. (A) Map yielded by the first (PC1) and second (PC2) principal component. (B) Map yielded by the first (PC1) and third (PC3) principal component.

S1 Table. List of the 80 viral proven overlapping genes assembled in S1 Dataset. The genes are grouped in 7 tables (from S1a to S1g) in accordance to the nature of the virus genome.

S2 Table. List of the 8 viral overlapping genes for which there is only partial experimental evidence.

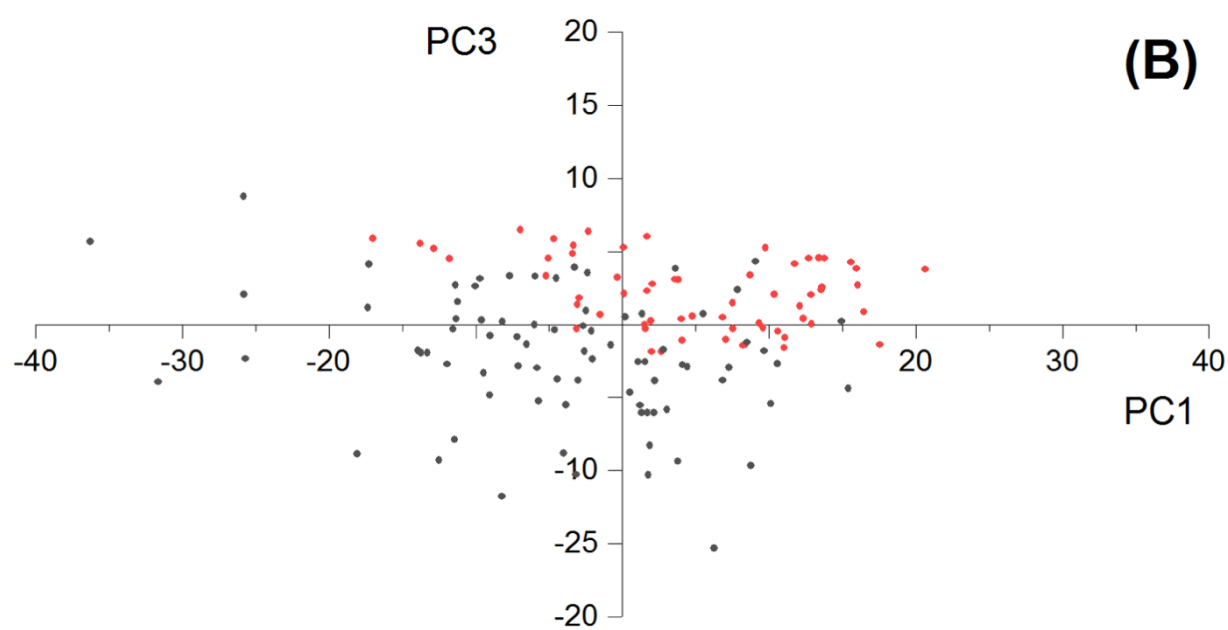
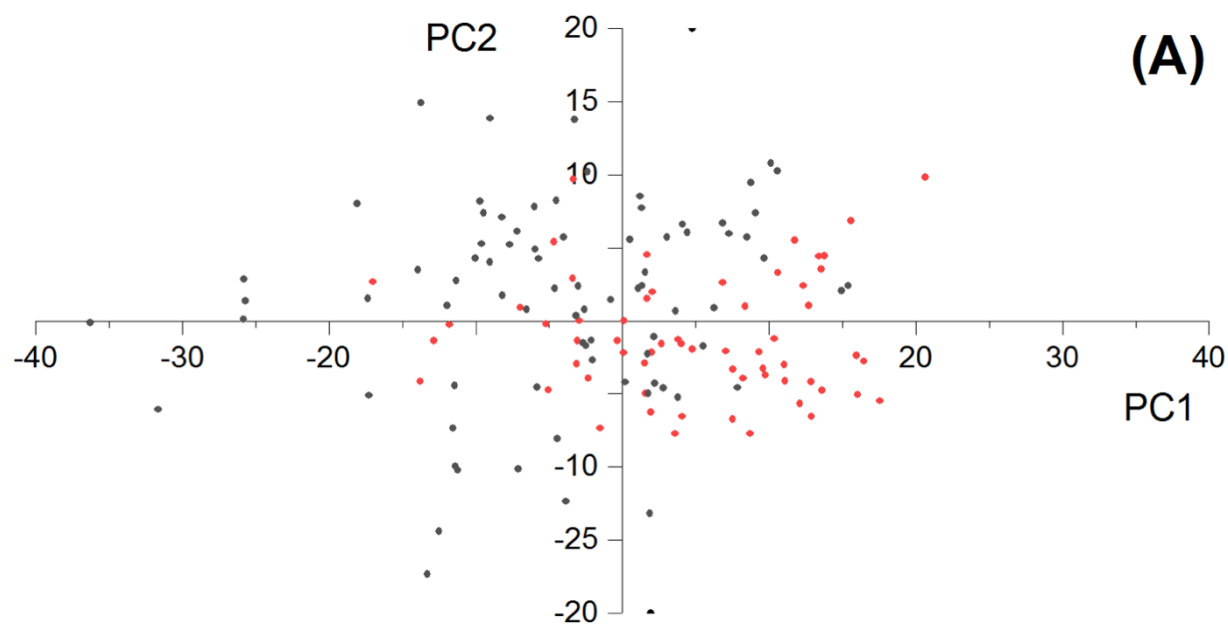
S3 Table. List of the 6 experimentally proven mammalian overlapping genes assembled and analysed in this work.

S4 Table. Details of the comparative analysis of overlapping and non-overlapping genes.

S4a Table shows the comparison of the pooled dataset of overlapping regions with that of the non-overlapping regions for 5 composition features. S4b Table lists the the 20 critical composition features peculiar to the overlapping gene dataset (chi-square > 100.0; 1 degree of freedom; $P < 0.00001$; z score < -2.55; $P < 0.01$).

S1 Dataset. Dataset of 80 proven overlapping genes from eukaryotic viruses.

S1 Figure



S1 Table. List of the 80 viral proven overlapping genes assembled in S1 Dataset. The genes are grouped in 7 tables (from S1a to S1g) in accordance to the nature of the virus genome.

S1a Table. 37 pairs of overlapping genes from 26 ssRNA+ viruses. The (§) symbol in the column “Genome Ac number” indicates the overlapping genes that were added by the authors to the NCBI reference genome database.

Family (-viridae)	Genus (-virus)	Virus species	Genome Ac number	Protein products	Protein Ac numbers
Arteri	Simarteri	Simian hemorrhagic fever virus	NC_003092	nsp2/nsp2TF	YP_009109556/YP_009172490
Arteri	Simarteri	Simian hemorrhagic fever virus	NC_003092	E/GP2	YP_009037600/NP_203547
Arteri	Simarteri	Simian hemorrhagic fever virus	NC_003092	GP3/GP4	NP_203548/NP_203549
Arteri	Simarteri	Simian hemorrhagic fever virus	NC_003092	GP5/5a	NP_203550/YP_009037601
Betaflexi	Tricho	Apple chlorotic leaf spot virus	NC_001409	movement protein/ capsid protein	NP_040552/NP_040553
Betaflexi	Capillo	Apple stem grooving virus	NC_001749	polyprotein/movement protein (36kDa protein)	NP_044335/NP_044336
Bromo	Ilar	Spinach latent virus	NC_003809	RNA-dependent RNA polymerase (2a) /2b	NP_620678/NP_620679
Bromo	Cucumo	Cucumber mosaic virus	NC_002035	RNA-dependent RNA polymerase (2a)/ 2b	NP_049324/NP_619631
Calici	Noro	Murine norovirus	NC_008311 (§)	capsid protein (VP1)/ VF1 (virulence factor 1)	YP_720002/YP_006390081
Carmotetra	Alphacarmo-tetra	Providencia virus	NC_014126	p130/replicase (p104)	YP_003620396/YP_003620397
Corona	Betacorona	SARS coronavirus	NC_004718	3a/3b	NP_828852/NP_828853
Corona	Betacorona	SARS coronavirus	NC_004718	nucleocapsid protein/ 9b	NP_828858/NP_828859
Dicistro	Apara	Israel acute paralysis virus	NC_009025 (§)	capsid protein/ORF _x (Pog)	YP_001040003/YP_006390080
Flavi	Hepaci	Hepatitis C virus	NC_004102	polyprotein/F (ARFP)	NP_671491/NP_803170
Hepe	Orthohepe	Hepatitis E virus	NC_001434	phosphoprotein (ORF3)/ capsid protein (ORF2)	YP_003864075/NP_056788

Luteo	Polero	Potato leafroll virus	NC_001747	P0/RNA-dependent RNA polymerase	NP_056746/NP_056748
Luteo	Polero	Potato leafroll virus	NC_001747	P1/RNA-dependent RNA polymerase	NP_056747/NP_056748
Luteo	Polero	Potato leafroll virus	NC_001747	capsid protein (P3)/movement protein (P4)	NP_056749/NP_056750
Luteo	Enamo	Pea enation mosaic virus	NC_003629	P0/RNA-dependent RNA polymerase	NP_619735/NP_620026
Noda	Alphanoda	Flock house virus	NC_004146	RNA-dependent RNA polymerase (A) /B2	NP_689444/NP_689446
Noda	Betanoda	Striped Jack nervous necrosis virus	NC_003448	RNA-dependent RNA polymerase (A)/B2 (B)	NP_599247/NP_599248
Picorna	Cardio	Encephalomyocarditis virus	NC_001479 (§)	polyprotein/2B*	NP_056777/YP_006383903
Picorna	Cardio	Theiler's murine encephalomyelitis virus	NC_001366	polyprotein/L*	NP_040350/YP_003587920
Poty	Poty	Sweet potato feathery mottle virus	NC_001841 (§)	polyprotein/P1N-PISPO	NP_045216/YP_009440977
Poty	Poty	Turnip mosaic virus	NC_002509	polyprotein/P3N-PIPO	NP_062866/YP_003587806
Tombus	Betacarmo	Hibiscus chlorotic ringspot virus	NC_003608	p28/p23	NP_619672/NP_619673
Tombus	Betacarmo	Hibiscus chlorotic ringspot virus	NC_003608	capsid protein/p25	NP_619676/NP_619677
Tombus	Machlomo	Maize chlorotic mottle virus	NC_003627	p32/p50	NP_619717/NP_619719
Tombus	Machlomo	Maize chlorotic mottle virus	NC_003627	p31/p7b	NP_619720/YP_009237216
Tombus	Machlomo	Maize chlorotic mottle virus	NC_003627	p31/capsid protein	NP_619720/NP_619722
Tombus	Panico	Panicum mosaic virus	NC_002598	capsid protein (p26)/p15	NP_068346/NP_068347
Tombus	Tombus	Tomato bushy stunt virus	NC_001554	p22/p19	NP_062900/NP_062901
Tombus	Umbra	Tobacco bushy top virus	NC_004366	movement protein (ORF3)/movement protein (ORF4)	NP_733849/NP_733850

Tymo	Tymo	Turnip yellow mosaic virus	NC_004063	movement protein (p69)/replicase	NP_663296/NP_663297
Unassigned	Sobemo	Sesbania mosaic virus	NC_002568	Px/polyprotein P2ab (protease domain)	YP_008873690/NP_066393
Unassigned	Sobemo	Sesbania mosaic virus	NC_002568	polyprotein P2a (ATPase P10 domain)/polyprotein P2ab (RdRp domain)	NP_066392/NP_066393
Unassigned	Sobemo	Sesbania mosaic virus	NC_002568	polyprotein P2ab (RdRp domain)/capsid protein	NP_066393/NP_066394

S1b Table. 15 pairs of overlapping genes from 13 ssRNA- viruses. The (§) symbol in the column “Genome Ac number” indicates the overlapping genes that were added by the authors to the NCBI reference genome database.

Family (-viridae)	Genus (-virus)	Virus species	Genome Ac number	Protein products	Protein Ac numbers
Borna	Borna	Borna disease virus 1	NC_001607	X protein/ phosphoprotein (P)	YP_009272535/NP_042021
Filo	Ebola	Zaire ebolavirus	NC_002549	secreted glycoprotein (sGP)/ envelope glycoprotein (GP1,2)	NP_066247/NP_066246
Hantaviridae	Orthohanta	Puumala virus	NC_005224	nucleocapsid protein/non-structural protein NSs	NP_941984/YP_004928150
Orthomyxo	InfluenzaA	Influenza A virus	NC_002021	RNA-dependent RNA polymerase (subunit PB1)/PB1-F2	NP_040985/YP_418248
Orthomyxo	InfluenzaA	Influenza A virus	NC_002022 (§)	RNA-dependent RNA polymerase (subunit PA)/PA-X	NP_040986/YP_006495785
Orthomyxo	InfluenzaB	Influenza B virus	NC_002209	glycoprotein NB/neuraminidase	NP_056662/NP_056663
Orthomyxo	Isa	Infectious salmon anemia virus	NC_006497	P6 (ORF2)/P7 (ORF1)	YP_145796/YP_145797
Paramyxo	Morbilli	Measles virus	NC_001498	phosphoprotein (P)/C	NP_056919/NP_056920
Paramyxo	Morbilli	Measles virus	NC_001498 (§)	phosphoprotein (P)/V	NP_056919/YP_003873249
Paramyxo	Respiro	Sendai virus	NC_001552	C'/phosphoprotein (P)	NP_056872/NP_056873
Peribunya	Orthobunya	La Crosse virus	NC_004110	nucleocapsid protein/non-structural protein NSs	NP_671970/NP_671971
Pneumo	Orthopneumo	Pneumonia virus of mice J3666	NC_006579	phosphoprotein (P)/ P2	YP_173327/YP_173328
Rhabdo	Cytorhabdo	Barley yellow striate mosaic	NC_028244	ORF4 protein/ORF5	YP_009177225/YP_009177226

		virus		protein	
Rhabdo	Vesiculo	Vesicular stomatitis Indiana virus	NC_001560	phosphoprotein (P)/C'	NP_041713/YP_003587923
Rhabdo	Vesiculo	Vesicular stomatitis New Jersey virus	NC_024473 (§)	phosphoprotein (P)/C'	YP_009047082/YP_009440976

S1c Table. 14 pairs of overlapping genes from 9 ssDNA viruses. The (§) symbol in the column “Genome Ac number” indicate the overlapping genes that were added by the authors to the NCBI reference genome database. The (**) symbol in the column “Genome Ac number” indicates overlapping genes generated by alternative splicing but without interruption of the reading frame.

Family (-viridae)	Genus (-virus)	Virus species	Genome Ac number	Protein products	Protein Ac numbers
Anello	Gyro	Chicken anemia virus	NC_001427	capsid protein (VP2)/apoptin (VP3)	NP_056773/NP_056774
Anello	Gyro	Chicken anemia virus	NC_001427	capsid protein (VP2)/nucleocapsid protein	NP_056773/NP_056775
Gemini	Begomo	East African cassava mosaic virus	NC_004674	AV2 protein/capsid protein (AV1)	NP_817107/NP_817108
Gemini	Begomo	East African cassava mosaic virus	NC_004674	transcriptional activator (TrAP, AC2)/replication enhancer (Ren, AC3)	NP_817111/NP_817110
Gemini	Begomo	East African cassava mosaic virus	NC_004674	replication associated protein (Rep, AC1)/AC4	NP_817112/NP_817113
Gemini	Curto	Beet curly top virus	NC_001412	movement protein (V3)/V2	NP_899663/NP_040558
Gemini	Curto	Beet curly top virus	NC_001412	C1/C2	NP_040557/NP_040561
Parvo	Bocaparvo	Canine minute virus	NC_004442 (**)	NS1/NP1	NP_758521/NP_758522
Parvo	Brevidenso	Aedes albopictus densovirus	NC_004285	NS1/NS2	NP_694827/NP_694828
Parvo	Dependoparvo	Adeno-associated virus-2	NC_001401	capsid protein (VP1)/AAP (Assembly Activating Protein)	YP_680426/YP_004030758
Parvo	Dependoparvo	Adeno-associated virus-2	NC_001401	capsid protein (VP1)/X protein	YP_680426/YP_009110690

Parvo	Erythroparvo	Human parvovirus B19	NC_000883 (**)	NS1/7.5 kDa protein	YP_004928144/YP_004928145
Parvo	Protoparvo	Porcine parvovirus	NC_001718 (§)	capsid protein (VP2)/SAT	NP_757372/YP_006355433
Parvo	Iteradenso	Dendrolimus punctatus densovirus	NC_006555	NS1/NS2	YP_164339/YP_164340

S1d Table. 5 pairs of overlapping genes from 5 dsRNA viruses. The (§) symbol in the column “Genome Ac number” indicates the overlapping genes that were added by the authors to the NCBI reference genome database.

Family (-viridae)	Genus (-virus)	Virus species	Genome Ac number	Protein products	Protein Ac numbers
Birna	Aquabirna	Infectious pancreatic necrosis virus	NC_001915	VP5/polyprotein	NP_047195/NP_047196
Reo	Orthoreo	Mammalian orthoreovirus	NC_013231 (§)	sigma1 (outer capsid protein, hemagglutinin)/sigma1s (minor capsid cell-attachment protein)	NP_694682/YP_009344826
Reo	Orbi	Bluetongue virus	NC_006008 (§)	VP6/NS4	YP_052953/YP_006390083
Reo	Rota	Rotavirus A	NC_011505	phosphoprotein (NSP5)/NSP6	YP_002302224/YP_002302225
Reo	Phyto	Rice dwarf virus	NC_003768	Pns12/Pns12-OP	NP_620538/YP_003587924

S1e Table. 1 pair of overlapping genes from a dsDNA virus. The double asterisk in the column “Genome Ac number” indicates overlapping genes generated by alternative splicing but without frame interruption.

Family (-viridae)	Genus (-virus)	Virus species	Genome Ac number	Protein products	Protein Ac numbers
Papilloma	Alphapapilloma	Human papillomavirus type 16	NC_001526 (**)	E2/E4	NP_041328/YP_009268708

S1f Table. 6 pairs of overlapping genes from 6 ssRNA-RT viruses. The single asterisk in the column “Genome Ac number” indicates overlapping genes in which at least one frame is interrupted by splicing. The (**) symbol in the column “Genome Ac number” indicates overlapping genes generated by alternative splicing but without interruption of the reading frame.

Family	Genus	Virus species	Genome	Protein products	Protein
--------	-------	---------------	--------	------------------	---------

(-viridae)	(-virus)		Ac number		Ac numbers
Retro	Deltaretro	Bovine leukemia virus	NC_001414 (*)	rex protein/tax protein	NC_056898/NC_056900
Retro	Lenti	Human immunodeficiency virus type 1	NC_001802	gag protein (p6 domain)/ pol protein (p6* domain)	NP_057850/NP_057849
Retro	Lenti	Human immunodeficiency virus type 2	KU179861 (*)	env protein/rev protein	ALQ56962/ALQ56964
Retro	Lenti	Simian immunodeficiency virus	NC_001549 (**)	vif protein/vpx protein	NP_054370/NP_054371
Retro	Lenti	Simian immunodeficiency virus SIV-mnd 2	NC_004455 (**)	env protein/nef protein	NP_758892/NP_758893
Retro	Spuma	Feline foamy virus	NC_001871 (*)	bel protein/bet protein	NP_056917/NP_056916

S1g Table. 2 pairs of overlapping genes from a dsDNA-RT virus

Family	Genus	Virus species	Genome	Protein products	Protein
(-viridae)	(-virus)		Ac number		Ac numbers
Hepadna	Orthohepadna	Hepatitis B	NC_003977	polymerase (P) /X protein	YP_009173866/YP_009173867
Hepadna	Orthohepadna	Hepatitis B	NC_003977	polymerase (P) /large envelope protein (L)	YP_009173866/YP_009173869

S2 Table. List of the 8 viral overlapping genes for which there is only partial experimental evidence

Family (-viridae)	Genus (-virus)	Virus species	Genome Ac number	Protein products	Protein Ac numbers	Boundaries of overlapping genes ^a
Alphaflexi	Mandari	Indian citrus ringspot virus	NC_003093	capsid protein/putative 23 kDa nucleic acid binding protein	NC_203557/NC_203558	6854-7156 6856-7158
Alphatetra	Omegatetra	Dendrolimus punctatus tetravirus	NC_005899	p17/capsid protein p71	YP_025095/YP_025096	372-755 374-757
Arteri	Simarteri	Simian hemorrhagic fever virus	NC_003092	GP2'/E'	NP_203544/YP_009037599	11046-11333 11047-11331
Arteri	Simarteri	Simian hemorrhagic fever virus	NC_003092	GP2'/GP3'	NP_203544/NP_203545	11484-11798 11486-11800
Barna	Barna	Mushroom bacilliform virus	NC_001633	hypothetical protein (ORF1)/hypothetical protein (ORF2)	NP_042508/NP_042509	67-600 68-601
Barna	Barna	Mushroom bacilliform virus	NC_001633	hypothetical protein (ORF2)/putative RNA-dependent RNA polymerase (ORF3)	NP_042509/NP_042510	1592-2041 1594-2043
Flavi	Flavi	Culex flavivirus	NC_008604	polyprotein (NS2A-NS2B)/truncated polyprotein (NS2AN-FIFO)	YP_899469/YP_006470608	3410-4294 3412-4293
Reo	Oryza	Rice ragged stunt virus	NC_003771	RNA-dependent RNA polymerase/36.9 kDa protein (P4b)	NP_620541/NP_620542	489-1472 491-1471

^a. The boundaries are given as the nucleotide position in the viral genome.

S3 Table. List of the 6 experimentally proven mammalian overlapping genes assembled and analysed in this work.

Gene	mRNA RefSeq ac number (and splice variant ac number if applicable)	Mechanism of expression (§)	Length of mRNA (nt)	Overlapping genes	Length of overlap (nt)	Boundaries of overlapping coding regions (on mRNA)	Boundaries of non-overlapping coding regions (on mRNA)	References for existence of the overlap and its boundaries
LBHD1 - LBH domain containing 1 (human)	NM_024099 mRNA expressing LBDH1	Overlap generated by alternative splicing	1357	LBHD1/C11orf98	255		*217-753: LBDH1	[1-3]
	NM_001286086 Variant 1 expressing C11orf98		666			*86-340: LBDH1 87-341: C11orf98	*48-86: C11orf98	
ATXN1 - ataxin 1 (human)	NM_000332 Variant 1 representing ATXN1 NM_001357857 Variant 1 representing Alt-ATXN1	Overlap generated by alternative start codon	10636	ATXN1/Alt-ATXN1	561	999-1559: ATXN1 1001-1558: Alt-ATXN1	972-998: ATXN1 1560-3419: ATXN1	[4]
PRNP - prion protein (human)	NM_183079 Variant 2 representing PrP NM_001271561 Variant 2 representing Alt-PrP	Overlap generated by alternative start codon	2804	PrP/Alt-PrP	225	513-737: PrP 515-736: Alt-PrP	426-512: PrP 738-1187: PrP	[5]
GNAS - GNAS complex locus (human)	NM_080425 Variant 2 representing XLas NM_001077490 Variant 2 representing Alex	Overlap generated by alternative start codon	3784	XLas /Alex	1884	472-2355: XLas 473-2353: Alex	286-471: XLas 2356-3399: XLas	[6, 7]
Adora2a - adenosine A2a receptor (mouse)	NM_001357942 mRNA representing uORF5 NM_053294 mRNA representing Adora2a	Overlap generated by alternative start codon	2491	uORF5/Adora2a	342	427-768: uORF5 429-770: Adora2a	364-426: uORF5 771-1661: Adora2a	[8]

Cdkn2a - cyclic dependent kinase inhibitor 2A (mouse)	NM_001040654 Variant 2 expressing p16INK4a	Overlap generated by alternative splicing	850	P16INK4a/P19ARF	198	*205-528: P16INK4a 207-527: P19ARF	*82-204: P16INK4a 529-588: P16INK4a	[9]
	NM_009877 Variant 1 expressing p19ARF		929				*97-285: P19ARF	

([§]) We list here the ultimate mechanism that results in the expression of two proteins from the same DNA sequence (see Results, paragraph on mechanism expression).

* Since the overlap is generated by an alternative splicing event, the transcript expressing one of the two frames does not contain enough non-overlapping coding region to reliably compare the compositional features; thus, some or all of boundaries of the non-overlapping regions of the latter refer to another transcript expressing it, as indicated in the table.

References

1. Oyama M, Kozuka-Hata H, Suzuki Y, Semba K, Yamamoto T, Sugano S. Diversity of translation start sites may define increased complexity of the human short ORFeome. *Mol Cell Proteomics*. 2007;6(6):1000-6. doi: 10.1074/mcp.M600297-MCP200.
2. Michel AM, Choudhury KR, Firth AE, Ingolia NT, Atkins JF, Baranov PV. Observation of dually decoded regions of the human genome using ribosome profiling data. *Genome Res*. 2012;22(11):2219-29. doi: 10.1101/gr.133249.111.
3. Chu Q, Rathore A, Diedrich JK, Donaldson CJ, Yates JR, 3rd, Saghatelian A. Identification of Microprotein-Protein Interactions via APEX Tagging. *Biochem*. 2017;56(26):3299-306. doi: 10.1021/acs.biochem.7b00265.
4. 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(30):21824-35. doi: 10.1074/jbc.M113.472654.
5. Vanderperre B, Staskevicius AB, Tremblay G, McCoy M, O'Neill MA, Cashman NR, et al. An overlapping reading frame in the PRNP gene encodes a novel polypeptide distinct from the prion protein. *FASEB J*. 2011;25(7):2373-86. doi: 10.1096/fj.10-173815.
6. Klemke M, Kehlenbach RH, Huttner WB. Two overlapping reading frames in a single exon encode interacting proteins—a novel way of gene usage. *EMBO J*. 2001;20(14): 3849-60.
7. Abramowitz J, Grenet D, Birnbaumer M, Torres HN, Birnbaumer L. XLas, the extra-long form of the α -subunit of the Gs G protein, is significantly longer than suspected, and so is its companion Alex. *Proceedings of the National Academy of Sciences USA*. 2004;101(22):8366-71.
8. Lee CF, Lai HL, Lee YC, Chien CL, Chern Y. The A2A adenosine receptor is a dual coding gene: a novel mechanism of gene usage and signal transduction. *J Biol Chem*. 2014;289(3):1257-70. doi: 10.1074/jbc.M113.509059.
9. Quelle DE, Zindy F, Ashmun RA, Sherr CJ. Alternative Reading Frames of the INK4a Tumor Suppressor Gene Encode Two Unrelated Proteins Capable of Inducing Cell Cycle Arrest. *Cell*. 1995;83:993-1000.

S4 Table. Details of the comparative analysis of overlapping and non-overlapping genes.

S4a Table shows the comparison of the pooled dataset of overlapping regions with that of the non-overlapping regions for 5 composition features. S4b Table lists the 20 critical composition features peculiar to the overlapping gene dataset (chi-square >100.0; 1 degree of freedom; P<0.00001, z score <-2.55; P<0.01).

S4a Table.

Compositional feature	Chi-square	Degrees of freedom	P<
Nucleotides	745.1	3	0.00001
Dinucleotides	1678.5	15	0.00001
Amino acids	1125.0	19	0.00001
Amino acids (high, medium, or low codon degeneracy)	360.9	2	0.00001
Synonymous codons	2242.2	58	0.00001

S4b Table

Compositional feature	Percent content in the pooled overlapping dataset	Percent content in the pooled non-overlapping dataset	Percent difference (overlap – non-overlap)	Chi-square	z score (Wilcoxon test for paired data)
A	27.03	29.97	-2.94	136.4	-2.56
T	22.24	25.95	-3.72	238.6	-4.94
C	26.48	21.08	5.40	571.7	-6.30
AT	5.80	7.81	-2.01	187.7	-5.19
TA	3.81	5.93	-2.12	271.5	-5.78
TT	5.42	7.15	-1.72	128.0	-4.85
CG	4.70	3.02	1.68	306.3	-5.80
CC	7.35	4.85	2.50	431.6	-4.30
Arg	7.47	5.27	2.19	188.0	-5.62
Ser	9.51	7.27	2.24	146.5	-5.60
Pro	7.31	4.98	2.33	221.7	-3.90
Tyr	2.32	3.66	-1.34	109.3	-6.35
Ile	4.39	6.05	-1.66	102.5	-4.62
High-degeneracy amino acids	26.86	21.77	5.08	305.2	-7.17
Low-degeneracy amino acids	36.62	41.59	-4.97	209.5	-6.09
CGA (Arg)	1.36	0.59	0.77	171.3	-5.93
TCG (Ser)	1.26	0.61	0.64	116.8	-5.45
CCC (Pro)	2.07	1.08	0.99	160.5	-2.79
CCG (Pro)	1.37	0.65	0.71	135.1	-4.42
TAT (Tyr)	1.05	2.18	-1.13	124.9	-5.38

