PNAS
www.pnas.org

Supplementary Information for

Definitive demonstration by synthesis of genome annotation completeness

Paul R. Jaschke[b,1], Gabrielle A Dotson[a], Kay Hung[a], Diane Liu[a], and Drew Endy[a,1]

[a]Bioengineering Department, Stanford University, Stanford, CA, 94305
[b]Department of Molecular Sciences, Macquarie University, Sydney, NSW, 2066, Australia

[1]To whom correspondence may be addressed:

Paul R Jaschke
Email: paul.jaschke@mq.edu.au

Drew Endy
Email: endy@stanford.edu

This PDF file includes:

Figures S1 to S7
Tables S1 to S2

Other supplementary materials for this manuscript include the following:

Datasets S1 to S4

www.pnas.org/cgi/doi/10.1073/pnas.1905990116

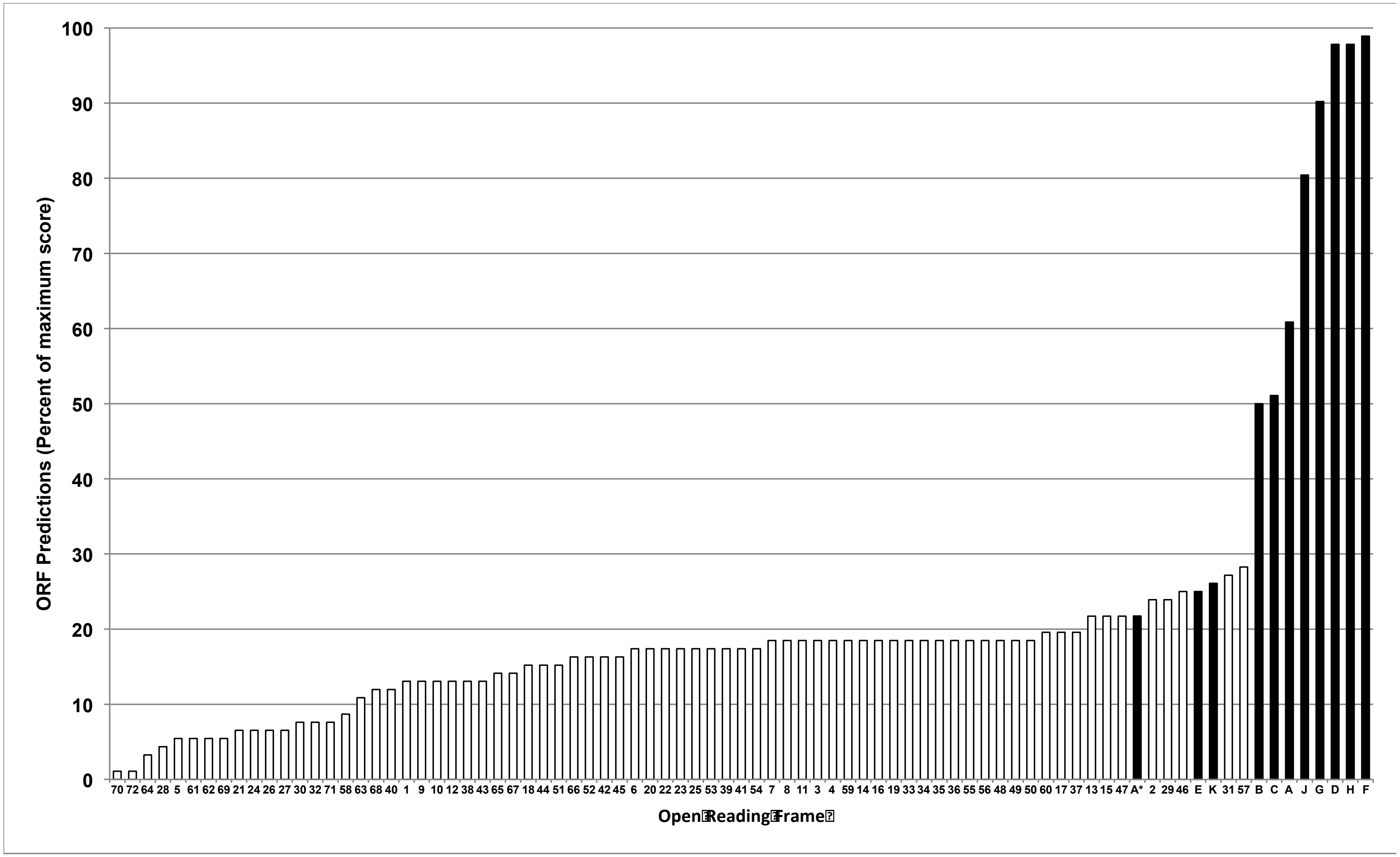

Fig. S1. Computational ORF predictions. ORFs were given a score based on the number of times it was identified across 23 *Bullavirinae* genomes (92 possible identifications for each ORF, based on four computational tools and 23 genomes). Seventy-two cryptic ORFs (white bars) and 11 previously discovered øX174 protein-coding ORFs (black bars).

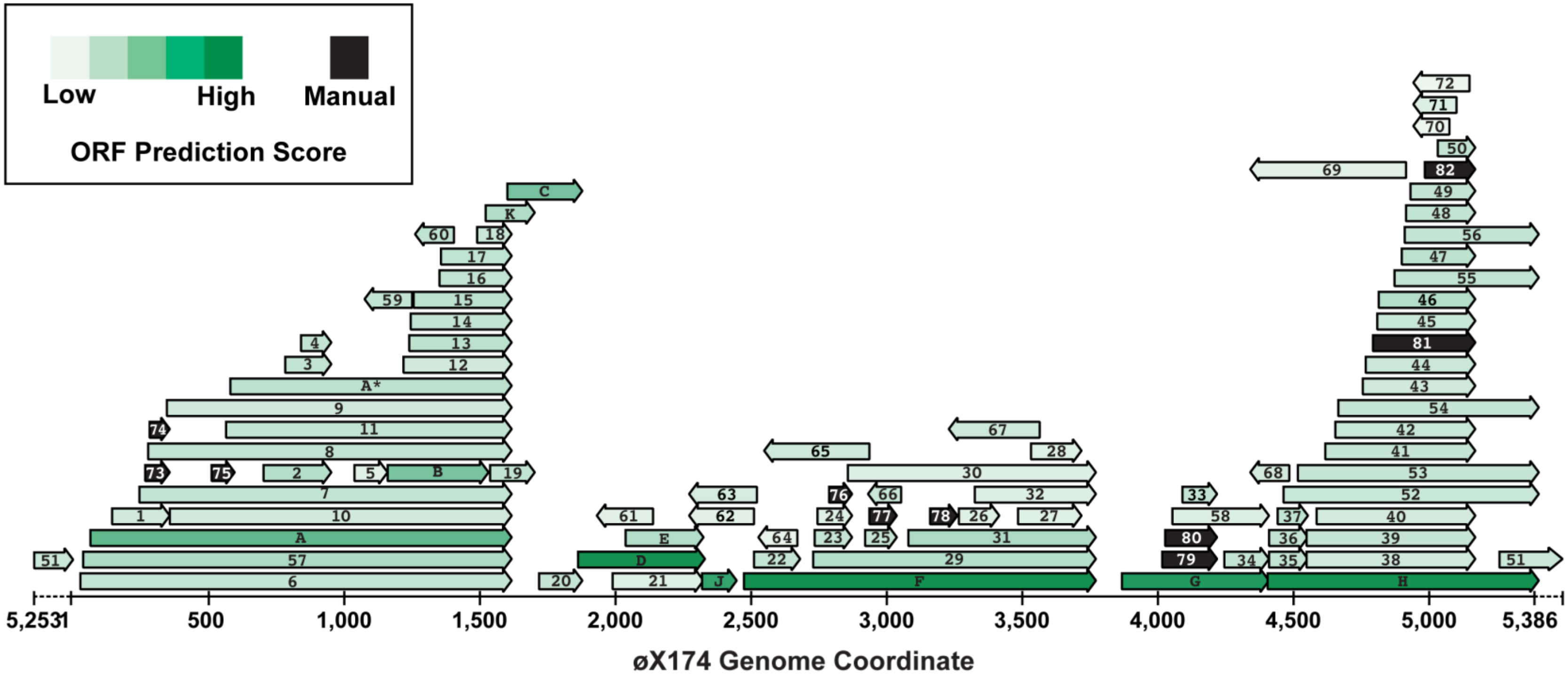

Fig. S2. Computationally identified ORF locations on øX174 genome. ORFs were given a score based on the number of times it was identified across 23 *Bullavirinae* genomes (92 possible identifications for each ORF, based on four computational tools and 23 genomes). Black ORFs indicate 10 expert-curated cryptic ORFs **(1)**.

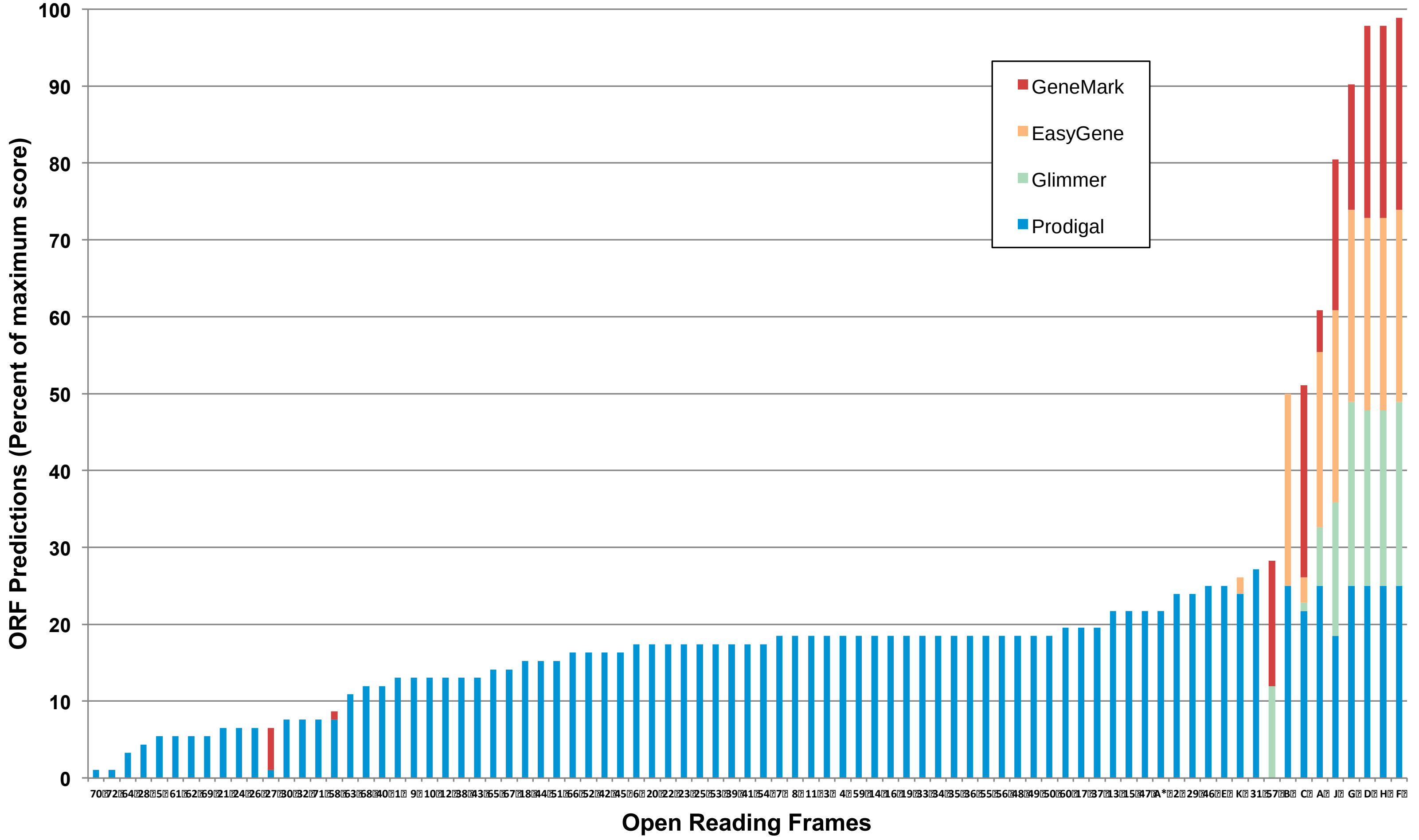

Fig. S3. Computational tool contributions to ORF prediction scores. Identified ORFs were given a score based on the number of times it was identified across 23 *Bullavirinae* genomes (92 possible identifications for each ORF, based on four computational tools and 23 genomes). Four standard gene prediction tools were used: GLIMMER (2), GeneMark (3) using the GeneMark.hmm PROKARYOTIC (Version 2.10b) algorithm, EasyGene (4), and Prodigal (5).

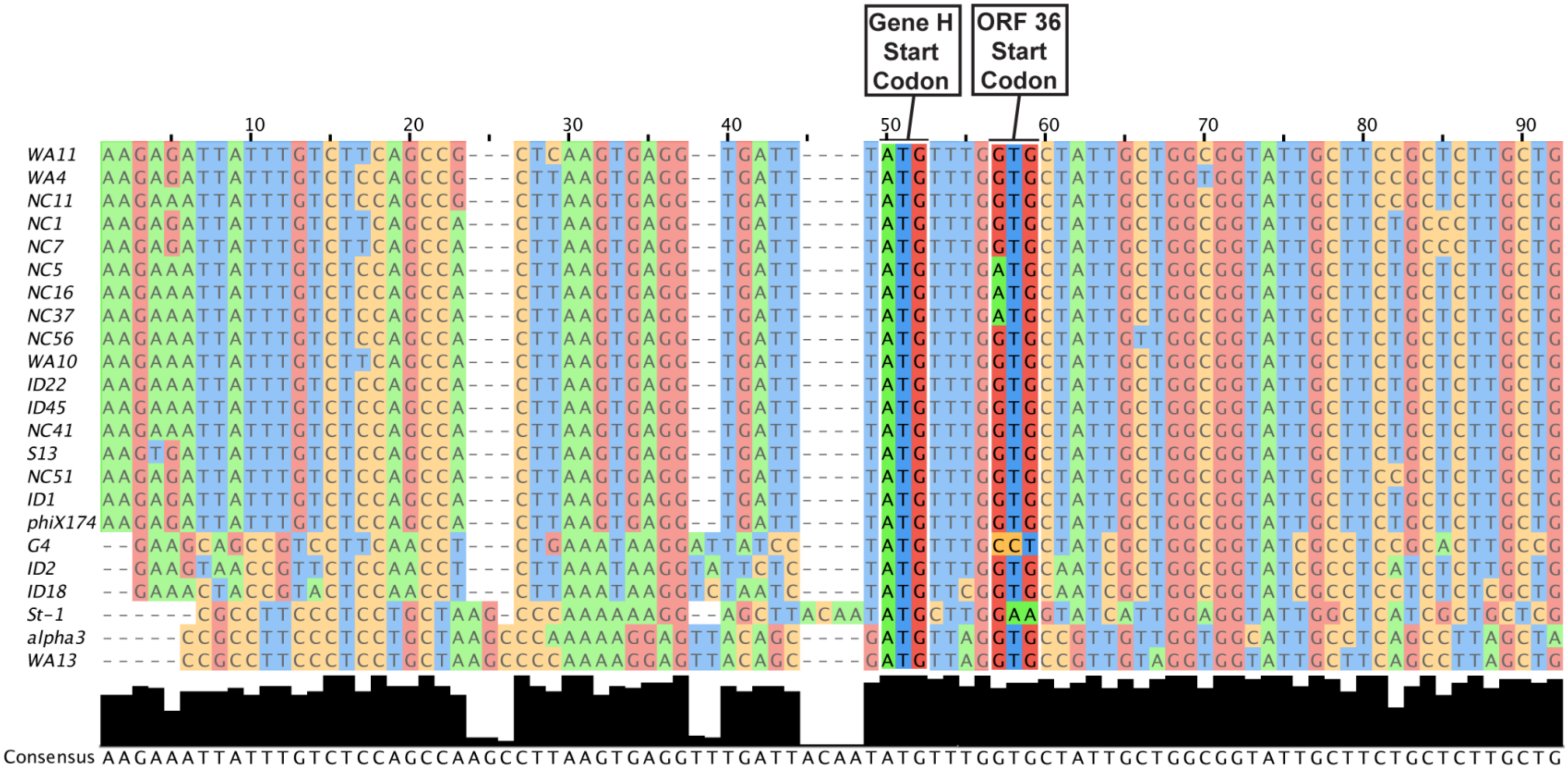

Fig. S4. Multiple sequence alignment of Gene H/ORF 36 regions from 23 *Bullavirinae* genomes. Multiple sequence alignment of 83 nt centered on the start codon of gene H performed with MUSCLE (6) multiple sequence alignment algorithm. Height of the black bars below each nucleotide represents degree of conservation within that column.

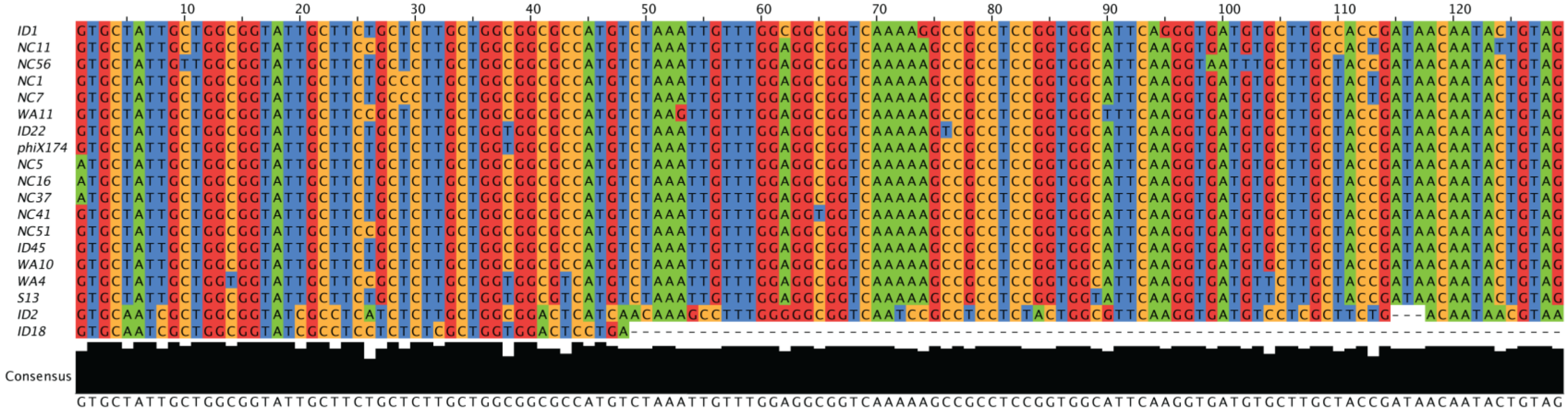
10
20
30
40
50
60
70
80
90
100
110
120
ID1
NC11
NC56
NC1
NC7
WA11
ID22
phiX174
NC5
NC16
NC37
NC41
NC51
ID45
WA10
WA4
S13
ID2
ID18
Consensus
GTGCTATTGCTGGCGGTATTGCTTCTGCTCTTGCTGGCGGCGCCATGTCTAAATTGTTTGGAGGCGGTCAAAAAGCCGCCTCCGGTGGCATTCAAGGTGATGTGCTTGCTACCGATAACAATACTGTAG

Fig. S5. Multiple sequence alignment of ORF 36 from 19 *Bullavirinae* genomes. Multiple sequence alignment of ORF 36 performed with MUSCLE multiple sequence alignment algorithm. WA13 and alpha3 ORF36 sequences not included in alignment because their short length disrupted the alignment. G4 and st-1 ORF36 sequences lack strong start codons and were not included in the alignment. Height of the black bars below each nucleotide represents degree of conservation within that column.

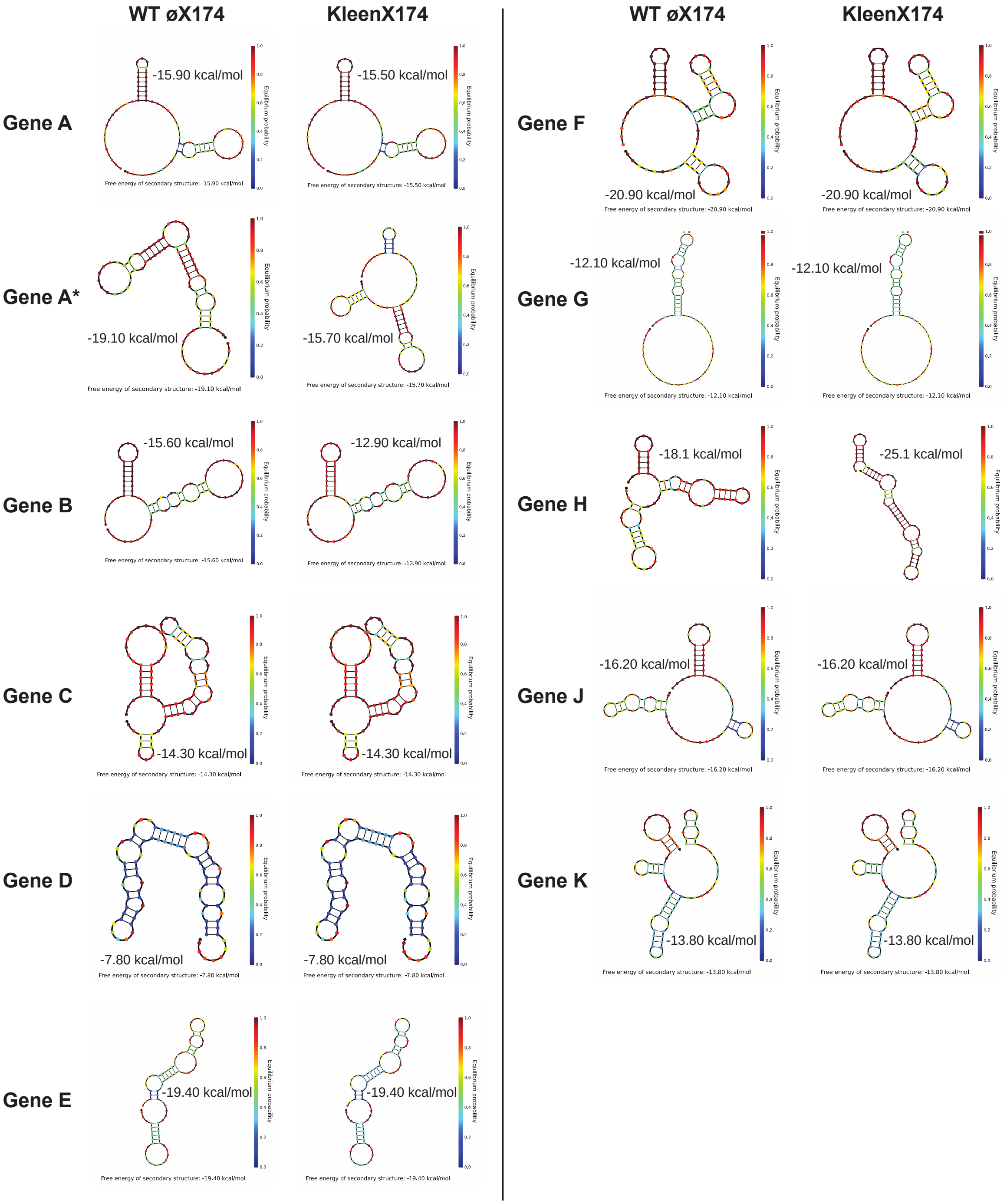
WT øX174
KleenX174
WT øX174
KleenX174
Gene A
-15.90 kcal/mol
Free energy of secondary structure: -15.90 kcal/mol
-15.50 kcal/mol
Free energy of secondary structure: -15.50 kcal/mol
Gene F
-20.90 kcal/mol
Free energy of secondary structure: -20.90 kcal/mol
-20.90 kcal/mol
Free energy of secondary structure: -20.90 kcal/mol
Gene A*
-19.10 kcal/mol
Free energy of secondary structure: -19.10 kcal/mol
-15.70 kcal/mol
Free energy of secondary structure: -15.70 kcal/mol
Gene G
-12.10 kcal/mol
Free energy of secondary structure: -12.10 kcal/mol
-12.10 kcal/mol
Free energy of secondary structure: -12.10 kcal/mol
Gene B
-15.60 kcal/mol
Free energy of secondary structure: -15.60 kcal/mol
-12.90 kcal/mol
Free energy of secondary structure: -12.90 kcal/mol
Gene H
-18.1 kcal/mol
-25.1 kcal/mol
Gene C
-14.30 kcal/mol
Free energy of secondary structure: -14.30 kcal/mol
-14.30 kcal/mol
Free energy of secondary structure: -14.30 kcal/mol
Gene J
-16.20 kcal/mol
Free energy of secondary structure: -16.20 kcal/mol
-16.20 kcal/mol
Free energy of secondary structure: -16.20 kcal/mol
Gene D
-7.80 kcal/mol
Free energy of secondary structure: -7.80 kcal/mol
-7.80 kcal/mol
Free energy of secondary structure: -7.80 kcal/mol
Gene K
-13.80 kcal/mol
Free energy of secondary structure: -13.80 kcal/mol
-13.80 kcal/mol
Free energy of secondary structure: -13.80 kcal/mol
Gene E
-19.40 kcal/mol
Free energy of secondary structure: -19.40 kcal/mol
-19.40 kcal/mol
Free energy of secondary structure: -19.40 kcal/mol
Equilibrium probability
1.0
0.8
0.6
0.4
0.2
0.0

Fig. S6. Simulated RNA folding structures of all known øX174 protein-coding ORFs in WT and kleenX174 genomes. NUPACK (7) lowest energy RNA structure from 83 nt window surrounding each known øX174 gene using sequence variants from WT øX174 and kleenX174 genome sequences.

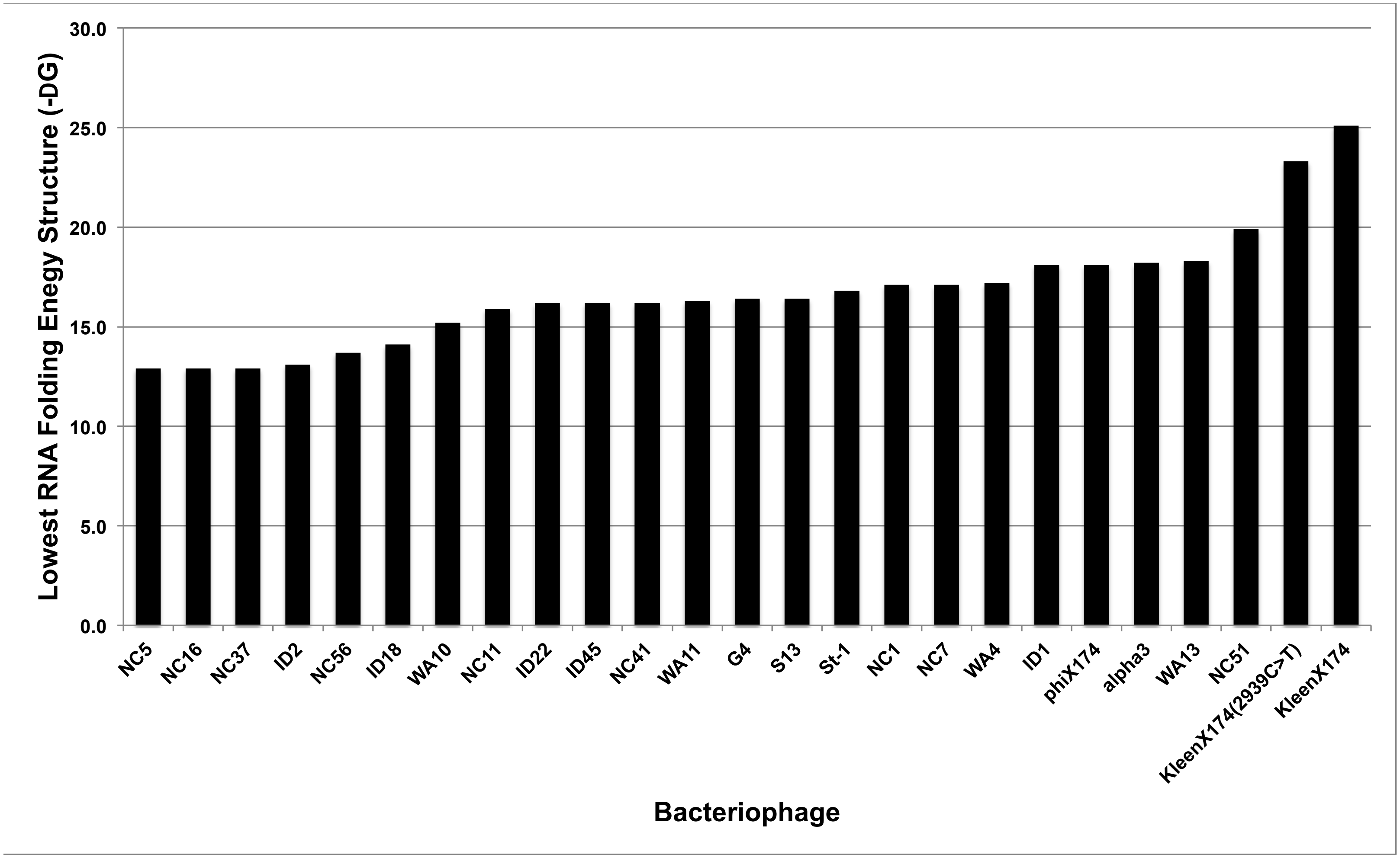

Fig. S7. Predicted gene H RNA structure in 23 *Bullavirinae* genomes shows kleenX174 and kleenX174(2939C>T) outside the normal range. Lowest energy structures of RNA folding simulation performed with NUPACK using default parameters. Folding window of 83 nt centered on gene H initiation codon.

Table S1. Changes made to wild-type øX174 genome to produce cryptX174 design.

Genome Coordinate[a]	WT Nucleotide	CryptX Nucleotide	Gene 1	Codon Affected[b]	Amino Acid Change	Gene 2	Codon Affected[b]	Amino Acid Change	Gene 3	Codon Affected[b]	Amino Acid Change
3982	T	C	A	`atg > aCg`	Start (Met) > Thr	-	-	-	-	-	-
4499	G	A	A/A*	`atg > atA`	Start (Met) > Ile	-	-	-	-	-	-
5076	T	C	B	`atg > aCg`	Start (Met) > Thr	A/A*	`tgg > Cgg`	Trp > Arg	-	-	-
53	G	A	K	`atg > atA`	Start (Met) > Ile	A/A*	`gag > Aag`	Glu > Lys	-	-	-
135	G	A	C	`atg > atA`	Start (Met) > Ile	K	`gag > Aag`	Glu > Lys	A/A*	`tga > tAa`	Stop > Stop
392	G	A	D	`atg > atA`	Start (Met) > Ile	C	`tga > tAa`	Stop > Stop	-	-	-
569	T	C	E	`atg > aCg`	Start (Met) > Thr	D	`tag > taC`	Tyr > Tyr	-	-	-
850	G	A	J	`atg > atA`	Start (Met) > Ile	-	-	-	-	-	-
1003	G	A	F	`atg > atA`	Start (Met) > Ile	-	-	-	-	-	-
2397	G	A	G	`atg > atA`	Start (Met) > Ile	-	-	-	-	-	-
2933	G	A	H	`atg > atA`	Start (Met) > Ile	-	-	-	-	-	-

[a] Coordinate in original Sanger øX174 genome sequence (Genbank NC_001422.1)

[b] Capital letters represent mutated base positions in CryptX174 design.

Table S2. Oligonucleotides used in this work.

Name	FORWARD/ REVERSE	Target	Sequence
Chimera_P1-FOR	FORWARD	WT_Part1/KX_Part1	GTCTAGGAAATAACCGTCAGGATTGACACCC
Chimera_P2-FOR	FORWARD	WT_Part2/KX_Part2	AAAATACGTGGCCTTATGGTTACAGTATGCCCATCG
Chimera_P3-FOR	FORWARD	WT_Part3/KX_Part3	GGAGTGATGTAATGTCTAAAGGTAAAAAACGTTCTGGCG
Chimera_P4-FOR	FORWARD	WT_Part4/KX_Part4	GGCACTATGTTTACTCTTGCGCTTGTTCG
Chimera_P5_WT-FOR	FORWARD	WT_Part5	GCCACTTAAGTGAGGTGATTTATGTTTGGTGCTATTGCTGGCG
Chimera_P5_KX-FOR	FORWARD	KX_Part5	GCCACTTAAGTGAGGTGATTTATGTTCGGCGCTATTGCTGG
Chimera_P1-REV	REVERSE	WT_Part1/KX_Part1	GCATACTGTAACCATAAGGCCACGTATTTTGCAAGC
Chimera_P2-REV	REVERSE	WT_Part2/KX_Part2	CGTTTTTTACCTTTAGACATTACATCACTCCTTCCGC
Chimera_P3_WT-REV	REVERSE	WT_Part3	GAACAAGCGCAAGAGTAAACATAGTGCCATGCTCAGGAACAAAG
Chimera_P3_KX-REV	REVERSE	KX_Part3	GAACAAGCGCAAGAGTAAACATAGTGCCGTGTTCGGGAACAAAGAAACG
Chimera_P4-REV	REVERSE	WT_Part4/KX_Part4	AACATAAATCACCTCACTTAAGTGGCTGG
Chimera_P5-REV	REVERSE	WT_Part5/KX_Part5	TCAATCCTGACGGTTATTTCCTAGACAAATTAGAGCCAATACC
KleenXSeq_1	FORWARD	KleenX174 Genome	CTGGCGACCCTGTTTTGTAT
KleenXSeq_2	FORWARD	KleenX174 Genome	CGGATATTTCTGATGAGTCGAA
KleenXSeq_3	FORWARD	KleenX174 Genome	CTACACGCAGGACGCTTTTTCA
KleenXSeq_4	FORWARD	KleenX174 Genome	TCTTTCTCAATCCCCAATGC
KleenXSeq_5	FORWARD	KleenX174 Genome	AAGTCACTTGGGGTTTCTGG
KleenXSeq_6	FORWARD	KleenX174 Genome	ATCTGTCAACGCCGCTAATC
KleenXSeq_7	FORWARD	KleenX174 Genome	GCCCCTAGTTTCGTTTCTGG
KleenXSeq_8	FORWARD	KleenX174 Genome	GAAATTATGCGCCAGATGCT

Dataset S1 (separate file). PhiX174 ORF characteristics and modifications to generate kleenX174 genome design. Excel file.

Dataset S2 (separate file). Wild type sequence gene H synthetic template for cell-free protein expression. Genbank file.

Dataset S3 (separate file). KleenX174 sequence gene H synthetic template for cell-free protein expression. Genbank file.

Dataset S4 (separate file). KleenX174(2939C>T) sequence gene H synthetic template for cell-free protein expression. Genbank file.

References

1. Godson GN, Fiddes JC, Barrell BG, Sanger F. Comparative DNA Sequence Analysis of the G4 and phiX174 Genomes. *The Single-Stranded DNA Phages* 8. (1978). doi:10.1101/087969122.8.51
2. Delcher AL, Bratke KA, Powers EC, Salzberg SL (2007) Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* 23 (6):673-679. doi:10.1093/bioinformatics/btm009
3. Besemer J, Borodovsky M (2005) GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic Acids Res* 33 (Web Server issue):W451-454. doi:10.1093/nar/gki487
4. Larsen TS, Krogh A (2003) EasyGene--a prokaryotic gene finder that ranks ORFs by statistical significance. *BMC Bioinformatics* 4:21. doi:10.1186/1471-2105-4-21
5. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11:119. doi:10.1186/1471-2105-11-119
6. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32 (5):1792-1797. doi:10.1093/nar/gkh340
7. Zadeh JN, Steenberg CD, Bois JS, Wolfe BR, Pierce MB, Khan AR, Dirks RM, Pierce NA (2011) *NUPACK: Analysis and design of nucleic acid systems.* J Comput Chem 32 (1):170-173. doi:10.1002/jcc.21596