Cell Reports, Volume *22*

Supplemental Information

Active Ebola Virus Replication and Heterogeneous Evolutionary Rates in EVD Survivors

Shannon L.M. Whitmer, Jason T. Ladner, Michael R. Wiley, Ketan Patel, Gytis Dudas, Andrew Rambaut, Foday Sahr, Karla Prieto, Samuel S. Shepard, Ellie Carmody, Barbara Knust, Dhamari Naidoo, Gibrilla Deen, Pierre Formenty, Stuart T. Nichol, Gustavo Palacios, Ute Ströher, and Ebola Virus Persistence Study Group

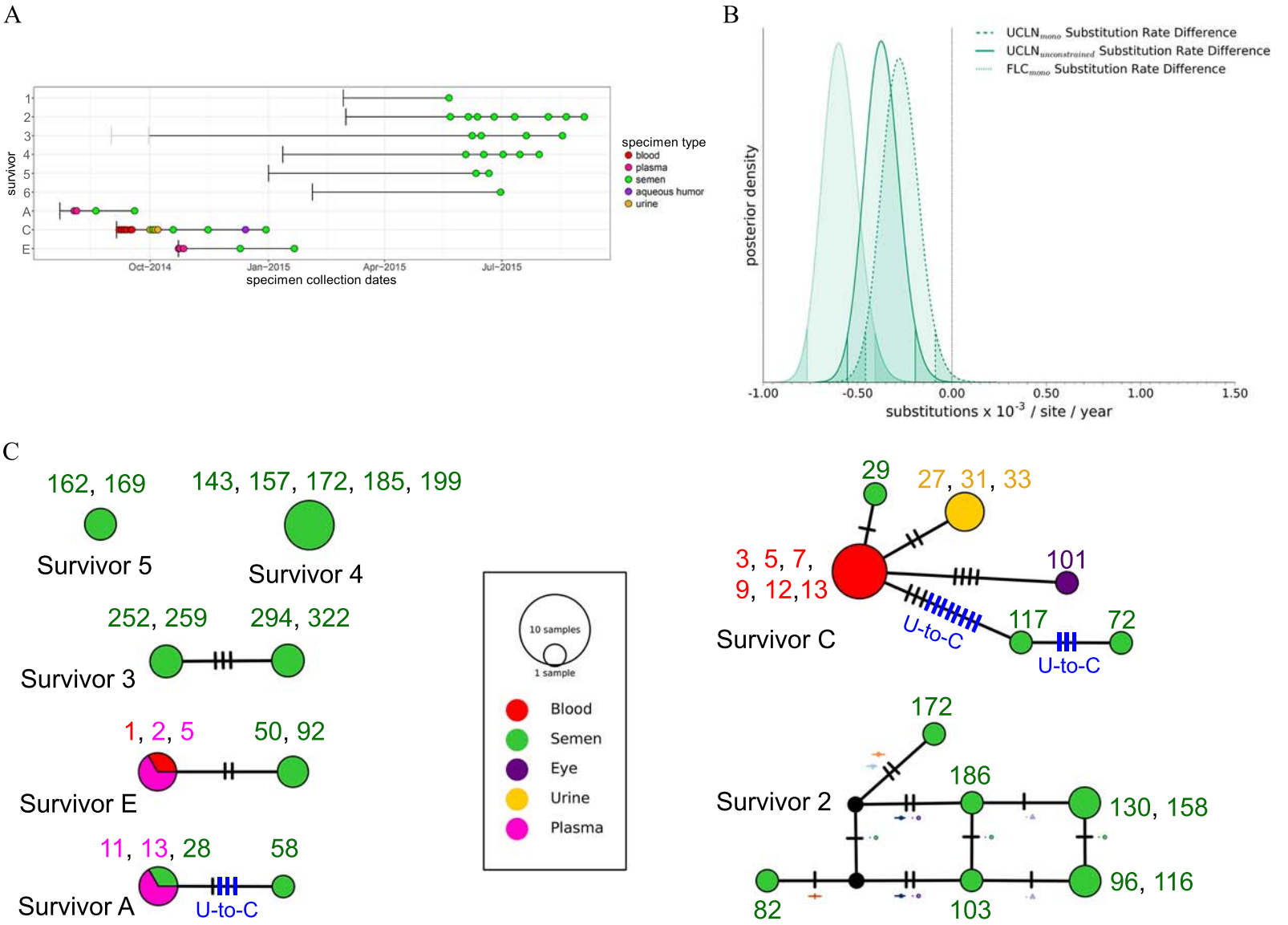

Figure S1: Overview of clinical specimens collected from Ebola virus disease (EVD) survivors, viral evolutionary rates and comparison of viral sequence changes per survivor, Related to Figures 1 and 2. **A)** Overview of clinical specimens collected from Ebola virus disease (EVD) survivors in Sierra Leone (survivors 1, 2, 3, 4, 5, and 6) and in the United States (survivors A, C, E). Survivor-reported symptom onset date is indicated with a black vertical bar, and survivor-reported ambiguity in onset is illustrated with a grey line (survivor 3). Clinical specimens from US EVD survivors were collected during acute and persistent infection, while clinical specimens from Sierra Leonean EVD survivors were collected only during persistent viral infection. Additional specimens were collected from survivors; here we only include specimens that produced a nearly-complete viral genome. **B)** Ebola virus in semen specimens from Sierra Leonean EVD survivors exhibits reduced evolutionary rates. Posterior distribution of evolutionary rate differences from serial semen specimens provided by EVD survivors relative to acute viral evolutionary rates calculated under FLC and UCLN clock models. FLC_{mono} and $UCLN_{mono}$ rates were calculated with SAVS constrained to survivor-specific monophyletic taxons (2, 3, 4, and 5), while $UCLN_{unconstratined}$ rates were calculated without prior assumptions on the tree. Regions within the shaded density tails indicate the 95% highest posterior density interval (HPD) and black dotted line indicates zero rate distribution difference. **C)** Comparison of AAVS and SAVS from EVD survivors. Median joining haplotype networks constructed using AAVS and SAVS from EVD survivors. Vertical bars indicate nucleotide changes (excluding regions that contain N, ? or -, representing less than 1.1% of consensus genomes. A single sequence with low coverage was removed from this figure (KY805812, survivor C)). Nodes are colored according to specimen matrix from which viral sequences were obtained and node size represents the number of clinical specimens. Numbers above nodes represent dayspost symptom onset. For survivor 2, symbols next to vertical bars coincide with iSNVs symbols in Supplemental Figure 2A. SAVS from survivors A (3 sites) and C (11 sites) exhibited potential evidence of human U-to-C hyper-editing following prolonged MGT persistence.

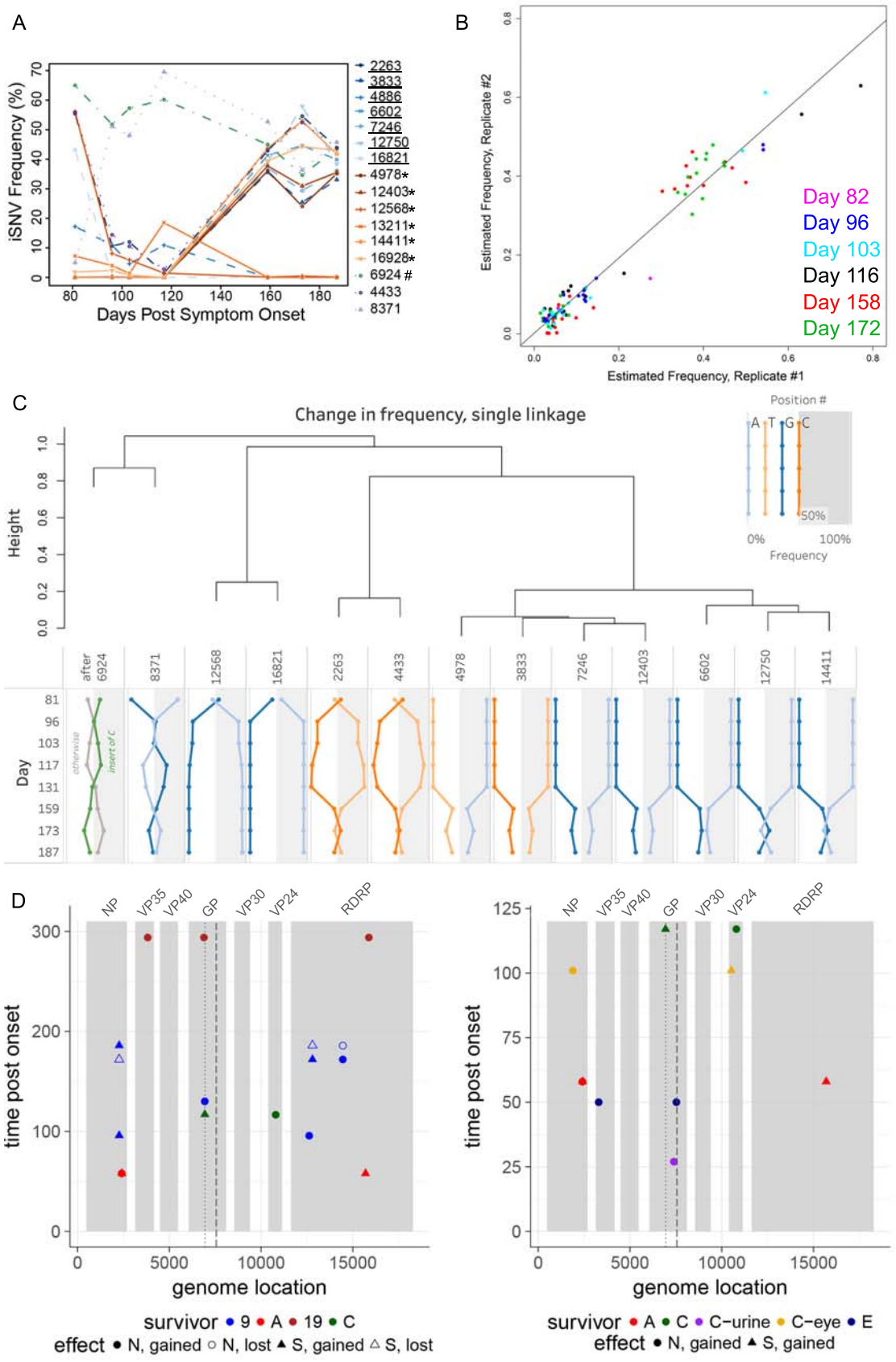

Figure S2: Comparison of AAVS and SAVS from EVD survivors, Related to Figures 2 and 3. **A)** Change in frequency for intrahost single nucleotide variants (iSNVs) (with greater than 15% frequency in a single specimen) versus time post symptom onset for Survivor 2. Sites that result in synonymous (underlined), nonsynonymous (starred), or frameshift (hash) mutations are highlighted and sites without annotations occur in noncoding regions. **B)** Resequencing of technical duplicates yields a similar correlation in iSNV frequencies for SAVS from survivor 2 (r^2=0.9515). **C)** Frequency of intrahost single nucleotide variants (iSNVs) versus time post symptom onset for Survivor 2. A pairwise (Manhattan) distance matrix was computed for each position-allele combination with the vector of the observed frequencies ordered by specimen date. The matrix was used to generate a single-linkage dendrogram (top). Frequency line graphs of iSNV positions, major/minor alleles, and specimen dates were ordered by their position in the dendrogram. Key in upper right-hand corner illustrates allele state (major or minor - grey shading) and value (A,T,C, or G). The presence of co-varying frequency changes suggests either: 1) distinct viral sub-populations, and/or 2) epistasis at the co-varying sites. **D-E)** Acquisition/Loss of synonymous (S) or nonsynonymous (N) changes in SAVS compared to earliest SAVS or AAVS from each survivor. **D)** Coding region changes for SAVS compared to earliest available SAVS from each survivor. Dotted line indicates glycoprotein editing site and dashed line indicates GP1/2 cleavage site. **E)** Coding region changes for SAVS compared to earliest available AAVS from each survivor. Dotted line indicates glycoprotein editing site and dashed line indicates GP1/2 cleavage site.

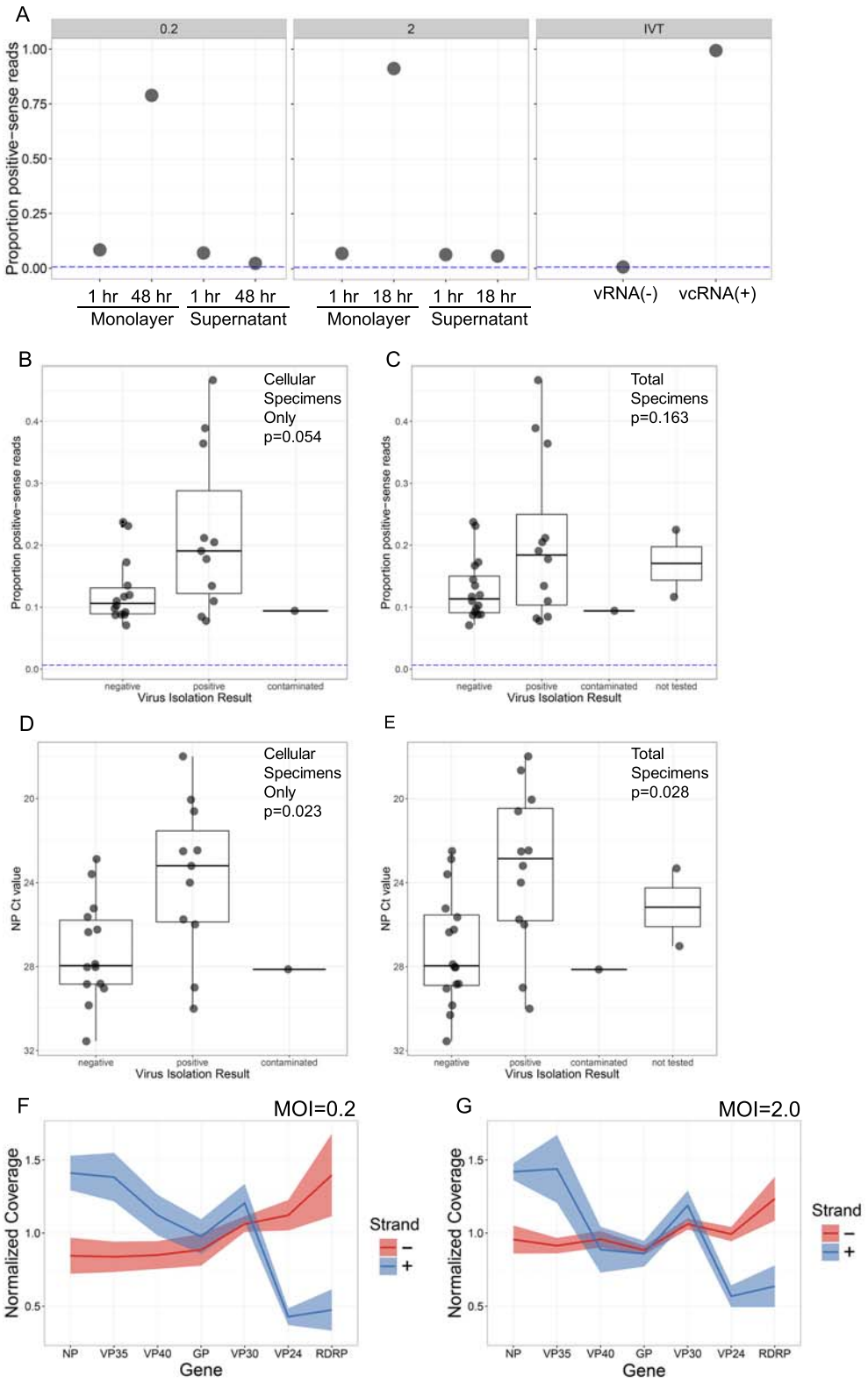

Figure S3: Supportive evidence for active viral replication during persistent infection, Related to Figure 4. To confirm the presence of positive-sense reads from SAVS, we validated our NGS assay with RNA extracted from Huh7 cells infected with recombinant Ebola virus encoding for ZsGreen protein (EBOV-ZsGreen) and *in vitro* transcribed RNA. **A)** Proportion of EBOV genome-wide positive-sense reads sequenced with NGS from an *in vitro* infection of Huh7 cells done at an MOI of 0.2 (1 and 48hpi) (left panel) or 2.0 (1 and 18hpi) (middle panel). Right panel indicates proportion of EBOV genome-wide positive-sense reads from the *in vitro* transcription(IVT) of a negative-sense (vRNA(-)) or positive (vcRNA(+)) viral transcript. **B)** One-sided ANOVA indicates a modest relationship between the proportion of positive-sense reads and virus isolation results ($p=0.054$). This analysis was conducted on clinical specimens containing only cellular material (blood and semen). Maxima and minima in boxplot illustrates the 25th and 75th percentiles, black line indicates median values, whiskers indicate the highest/lowest values within 1.5x the inter-quartile range. **C)** One-sided ANOVA indicates a limited relationship between the proportion of positive-sense reads and virus isolation results ($p=0.163$). This analysis was conducted on clinical specimens containing both acellular (urine, aqueous humor, and plasma) and cellular material (blood and semen). Boxplot values are described in panel B. **D)** One-sided ANOVA indicates a statistically significant ($p<0.05$) relationship between NP real-time polymerase chain reaction (RT-PCR) cycle threshold (Ct) value and virus isolation results ($p=0.023$). This analysis was only conducted on clinical specimens containing cellular material (blood and semen). Boxplot values are described in panel B. **E)** One-sided ANOVA indicates a statistically significant relationship between NP Ct value and virus isolation results ($p=0.028$). This analysis was conducted on clinical specimens containing both acellular (urine, aqueous humor, and plasma) and cellular material (blood and semen). Boxplot values are described in panel B. **F)** Proportion of strand-specific reads per EBOV gene (normalized to total positive- or negative-sense reads) from *in vitro* infection of Huh7 cells at MOI of 0.2. Data represents monolayer and supernatant samples collected after one hour and 48hrs post infection. Negative-sense reads in red, and positive-sense reads in blue. **G)** Proportion of strand-specific reads per EBOV gene (normalized to total positive- or negative-sense reads) from *in vitro* infection of Huh7 cells at MOI of 2. Data represents monolayer and supernatant samples collected after one hour and 18 hrs post infection. Negative-sense reads in red, and positive-sense reads in blue.

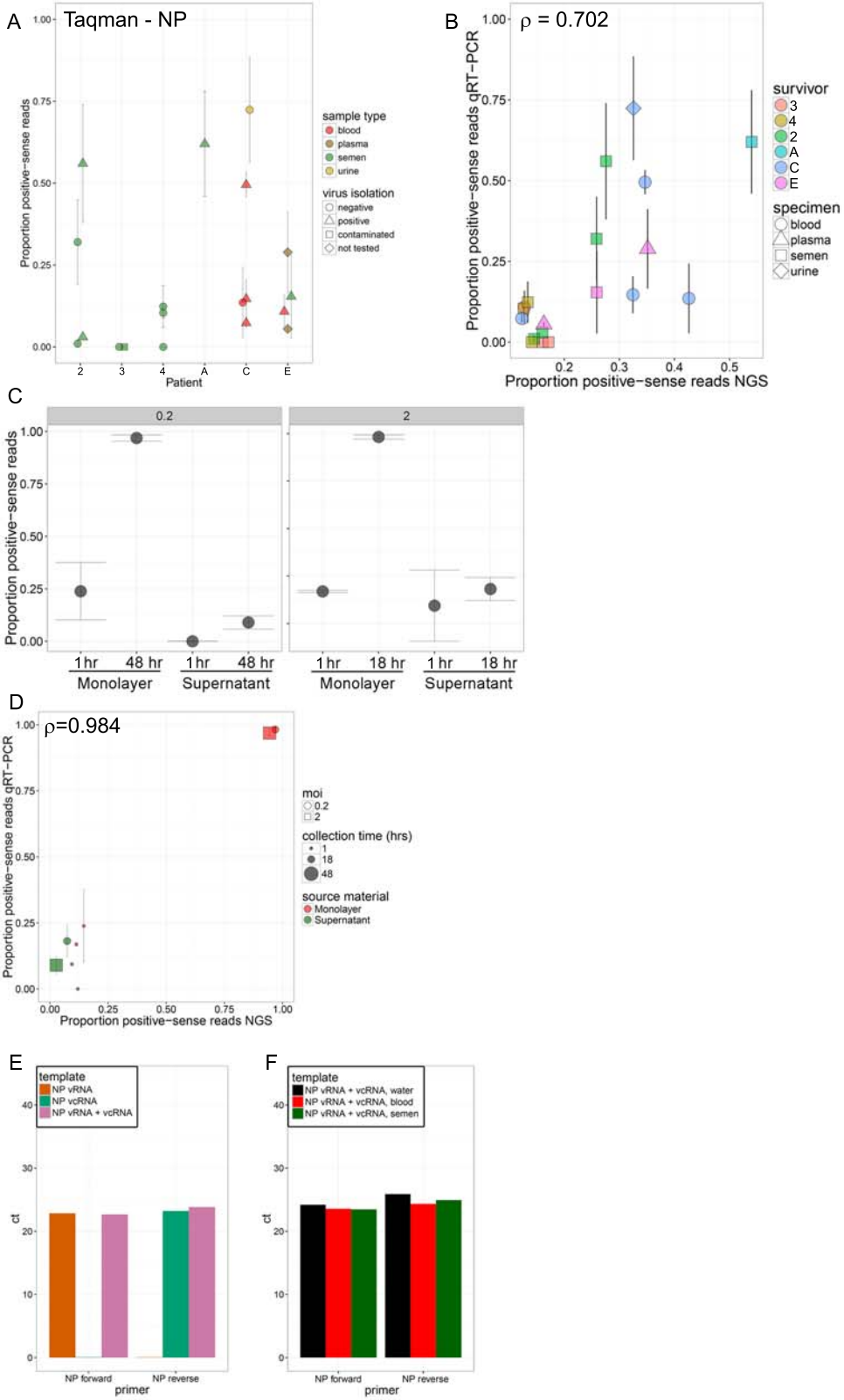

Figure S4: Supportive qRT-PCT evidence of active viral replication during persistent infection, Related to Figure 4. NGS strandedness assay results were further verified by performing strand-specific qRT-PCR with synthetic positive- and negative-sense RNA, RNA remaining from clinical specimens and RNA extracted from Huh7 cells infected with EBOV-ZsGreen. **A)** Proportion of EBOV NP-specific positive-sense reads from each EVD survivor specimen. Specimen types are highlighted with different colors and error bars indicate standard deviations between biological replicates. Virus isolation was attempted on most specimens,and point shape indicates virus isolation results. **B)** Correlation between the proportions of NP-specific positive-sense reads detected by stranded next-generation sequencing and qRT-PCR using RNA extracted from the semen of EVD survivors. Error bars indicate standard deviations in copy numbers detected by qRT-PCR between biological replicates. A positive monotonic relationship was detected as measured by Spearman's rank-order correlation (ρ=0.702). **C)** Proportion of EBOV NP-specific positive-sense reads from an in vitro infection of Huh7 cells done at an MOI of 0.2 (1 and 48hpi) or 2 (1 and 18hpi). Error bars indicate standard deviations in copy numbers detected by qRT-PCR between biological replicates. **D)** Correlation between the proportions of NP-specific positive-sense reads detected by stranded next-generation sequencing and qRT-PCR using RNA extracted from an in vitro infection of Huh7 cells. Error bars indicate standard deviations in copy numbers detected by qRT-PCR between biological replicates. A strongly positive monotonic relationship was detected as measured by Spearman's rank-order correlation (ρ=0.984). **E)** Specificity of NP stranded qRT-PCR assays. Specificity of forward and reverse qRT-PCT assays was assessed and confirmed using negative or positive-sense synthetic RNA, or a mixture of the two strands. **F)** Specificity of NP stranded qRT-PCR assay. Specificity of forward and reverse qRT-PCT assays was assessed and confirmed using negative and positive-sense synthetic RNA. Mixtures of synthetic RNA were spiked into water, or RNA extracted from normal human blood or semen.

Table S1: Evolutionary Rate Estimates from non-edited and edited SAVS from SLE and US EVD Survivors, Related to Figures 1 and 2.

SLE EVD SURVIVORS:

Clock Model	Rate Estimates (*10^3 subs/site/year) mean [95%HPD lower - upper]	
	Acute Rate	Latent Rate
No Edits	0.963 (0.863-1.066)	0.674 (0.504-0.846)
Hyper-edits removed	0.895 (0.805-0.991)	0.739 (0.580-0.904)

Clock Model	Rate Estimates (*10-3 subs/site/year) mean [95%HPD lower - upper]		
	Survivor 1	Survivor 2	Survivor 3
No Edits	0.500 (0.122-0.969)	0.631 (0.418-0.853)	0.862 (0.441-1.339)
Hyper-edits removed	0.631 (0.238-1.068)	0.710 (0.514-0.918)	0.856 (0.529-1.209)

Clock Model	Rate Estimates (*10-3 subs/site/year) mean [95%HPD lower - upper]		
	Survivor 4	Survivor 5	Survivor 6
No Edits	0.717 (0.391-1.071)	0.787 (0.230-1.502)	0.751 (0.298-1.261)
Hyper-edits removed	0.776 (0.498-1.042)	0.815 (0.367-1.334)	0.783 (0.383-1.225)

Clock Model - Uneditted	Loglikelihood Path Sampling		Loglikelihood Stepping Stone	
	Average	Standard Deviation	Average	Standard Deviation
Relaxed Clock (UCLN), Skygrid	*-34238.4*	*7.9*	*-34246.3*	*7.9*
Relaxed Clock (UCLN), constant population	-34359.0	4.4	-34363.1	5.1
Relaxed Clock (UCLN) - monophyly, constant population	-34331.5	4.4	-34340.3	5.1
Fixed Local Clock - monophyly, individual rates, constant population	-34427.0	10.1	-34434.7	6.5
Fixed Local Clock - monophyly, persistent rate, constant population	-34365.1	10.0	-34369.7	9.2

Clock Model - Hyper-edits removed	Loglikelihood Path Sampling		Loglikelihood Stepping Stone	
	Average	Standard Deviation	Average	Standard Deviation
Relaxed Clock (UCLN), Skygrid	*-33822.7*	*4.9*	*-33830.6*	*6.6*
Relaxed Clock (UCLN), constant population	-33939.6	3.7	-33946.7	3.8
Relaxed Clock (UCLN) - monophyly, constant population	-33901.5	7.3	-33912.0	13.6
Fixed Local Clock - monophyly, individual rates, constant population	-33928.6	2.2	-33940.2	4.7
Fixed Local Clock - monophyly, persistent rate, constant population	ND	ND	ND	ND

US EVD SURVIVORS:

Clock Model	Rate Estimates (subs/site/year) mean (95%HPD lower - upper)		
	Acute Rate	Blood Rate	Persistence Rate
No Edits	1.152 (1.043-1.267)	0.884 (0.518-1.290)	1.290 (0.903-1.713)
Hyper-edits removed	1.041 (0.937-1.138)	0.888 (0.577-1.235)	0.859 (0.617-1.133)

Clock Model	Rate Estimates (subs/site/year) mean (95%HPD lower - upper)	
	Survivor A - Acute	Survivor A - Semen
No Edits	1.037 (0.257-2.103)	1.378 (0.426-2.619)
Hyper-edits removed	0.979 (0.331-1.805)	0.816 (0.264-1.499)

Clock Model	Rate Estimates (subs/site/year) mean (95%HPD lower - upper)	
	Survivor E - Acute	Survivor E - Semen
No Edits	0.962 (0.530-1.475)	0.571 (0.132-1.137)
Hyper-edits removed	0.848 (0.361-1.419)	0.634 (0.187-1.174)

Clock Model	Rate Estimates (subs/site/year) mean (95%HPD lower - upper)		
	Survivor C - Acute	Survivor C - Semen and Urine	Survivor C - Eye
No Edits	0.962 (0.530-1.475)	1.431 (0.940-1.956)	0.903 (0.264-1.662)
Hyper-edits removed	0.948 (0.572-1.365)	0.916 (0.619-1.241)	0.930 (0.306-1.669)

Clock Model - uneditted	Loglikelihood Path Sampling		Loglikelihood Stepping Stone	
	Average	Standard Deviation	Average	Standard Deviation
Relaxed Clock (UCLN), Skygrid	*-38288.5*	*9.3*	*-38302.7*	*8.7*
Relaxed Clock (UCLN), constant population	-38456.4	8.8	-38470.6	8.9
Relaxed Clock (UCLN) - monophyly, constant population	-38406.7	5.7	-38420.0	8.9
Fixed Local Clock - monophyly, individual rates, constant population	-38621.5	12.5	-38643.5	24.2
Fixed Local Clock - monophyly, blood/persistent rates, constant population	-38547.7	1.7	-38561.8	9.8

Clock Model - hyper-edits removed	Loglikelihood Path Sampling		Loglikelihood Stepping Stone	
	Average	Standard Deviation	Average	Standard Deviation
Relaxed Clock (UCLN), Skygrid	*-37424.7*	*9.0*	*-37438.1*	*8.7*
Relaxed Clock (UCLN), constant population	-37578.0	8.5	-37591.7	8.5
Relaxed Clock (UCLN) - monophyly, constant population	-37503.8	9.8	-37529.9	16.4
Fixed Local Clock - monophyly, individual rates, constant population	-37680.3	12.0	-37696.0	5.5
Fixed Local Clock - monophyly, blood/persistent rates, constant population	ND	ND	ND	ND

Table S1: Evolutionary Rate Estimates from non-edited and edited SAVS from SLE and US EVD Survivors, Related to Figures 1 and 2.**(TOP)** Bayesian analysis conducted using UCLNunconstrained clock models with un-edited viral sequences and U-to-C hyper-edits removed from viral sequences. Marginal likelihood values from path sampling and stepping stone analysis with different clock models and prior tree assumptions (Relaxed UCLNunconstrained, Relaxed UCLNmonophyletic, Fixed local clockmonophyletic-individual rates, and Fixed local clockmonophyletic-latent rates) are included on lower half. **(BOTTOM)** Evolutionary Rate Estimates from non-edited and edited SAVS using AAVS and SAVS from US EVD Survivors. Bayesian analysis conducted using UCLNunconstrained clock models with un-edited viral sequences and U-to-C hyper-edits removed from viral sequences. Marginal likelihood values from path sampling and stepping stone analysis using different clock models and prior tree assumptions (Relaxed UCLNunconstrained, Relaxed UCLNmonophyletic, Fixed local clockmonophyletic-individual rates, and Fixed local clockmonophyletic-latent rates) are included on lower half.

Table S2: Evolutionary Preassure and iSNV Analysis, Related to Figure 3.

"branch models" (codeml) AAVS vs SAVS

Gene	2ΔlnL	Degrees of Freedom	P value (Bonferroni corrected)
NP	0.67	1	0.4118
VP35	3.04	1	0.0814
VP40	4.29	1	0.0384
NGP	0.01	1	0.9395
Mucin	0.05	1	0.8258
CGP	1.39	1	0.2377
SGP without p19 tail	16.06	1	**6.13E-05**
SGP with p19 tail	5.03	1	0.0249
VP30	0.00	1	0.9517
VP24	2.32	1	0.1276
RDRP	-61.67	1	0.0000

"branch models" (codeml) AAVS vs $SAVS_{fast}$ vs $SAVS_{slow}$

Gene	2ΔlnL	Degrees of Freedom	P value (Bonferroni corrected)
NP	0.73	1	0.3921
VP35	3.05	1	0.0806
VP40	4.01	1	0.0451
NGP	0.15	1	0.7030
Mucin	1.09	1	0.2957
CGP	2.40	1	0.1213
SGP without p19 tail	0.92	1	0.3382
SGP with p19 tail	5.03	1	0.0249
VP30	0.00	1	0.9495
VP24	2.32	1	0.1276
RDRP	-65.01	1	0.0000

"branch-site models" (codeml) AAVS vs SAVS

Gene	2ΔlnL	Degrees of Freedom	P value	Sites under positive selection (NEB)
NP	1.79	1	0.1812	101 E 0.940, 376 Y 0.939
VP35	0.75	1	0.3856	51 P 0.880
VP40	1.67	1	0.1958	131 Q 0.958*, 252 V 0.960*
NGP	5.40E-04	1	0.9815	
Mucin	1.14	1	0.2857	213 P 0.879, 264 T 0.879
CGP	24.91	1	**6.00E-07**	**296 N 0.999****
SGP without p19 tail	24.89	1	**6.06E-07**	**296 T 0.999**, 315 P 0.782**
SGP with p19 tail	24.935648	1	**5.93E-07**	**296 T 0.999**, 315 P 0.766**
VP30	0.01	1	0.9166	
VP24	0.71	1	0.3988	117 R 0.918
RDRP	-3.02E-03	1	0.0000	

Position	Gene	Major Variant	Minor Variant	Effect	Major Amino Acid	Minor Amino Acid
2263	NP	TC**C**	TC**T**	synonymous	S598	S
3833	VP35	TT**T**	TT**C**	synonymous	F235	F
4433	noncoding	C	T	N/A		
4886	VP40	AA**T**	AA**C**	synonymous	N136	N
4978	VP40	C**A**A	C**T**A	nonsynonymous	Q167	L
6602	GP - shared with FL and sGP	CA**A**	CA**G**	synonymous	Q188	Q
6924	GP - shared with FL and sGP	AAA AAA ACC CTC	AAA AAA ACC -TC A	nonsynonymous - results in frame shift from full length GP to sGP	KKTL, full length GP tail	KKTS, sGP tail
7246	full length GP	CA**A**	CA**G**	synonymous	Q108	Q
8371	noncoding	A	G	N/A		
12403	polymerase	**A**TG	**G**TG	nonsynonymous	M275	V
12568	polymerase	**G**CC	**A**CC	nonsynonymous	A330	T
12750	polymerase	AA**A**	AA**G**	synonymous	K390	
13211	polymerase	C**A**A	C**G**A	nonsynonymous	Q544	R
14411	polymerase	G**A**G	G**G**G	nonsynonymous	E944	G
16821	polymerase	TC**A**	TC**G**	synonymous	S1747	S
16928	polymerase	**A**CC	**A**TC	nonsynonymous	T1783	I

Table S2: Evolutionary Preassure and iSNV Analysis, Related to Figure 3. **(TOP)** Likelihood ratio test statistics from PAML branch- and branch-site models. **(BOTTOM)** Effect of iSNV's from SLE Survivor 2 on viral coding and noncoding regions.

Table S3: Chimeric Reads from Sierra Leone and US EVD Survivors, and Cell Culture *in vitro* Infections, Related to Figure 4.

SIERRA LEONE EVD SURVIVORS:

Survivor:	Days post Onset:	Specimen Number:	Chimera Type:	#Unique Deletions:	#Reads Chimeric:	Avg. # Reads Per Chimera:	Standard Deviation Reads Per Chimera	Total # Mapped Reads:	Proportion Mapped Chimeric Reads:
3	252	VP1201500050	Deletions	11	57	5.1818	4.9326	225885	0.0003
3	252	VP1201500050	SmallDups	4	14	3.5	1.5	225885	0.0001
3	252	VP1201500050	LargeDups	8	33	4.125	4.3714	225885	0.0001
3	252	VP1201500050	CopyBacks	1	4	4	0	225885	0
3	259	VP1201500100	Deletions	9	47	5.2222	5.0723	2210508	0
3	259	VP1201500100	LargeDups	13	44	3.3846	2.1318	2210508	0
3	259	VP1201500100	CopyBacks	2	2	1	0	2210508	0
3	294	VP1201500247	Deletions	2	21	10.5	6.5	45989	0.0005
3	294	VP1201500247	SmallDups	2	29	14.5	4.5	45989	0.0006
3	294	VP1201500247	LargeDups	1	2	2	0	45989	0
3	322	VP1201500357	Deletions	2	23	11.5	9.5	24405	0.0009
3	322	VP1201500357	LargeDups	1	4	4	0	24405	0.0002
4	143	VP1201500033	Deletions	3	97	32.3333	28.1227	397245	0.0002
4	143	VP1201500033	SmallDups	1	20	20	0	397245	0.0001
4	143	VP1201500033	LargeDups	7	198	28.2857	27.7731	397245	0.0005
4	143	VP1201500033	CopyBacks	2	54	27	24	397245	0.0001
4	157	VP1201500118	Deletions	1	15	15	0	328497	0
4	157	VP1201500118	SmallDups	6	150	25	21.7486	328497	0.0005
4	157	VP1201500118	LargeDups	3	71	23.6667	30.6522	328497	0.0002
4	172	VP1201500193	Deletions	12	53	4.4167	3.0127	310225	0.0002
4	172	VP1201500193	SmallDups	7	37	5.2857	3.3685	310225	0.0001
4	172	VP1201500193	LargeDups	23	121	5.2609	5.4231	310225	0.0004
4	172	VP1201500193	CopyBacks	3	64	21.3333	18.625	310225	0.0002
4	185	VP1201500235	Deletions	6	10	1.6667	1.1055	5016545	0
4	185	VP1201500235	SmallDups	27	154	5.7037	4.8672	5016545	0
4	185	VP1201500235	LargeDups	28	215	7.6786	10.1526	5016545	0
4	185	VP1201500235	CopyBacks	5	43	8.6	7.3103	5016545	0
4	199	VP1201500293	SmallDups	1	28	28	0	19305	0.0015
5	169	VP1201500132	Deletions	7	35	5	9.396	3363725	0
5	169	VP1201500132	SmallDups	2	43	21.5	1.5	3363725	0
5	169	VP1201500132	LargeDups	3	4	1.3333	0.4714	3363725	0
5	169	VP1201500132	CopyBacks	6	6	1	0	3363725	0
6	178	VP1201500297	SmallDups	1	2	2	0	28892	0.0001
6	178	VP1201500297	LargeDups	2	11	5.5	4.5	28892	0.0004
2	82	VP1201500009	Deletions	2	11	5.5	0.5	204315	0.0001
2	82	VP1201500009	SmallDups	1	11	11	0	204315	0.0001
2	82	VP1201500009	LargeDups	6	59	9.8333	8.1938	204315	0.0003
2	82	VP1201500009	CopyBacks	1	1	1	0	204315	0
2	96	VP1201500046	Deletions	14	76	5.4286	5.0244	8123282	0
2	96	VP1201500046	SmallDups	7	49	7	10.1419	8123282	0
2	96	VP1201500046	LargeDups	20	111	5.55	5.6963	8123282	0
2	96	VP1201500046	CopyBacks	14	81	5.7857	8.6204	8123282	0
2	103	VP1201500084	Deletions	35	748	21.3714	31.5722	1153901	0.0006
2	103	VP1201500084	SmallDups	71	1262	17.7746	28.1085	1153901	0.0011
2	103	VP1201500084	LargeDups	43	682	15.8605	17.2742	1153901	0.0006
2	103	VP1201500084	CopyBacks	72	1208	16.7778	44.6555	1153901	0.001
2	116	VP1201500163	Deletions	7	78	11.1429	7.8272	550976	0.0001
2	116	VP1201500163	SmallDups	6	63	10.5	9.4472	550976	0.0001
2	116	VP1201500163	LargeDups	4	93	23.25	37.963	550976	0.0002
2	116	VP1201500163	CopyBacks	1	3	3	0	550976	0
2	158	VP1201500320	Deletions	38	80	2.1053	1.5181	5205496	0
2	158	VP1201500320	SmallDups	29	154	5.3103	8.9022	5205496	0
2	158	VP1201500320	LargeDups	43	379	8.814	37.1911	5205496	0.0001
2	158	VP1201500320	CopyBacks	14	399	28.5	70.5111	5205496	0.0001
2	172	VP1201500374	Deletions	12	319	26.5833	56.1508	5339682	0.0001
2	172	VP1201500374	SmallDups	16	1187	74.1875	90.4898	5339682	0.0002
2	172	VP1201500374	LargeDups	20	573	28.65	35.1885	5339682	0.0001
2	172	VP1201500374	CopyBacks	11	152	13.8182	38.6472	5339682	0
2	186	VP1201500423	Deletions	12	69	5.75	5.4333	3188103	0
2	186	VP1201500423	SmallDups	15	105	7	5.379	3188103	0
2	186	VP1201500423	LargeDups	13	79	6.0769	6.9222	3188103	0
2	186	VP1201500423	CopyBacks	13	42	3.2308	4.509	3188103	0

US EVD SURVIVORS:

Survivor:	Days post Onset:	Specimen Type:	Specimen Number:	Chimera Type:	#Unique Deletions:	#Reads Chimeric:	Avg. # Reads Per Chimera:	Standard Deviation Reads Per Chimera	Total # Mapped Reads:	Proportion Mapped Chimeric Reads:
A	28	semen	201403120	Deletions	1	4	4	0	145785	0
A	28	semen	201403120	SmallDups	2	33	16.5	1.5	145785	0.0002
A	28	semen	201403120	LargeDups	8	57	7.125	3.8871	145785	0.0004
A	58	semen	201403184	LargeDups	3	266	88.6667	62.4304	192147	0.0014
C	5	blood	201403131	Deletions	42	84	2	1.291	606711	0.0001
C	5	blood	201403131	SmallDups	45	123	2.7333	2.2549	606711	0.0002
C	5	blood	201403131	LargeDups	57	142	2.4912	1.6975	606711	0.0002
C	5	blood	201403131	CopyBacks	13	13	1	0	606711	0
C	7	blood	201403142	Deletions	71	102	1.4366	1.1596	583227	0.0002
C	7	blood	201403142	SmallDups	82	107	1.3049	0.7104	583227	0.0002
C	7	blood	201403142	LargeDups	134	175	1.306	0.7353	583227	0.0003
C	7	blood	201403142	CopyBacks	34	36	1.0588	0.3379	583227	0.0001
C	9	blood	201403147	Deletions	31	57	1.8387	0.8461	524944	0.0001
C	9	blood	201403147	SmallDups	68	129	1.8971	2.3272	524944	0.0002
C	9	blood	201403147	LargeDups	101	168	1.6634	0.9676	524944	0.0003
C	9	blood	201403147	CopyBacks	17	18	1.0588	0.2353	524944	0
C	12	blood	201403162	Deletions	13	23	1.7692	1.2499	91224	0.0003
C	12	blood	201403162	SmallDups	7	7	1	0	91224	0.0001
C	12	blood	201403162	LargeDups	33	39	1.1818	0.3857	91224	0.0004
C	12	blood	201403162	CopyBacks	1	1	1	0	91224	0
C	27	urine	201403234	Deletions	11	40	3.6364	3.7725	1201902	0
C	27	urine	201403234	SmallDups	17	452	26.5882	61.5296	1201902	0.0004
C	27	urine	201403234	LargeDups	38	590	15.5263	42.5989	1201902	0.0005
C	27	urine	201403234	CopyBacks	1	1	1	0	1201902	0
C	33	urine	201403258	SmallDups	1	7	7	0	877878	0
C	45	semen	201403360	Deletions	2	2	1	0	837349	0
C	45	semen	201403360	SmallDups	4	4254	1063.5	1832.2228	837349	0.0051
C	45	semen	201403360	LargeDups	4	4832	1208	2064.102	837349	0.0058
C	72	semen	201403439	Deletions	1	2	2	0	15350	0.0001
C	72	semen	201403439	SmallDups	1	4	4	0	15350	0.0003
C	72	semen	201403439	LargeDups	1	3	3	0	15350	0.0002
C	101	eye	201403522	Deletions	108	151	1.3981	0.8044	1264908	0.0001
C	101	eye	201403522	SmallDups	92	208	2.2609	3.4385	1264908	0.0002
C	101	eye	201403522	LargeDups	158	233	1.4747	1.8235	1264908	0.0002
C	101	eye	201403522	CopyBacks	2	3	1.5	0.5	1264908	0
C	117	semen	201403557	SmallDups	1	5	5	0	7479	0.0007
E	1	blood	201403368	Deletions	1	1	1	0	197384	0
E	1	blood	201403368	LargeDups	2	54	27	6	197384	0.0003
E	1	blood	201403368	CopyBacks	1	1	1	0	197384	0
E	2	plasma	201403391	Deletions	23	75	3.2609	2.6738	631517	0.0001
E	2	plasma	201403391	SmallDups	21	101	4.8095	4.0897	631517	0.0002
E	2	plasma	201403391	LargeDups	43	149	3.4651	2.5183	631517	0.0002
E	2	plasma	201403391	CopyBacks	8	8	1	0	631517	0
E	5	plasma	201403394	Deletions	5	136	27.2	18.7553	297798	0.0005
E	5	plasma	201403394	SmallDups	9	177	19.6667	19.1079	297798	0.0006
E	5	plasma	201403394	LargeDups	12	128	10.6667	10.3789	297798	0.0004
E	5	plasma	201403394	CopyBacks	1	1	1	0	297798	0
E	50	semen	201403509	Deletions	3	9	3	1.4142	58765	0.0002
E	50	semen	201403509	SmallDups	7	11	1.5714	0.9035	58765	0.0002
E	50	semen	201403509	LargeDups	8	12	1.5	1	58765	0.0002
E	50	semen	201403509	CopyBacks	1	2	2	0	58765	0

CELL CULTURE *in vitro* INFECTION

Time Point:	MOI:	Chimera Type:	#Unique Deletions:	#Reads Chimeric:	Avg. # Reads Per Chimera:	Standard Deviation Reads Per Chimera:	Total # Mapped Reads:	Proportion Mapped Chimeric Reads:
1 hr	2.0	Deletions	23	298	12.9565	48.7981	159328	0.0019
1 hr	2.0	SmallDups	22	32	1.4545	0.6556	159328	0.0002
1 hr	2.0	LargeDups	16	20	1.25	0.75	159328	0.0001
1 hr	2.0	CopyBacks	2	3	1.5	0.5	159328	0
18 hr	2.0	Deletions	150	2865	19.1	149.923	438530	0.0065
18 hr	2.0	SmallDups	76	87	1.1447	0.622	438530	0.0002
18 hr	2.0	LargeDups	79	95	1.2025	0.5126	438530	0.0002
18 hr	2.0	CopyBacks	58	58	1	0	438530	0.0001
1 hr	0.2	Deletions	4	13	3.25	3.3448	14540	0.0009
1 hr	0.2	SmallDups	3	6	2	1.4142	14540	0.0004
1 hr	0.2	LargeDups	1	1	1	0	14540	0.0001
48 hr	0.2	Deletions	101	581	5.7525	30.2145	401382	0.0014
48 hr	0.2	SmallDups	82	96	1.1707	0.5589	401382	0.0002
48 hr	0.2	LargeDups	99	105	1.0606	0.2386	401382	0.0003
48 hr	0.2	CopyBacks	143	149	1.042	0.2005	401382	0.0004

SUPPLEMENTARY EXPERIMENTAL METHODS

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for reagents may be directed to, and will be fulfilled by the corresponding authors, Ute Ströher (ute.stroeher@gmail.com) and Gustavo Palacios (gustavo.f.palacios.ctr@mail.mil).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Human Subjects

Through the joint Sierra Leone Ebola Virus Persistence study (SLEVPS) with the Ministry of Health and Sanitation (MoHS) in Sierra Leone, WHO, China-CDC, and CDC, we had access to semen specimens collected from EVD survivors (Deen et al., 2015). Through this study we did not have access to direct patient data, such as patient age. Male study participates were stratified and selected for sequencing based on their NP Ct value and number/time span of serial semen specimens. As the SLEVPS only focused on specimen collection from EVD survivors, we did not have access to acute specimens from these participants. The SLEVPS was reviewed and approved by the Sierra Leone Institutional Review Board and the World Health Organization Ethical Review Committee. Following clinical diagnostic testing in the US, we did have access to paired acute blood and persistent semen specimens collected from US EVD patients. Acute and persistent specimens from US EVD survivors were collected by their treating physicians and transported to the CDC for detection of viral RNA (Kraft et al., 2015; Lyon et al., 2014; McElroy et al., 2015; Varkey et al., 2015). This sequencing project was determined by the CDC institutional human subject advisor to be a non-research public health response activity, and institutional review board review was not required.

METHOD DETAILS

Whole Genome Sequencing and Bioinformatics

RNA was extracted from blood and semen specimens using MagMAX Pathogen RNA/DNA isolation kit (Invitrogen) and BeadRetriever (Invitrogen) and treated with recombinant DNase I RNase-free (Roche). Ribosomal and carrier RNA were removed as previously described (Matranga et al., 2014). Non-depleted and rRNA/carrier RNA depleted specimens were prepared for sequencing using a modified version of the Illumina TruSeq RNA Access Library Prep kit as described previously with some minor variations (Blackley et al., 2016; Levin et al., 2010; Mate et al., 2015; Parkhomchuk et al., 2009; Sultan et al., 2012; Wang et al., 2011). RNA was fragmented for one minute prior to cDNA synthesis and custom dual indexes were used to avoid any sequencer bleed-through (Kircher et al., 2012). All specimens were enriched separately to avoid any bias of enriching one or two libraries over others in a pool. Specimens were sequenced using a Illumina MiSeq (version 3, 2x151 cycles), an Illumina Nextseq500 (midoutput kit, 2x151 cycles), and an Illumina HiSeq 2500 (rapid run v2, 2x151 cycles).

EBOV genomes were assembled by aligning reads to Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3864.1 (KR013754, missing bases in the reference were replaced with consensus calls from complete EBOV genomes); this reference is equivalent to the basal SL2 haplotype (Gire et al., 2014). The priming sites of the random hexamer and Illumina TruSeq adaptors were removed from the sequencing reads using Cutadapt v1.21 (Martin, 2011) and low quality reads/bases were filtered using Prinseq-lite v0.20.4 (-min_qual_mean 25 -trim_qual_right 20 -min_len 50) (Schmeider, 2011). Reads were aligned to the reference using Bowtie2 (Langmead and Salzberg, 2012), duplicates were removed with Picard (broadinstitute.github.io/picard) and a new consensus was generated using a combination of Samtools v0.1.18 (Li et al., 2009) and custom scripts. Only bases with Phred quality score ≥20 were utilized in consensus calling, and a minimum of 3x read-depth coverage, in support of the consensus, was required to make a call; positions lacking this depth of coverage were treated as missing (i.e., called as 'N'). Genomes acquired from clinical specimens were deposited into Genbank: KY401638-KY401675, KY805810-2.

Analysis of Viral Evolutionary Rates

Viral evolutionary rate estimates were conducted using both linear regression modeling and time-structured phylogenies. For SAVS from SLE survivors, 1,058 EBOV genomes from Sierra Leone were analyzed using Path-O-Gen (now called TempEst (Rambaut, 2016)) and a maximum likelihood tree (GTR+G) rooted on the earliest available Sierra Leone sequence. For SAVS from US EVD survivors, 1498 genomes, representing a majority of sequences from Sierra Leone, Guinea and Liberia, were analyzed using Path-O-Gen (now called TempEst (Rambaut, 2016)) and a maximum likelihood tree (GTR+G) rooted on the earliest available Guinea sequence. Evolutionary rates and residual density plots were analyzed using R and custom python scripts from (Park et al., 2015). Evolutionary rate estimates for SAVS were also obtained using BEAST/v1.8.2,1.8.3, 1.8.4 (Drummond et al., 2012). A random selection of viral sequences, representing 25% of available sequences from SLE, or SLE/LBR/GIN, were used for the Bayesian analysis by partitioning into concatenated coding and noncoding sites. Rate estimates were modeled using unlinked HKY nucleotide evolutionary models with 4-independent Γ distributions, Bayesian skygrid demographic model (with variable population model estimated between January 1, 2014 and January 1, 2016, ie – "Time at last point:2"; or constant population, ie – "Time at last point:0"), and fixed local clock (Yoder and Yang, 2000) or uncorrelated lognormal local clock (Drummond et al., 2006) set with an initial prior of $1.1*10^{-3}$ subs/site/year. Model comparisons were conducted using: 1) relaxed uncorrelated lognormal clock with no constraints on the tree prior, variable Skygrid population; 1) relaxed uncorrelated lognormal clock with no constraints on the tree prior, constant population: "$UCLN_{unconstrained}$"; 2) relaxed uncorrelated lognormal clock with individual survivor blood and/or semen sequences constrained to survivor-specific monophyletic blood/semen taxons, constant population: "$UCLN_{monophyletic}$", 3) Fixed local clock with individual survivor blood and/or semen sequences constrained to survivor-specific monophyletic blood/semen taxons, constant population: and 4) Fixed local clock with survivor blood and/or semen sequences

constrained to blood-specific and semen-specific taxons, constant population: "$FLC_{monophyletic}$". The MCMC analysis was conducted for 800 million generations, which represents a compilation of 8-independent replicates of 100 million generations (sampled every $10{,}000^{th}$ state). Convergence was obtained for the majority of replicates and burn-in was removed (usually 5-10% of total states) by examining the trace and effective sample size statistics (min ESS > 200 for all models) using tracer/v1.6. Strength of model fit was evaluated by performing path- and stepping stone-sampling with default values and best-of-fit was evaluated by calculating Bayes Factors. Survivor and acute rate estimates from Bayesian analysis conducted with the UCLN clock models were estimated using custom-modified samogitia.py scripts (Dudas, 2017).

Sequence Analysis

Additional sequence analysis was conducted using CLC Genomics/v9.0. Potential hyper-edited sites due to host-encoded adenosine deaminases acting on RNA (ADARs) can result in the rapid accumulation of clustered T(U)-to-C substitutions (on the positive strand) in the EBOV genome (Dudas et al., 2017). We identified clusters of substitutions consistent with ADAR-mediated editing (≥3 phylogenetically-linked T(U)-to-C substitutions within a 200 nt window), and these substitutions were masked for evolutionary rate analyses (i.e., C genotypes were converted to T at these positions) in Figure 2A-B, and Figure 3A-B. Histograms of U-to-C hyper-editing were generated using R. Median joining networks were constructed using sequence alignments from each EVD survivor with PopART/v1.7.2. Intrahost variants (iSNVs) were detected with FreeBayes v1.0.2 (Garrison, 2012). For iSNV detection, we only used reads with mapping quality ≥30 and positions with base quality ≥30. An iSNV was only considered if the alternate allele was represented by ≥5 reads and present at a frequency ≥3%. We estimated SNV and insertion frequencies for the longitudinal phasing analysis by first performing a read-pair merging of the assemblies in IRMA v0.6.5 (Shepard et al., 2016) and computing allele frequencies for each selected site using IRMA's *call.pl* script (–B option). A pairwise (Manhattan)

distance matrix was computed in R v3.3.1 for each position-allele combination with the vector of the observed frequencies ordered by specimen date. The matrix was used to generate a single linkage dendrogram, also in R. SNVs were divided into two clusters based on two near-symmetrical branches in the tree. The insertion frequency of C at upstream position 6924 and its complement were added to the dataset and a second dendrogram produced. The 6924 C-insert and 6924 non-C-insert frequencies were assigned an SNV cluster according to their nearest neighboring SNV in the second tree: 8371G and 8371A respectively. A final tree and distance matrix was produced for each variant position by ordering the frequency vectors by specimen date as well as variant cluster (one allele or insertion state was assigned to each cluster for each site). Frequency line graphs of positions, alleles, and specimen dates were created using Tableau v10.0 and the positions in the graph ordered and composited with the final R dendrogram in Supplementary Figure 3. Identification of chimeric reads were performed by mapping reads to Ebola virus/H.sapiens-wt/SLE/2014/Makona-G3864.1 (KR013754, missing bases in the reference were replaced with consensus calls from complete EBOV genomes) using bwa and chimeric reads were further defined using custom scripts. US EVD survivors received multiple thereapeutic treatments (A: whole blood transfusion, convalescent whole blood transfusion, ZMAPP; C: TKM-Ebola, convalescent plasma; E: Convalescent plasma, brincidofovir)(Kraft et al., 2015; Lyon et al., 2014; McElroy et al., 2015; Varkey et al., 2015) and we confirmed that viral regions targeted by these compounds (GP, VP35, polymerase) did not mutate through comparison of serial consensus viral sequences.

Estimation of Selective Pressure

Evolutionary selective pressures were estimated using the renaissance counting method in beast/v1.8.2 (Lemey et al., 2012; O'Brien et al., 2009) with a subset of genomes representing 25% of available random sequences from Sierra Leone, Guinea, and Liberia that did not contain codon frame shifts.

123 Codon alignments for each gene were partitioning into coding and concatenated total noncoding sites.
124 Rate estimates were modeled using unlinked HKY nucleotide substitution models, Bayesian skygrid
125 demographic model, and uncorrelated lognormal relaxed clock set with an initial prior of $1.1*10^{-3}$
126 subs/site/year. The MCMC analysis was conducted for 400 million generations for each gene, which
127 represents a compilation of 4-independent replicates of 100 million generations (sampled every 1000^{th}
128 state). Due to time constraints, MCMC analysis for the VP40 and polymerase gene were stopped at
129 ~200 million or ~120 million iterations, which easily reached convergence. For all replicates,
130 convergence was obtained and burn-in was removed (usually 10% of total states) by examining the trace
131 and effective sample size statistics (>200 for all MCMC analyses) using tracer/v1.6. Only one MCMC
132 replicate for the CGP tail did not converge, and it was removed from additional analysis. Omega
133 estimates were calculated by using the conditioned and unconditioned N and S estimates and equation 1
134 ((total_N/total_S) / (unconditioned_N/unconditioned_S)) from Lemey *et al.* (Lemey et al., 2012) and
135 scripts from Park *et al*. (Park et al., 2015). To prevent rate overestimation by double-counting shared
136 amino acids, the glycoprotein was split at the transcriptional editing site (nucleotide 6923) into N-
137 terminal (nucleotides 6039-6923, "NGP"), C-terminal full length (nucleotides 6923-8068 - containing
138 the GP1 carboxy-terminus and GP2, "CGP") and secreted GP (nucleotides 6924-7157, "SGP_c"). For
139 secreted GP (nucleotides 6924-7157, "SGP_c") rate estimates, approximately 9.6% of unconditioned S
140 estimates and 0.2% of unconditioned N estimates were 0.0; thus to bypass undefined ω estimates these
141 values were converted to 1. For polymerase rate estimates, approximately 3% of N or S estimates were
142 undefined (NaN) and to bypass undefined ω estimates these states were removed from the analysis.
143
144 Selective pressure hypothesis testing was performed using the codeml model in paml/v4.5 with a subset
145 of 231 genomes, representing approximately 25% of available random sequences from Sierra Leone,
146 Guinea and Liberia that did not contain reading frame shifts. We constructed a Maximum Clade

147 Credibility tree using beast/v.1.8.2 by partitioning the alignments into concatenated coding and
148 noncoding sites and trees were modeled using unlinked HKY nucleotide substitution models, Bayesian
149 skygrid demographic model, and uncorrelated lognormal relaxed clock set with an initial prior of
150 $1.1*10^{-3}$ subs/site/year. The MCMC analysis was conducted for 50 million generations (sampled every
151 1000^{th} state), which easily reached convergence. The cladogram of the MCC tree was used as input for
152 paml codeml. Branch model testing was performed using model0 and model2 and branch-site testing
153 was performed using modelA and A_null with codon frequencies F3x4. For branch testing, kappa and
154 omega estimates from model0 were set as initial estimates for model2 (acute sequences vs. SAVS) and
155 model2 (acute sequences vs. SAVS_acute_rate vs. SAVS_slow_rate). Strength of statistical support for
156 models2 (alternative hypotheses) vs. model0 (null hypothesis) was measured using the 2Δlog-likelihood
157 method with degrees of freedom=1 and further corrected according to Bonferroni (p = 0.05/ 2 tests
158 conducted with same sequence alignment) (Anisimova and Yang, 2007; Yang, 2007). The ratio of N to
159 S was calculated by summing the total N (N*dN from PAML model2 output) and S (S*dS from PAML
160 model2 output) estimates for all acute and SAVS branches and dividing by the total N and S count. For
161 branch-site testing, semen-specific branches were set as foreground branches and modelA was
162 performed using NSsites=2 with kappa and omega estimates set at initial values from model0.
163 ModelA_null testing was performed with NSsites=2, kappa and omega estimates set at initial values
164 from model0, and omega fixed at 1. Significance values were calculated using the 2Δlog-likelihood
165 method and significance was established with p values below 0.05.

166

167 *Ebola virus* in vitro *Infection*

168 All work with EBOV was performed in a biosafety level 4 (BSL-4) facility. Huh7 cells were cultured in
169 Dulbecco's modified Eagle's medium high glucose (DMEM) (item number 11960-044, Invitrogen)
170 supplemented with 10% heat inactivated HyClone fetal bovine serum (Thermo Scientific), 1x non-

essential amino acids (Invitrogen), 1x penicillin-streptomycin (Invitrogen), and 1x Glutamax (Invitrogen) at 37°C with 5% CO_2. Prior to viral infection, cells were seeded into triplicate wells in a 12-well plate and media was replaced with FluorBrite Dulbecco's modified Eagle's medium (DMEM) (item number A1896701, Invitrogen) supplemented with 10% heat inactivated HyClone fetal bovine serum (Thermo Scientific), 1x non-essential amino acids (Invitrogen), 1x penicillin-streptomycin (Invitrogen), and 1x Glutamax (Invitrogen). Immediately before infection cells were counted using the Moxi Z cell counter with M cassettes (Orflo, Technologies). Cells were infected at MOI of 2 or 0.2 with rEBOV-L2014/ZsG (Albarino et al., 2016) in 200uL of FluorBrite media with supplements for 1 hour at 37°C with 5% CO_2. After absorption, inoculum was removed and cells were washed twice with 1mL of PBS. Media was replaced with FluorBrite media for infection duration. At specified time points supernatant and monolayers were inactivated with TriPure isolation reagent (Roche). Prior to inactivation supernatants were spun at 200xg for 10 minutes to remove cellular debris. RNA was extracted from Tripure using the Direct-zol-96 MagBead RNA isolation kit (Zymo Research). Active viral infection of cells was confirmed by visualization of ZsGreen fluorescence at 1, 18, and 48 hours post infection.

Ebola virus RNA Strandedness Analysis

The TruSeq RNA Access Library Prep kit results in stranded data (i.e., read 1 is complementary to the original RNA molecule). Using custom scripts we quantified the proportion of positive- and negative-sense RNA molecules present in each specimen. Independently for each strand and each specimen, we also calculated relative depth of coverage for every EBOV ORF as

$$\widehat{D_j} = \frac{\frac{1}{n_j}\sum_{i=1}^{n_j} D_{ji}}{\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sum_{i=1}^{n_j} D_{ji}}$$

193 N is the total number of ORFs, n_j is the length in nucleotides of the j[th] ORF and D_{ji} is the read 1 depth
194 of coverage at the i[th] nucleotide of the j[th] ORF. Only read 1 was used to avoid double counting at
195 positions where reads 1 and 2 overlap. Regions of the genome included in multiple ORFs were
196 excluded. In Figure 4, strandedness data from next generation sequencing is presented either for 1)
197 specific genes (Figures 4D, Supplementary Figures 5F and 5G), 2) whole genome sequences (Figures
198 4A-C, and Supplementary Figures 5A-C), or 3) for the NP-specific open reading frame (Supplementary
199 Figures 6A-D). Cutoff criteria was ≥50x average paired end coverage across the genome for Figures
200 4A-C and Supplementary Figures 5A-C, ≥25x average read 1 coverage across the coding portions of the
201 genome (Figure 4D and Supplementary Figure 5F), and ≥25x average read 1 coverage across NP
202 (Supplementary Figure 6A-D).

203 The EBOV NP strand-specific qRT-PCR assay was performed by using separate first and second
204 strand reactions. The first-strand reaction was conducted with 2.5uL of input RNA, 1uL 10mM dNTPs
205 (Invitrogen), 1uL of 2uM gene-specific tagged stranded primer and 5uL of nuclease-free water
206 (Ambion). This mixture was heated to 65°C for 5 minutes and placed on ice for 2 minutes. The reverse
207 transcription reaction followed with 4uL of 5x first-strand reaction buffer (Invitrogen), 1uL SUPERase-
208 In (Invitrogen), 1uL superscript III reverse transcriptase (Invitrogen), 1uL 0.1M DTT (Invitrogen) and
209 3.5 uL of nuclease-free water (Ambion). The reaction was heated at 55°C for 15 minutes and cooled on
210 ice for 2 minutes. First strand reactions were cleaned with the QiaQuick PCR cleanup kit (Qiagen) and
211 ssDNA was eluted with 30uL of nuclease-free water (Ambion). The second strand reaction proceeded
212 with 5uL of input cDNA, 2.5uL of AmpliTaq 10x buffer I, 0.5uL of 10mM dNTP's (Invitrogen), 2.25uL
213 of 10uM tag-specific primer, 2.25uL of 10uM gene-specific primer, 0.625uL of 10uM NP probe,
214 0.125uL of AmpliTaq DNA polymerase (Invitrogen), and 11.75uL of nuclease-free water (Ambion).
215 Thermocycler conditions consisted of 50°C for 15 minutes, 95°C for 2 minutes, 95°C for 15s and 55°C
216 for 45s (44 cycles). To convert Ct values into strand copy numbers, we established a Ct versus molarity

concentration curves for both positive- and negative-sense synthetic RNA's. Goodness-of-fit values for these curves (r^2) were all greater than 0.988. Using the same first-strand cDNA products, we also established Ct versus copy number using the Bio-Rad QX200 digital droplet PCR and r^2 values for these curves were all greater than 0.979. Reaction conditions for ddPCR consisted of 10uL of 2x ddPCR Supermix for Probes (Bio-Rad), 1.8uL of 10uM tag-specific primer, 1.8uL of 10uM gene-specific primer, 0.5uL of 10 uM NP-specific probe, uL of cDNA, and 0.9uL of nuclease-free water (Ambion). Thermocycler conditions consisted of 95°C for 10 minutes, 94°C for 30s, 60°C for 1 minute (39 cycles), and 98°C for 10 minutes with a ramp speeds done at 2°C/sec. Final Ct to copy number conversions for *in vitro* infections and EVD survivors clinical specimens were calculated using the Ct versus molarity concentration curves corrected for copy numbers as estimated using ddPCR.

QUANTIFICATION AND STATISTICAL ANALYSIS

For evolutionary rate estimates using Bayesian analysis we present the mean and 95% highest posterior density estimates calculated from all total combined states (after removal of burn-in, in most cases 10%) using scripts from Park et al. and custom-modified samogitia.py scripts (Dudas, 2017; Park et al., 2015). Evolutionary rates estimates from RTT's are presented as the line of best fit with 95% confidence intervals shaded in grey. Residual comparisons from linear regressions display the 2-fold standard deviations of the acute residual density in grey. Strength of statistical support for paml estimation of selective pressure was measured using the likelihood ratio test with degrees of freedom=1 comparing model0 (null hypothesis) with model2 (alternative hypotheses). Significance values for modelA and modelA_null branch-site testing with PAML were calculated using the 2Δlog-likelihood method and significance was established with p values below 0.05. A one-way analysis of variance (ANOVA) for the association of proportion of positive-sense reads or NP Ct values vs. virus isolation result was assessed using the lm() and anova() functions from R v3.3.1.

DATA AND SOFTWARE AVAILABILITY

Most software utilized is freely available, and when possible we include the version number and reference for the software used. Custom scripts have been submitted to github (https://github.com/jtladner/Scripts and https://github.com/evk3/EBOV_semen_sequencing). Genomes acquired from clinical specimens were deposited into Genbank: KY401638-KY401675 and KY805810-2.

REFERENCES:

Albarino, C.G., Guerrero, L.W., Chakrabarti, A.K., Kainulainen, M.H., Whitmer, S.L., Welch, S.R., and Nichol, S.T. (2016). Virus fitness differences observed between two naturally occurring isolates of Ebola virus Makona variant using a reverse genetics approach. Virology *496*, 237-243.

Anisimova, M., and Yang, Z. (2007). Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites. Mol Biol Evol *24*, 1219-1228.

Blackley, D.J., Wiley, M.R., Ladner, J.T., Fallah, M., Lo, T., Gilbert, M.L., Gregory, C., D'Ambrozio, J., Coulter, S., Mate, S.*, et al.* (2016). Reduced evolutionary rate in reemerged Ebola virus transmission chains. Sci Adv *2*, e1600378.

Deen, G.F., Knust, B., Broutet, N., Sesay, F.R., Formenty, P., Ross, C., Thorson, A.E., Massaquoi, T.A., Marrinan, J.E., Ervin, E.*, et al.* (2015). Ebola RNA Persistence in Semen of Ebola Virus Disease Survivors - Preliminary Report. The New England journal of medicine.

Drummond, A.J., Ho, S.Y., Phillips, M.J., and Rambaut, A. (2006). Relaxed phylogenetics and dating with confidence. PLoS Biol *4*, e88.

Drummond, A.J., Suchard, M.A., Xie, D., and Rambaut, A. (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol *29*, 1969-1973.

Dudas, G. (2017). BALTIC - adaptable lightweight tree import code for molecular phylogeny manipulation, analysis and visualisation (github.com).

Dudas, G., Carvalho, L.M., Bedford, T., Tatem, A.J., Baele, G., Faria, N.R., Park, D.J., Ladner, J.T., Arias, A., Asogun, D.*, et al.* (2017). Virus genomes reveal factors that spread and sustained the Ebola epidemic. Nature *544*, 309-315.

Garrison, E.a.M., G. (2012). Haplotype-based variant detection from short-read sequencing. arXiv *1207.3907 [q-bio.GN]*

Gire, S.K., Goba, A., Andersen, K.G., Sealfon, R.S., Park, D.J., Kanneh, L., Jalloh, S., Momoh, M., Fullah, M., Dudas, G.*, et al.* (2014). Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. Science *345*, 1369-1372.

Kircher, M., Sawyer, S., and Meyer, M. (2012). Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic Acids Res *40*, e3.

Kraft, C.S., Hewlett, A.L., Koepsell, S., Winkler, A.M., Kratochvil, C.J., Larson, L., Varkey, J.B., Mehta, A.K., Lyon, G.M., 3rd, Friedman-Moraco, R.J.*, et al.* (2015). The Use of TKM-100802 and Convalescent Plasma in 2 Patients With Ebola Virus Disease in the United States. Clin Infect Dis *61*, 496-502.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Methods *9*, 357-359.

Lemey, P., Minin, V.N., Bielejec, F., Kosakovsky Pond, S.L., and Suchard, M.A. (2012). A counting renaissance: combining stochastic mapping and empirical Bayes to quickly detect amino acid sites under positive selection. Bioinformatics *28*, 3248-3256.
Levin, J.Z., Yassour, M., Adiconis, X., Nusbaum, C., Thompson, D.A., Friedman, N., Gnirke, A., and Regev, A. (2010). Comprehensive comparative analysis of strand-specific RNA sequencing methods. Nat Methods *7*, 709-715.
Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079.
Lyon, G.M., Mehta, A.K., Varkey, J.B., Brantly, K., Plyler, L., McElroy, A.K., Kraft, C.S., Towner, J.S., Spiropoulou, C., Stroher, U.*, et al.* (2014). Clinical care of two patients with Ebola virus disease in the United States. The New England journal of medicine *371*, 2402-2409.
Martin, M. (2011). Cutadapt removes adaptor sequences from high-throughput sequences reads. EMBnetjournal *17*, 10-12.
Mate, S.E., Kugelman, J.R., Nyenswah, T.G., Ladner, J.T., Wiley, M.R., Cordier-Lassalle, T., Christie, A., Schroth, G.P., Gross, S.M., Davies-Wayne, G.J.*, et al.* (2015). Molecular Evidence of Sexual Transmission of Ebola Virus. The New England journal of medicine *373*, 2448-2454.
Matranga, C.B., Andersen, K.G., Winnicki, S., Busby, M., Gladden, A.D., Tewhey, R., Stremlau, M., Berlin, A., Gire, S.K., England, E.*, et al.* (2014). Enhanced methods for unbiased deep sequencing of Lassa and Ebola RNA viruses from clinical and biological samples. Genome Biol *15*, 519.
McElroy, A.K., Akondy, R.S., Davis, C.W., Ellebedy, A.H., Mehta, A.K., Kraft, C.S., Lyon, G.M., Ribner, B.S., Varkey, J., Sidney, J.*, et al.* (2015). Human Ebola virus infection results in substantial immune activation. Proc Natl Acad Sci U S A *112*, 4719-4724.
O'Brien, J.D., Minin, V.N., and Suchard, M.A. (2009). Learning to count: robust estimates for labeled distances between molecular sequences. Mol Biol Evol *26*, 801-814.
Park, D.J., Dudas, G., Wohl, S., Goba, A., Whitmer, S.L., Andersen, K.G., Sealfon, R.S., Ladner, J.T., Kugelman, J.R., Matranga, C.B.*, et al.* (2015). Ebola Virus Epidemiology, Transmission, and Evolution during Seven Months in Sierra Leone. Cell *161*, 1516-1526.
Parkhomchuk, D., Borodina, T., Amstislavskiy, V., Banaru, M., Hallen, L., Krobitsch, S., Lehrach, H., and Soldatov, A. (2009). Transcriptome analysis by strand-specific sequencing of complementary DNA. Nucleic Acids Res *37*, e123.
Rambaut, A., Lam, T.T., Carvalho, L.M., Pybus, O.G. (2016). Exploring the temporal structure of heterchronus sequences using TempEst (formerly Path-O-Gen). Virus Evolution, vew007.
Schmeider, R.a.E., R (2011). Quality Control and preprocessing of metagenomics datasets. Bioinformatics *27*, 863-864.
Shepard, S.S., Meno, S., Bahl, J., Wilson, M.M., Barnes, J., and Neuhaus, E. (2016). Viral deep sequencing needs an adaptive approach: IRMA, the iterative refinement meta-assembler. BMC Genomics *17*, 708.
Sultan, M., Dokel, S., Amstislavskiy, V., Wuttig, D., Sultmann, H., Lehrach, H., and Yaspo, M.L. (2012). A simple strand-specific RNA-Seq library preparation protocol combining the Illumina TruSeq RNA and the dUTP methods. Biochem Biophys Res Commun *422*, 643-646.
Varkey, J.B., Shantha, J.G., Crozier, I., Kraft, C.S., Lyon, G.M., Mehta, A.K., Kumar, G., Smith, J.R., Kainulainen, M.H., Whitmer, S.*, et al.* (2015). Persistence of Ebola Virus in Ocular Fluid during Convalescence. The New England journal of medicine.
Wang, L., Si, Y., Dedow, L.K., Shao, Y., Liu, P., and Brutnell, T.P. (2011). A low-cost library construction protocol and data analysis pipeline for Illumina-based strand-specific multiplex RNA-seq. PLoS One *6*, e26426.
Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol *24*, 1586-1591.
Yoder, A.D., and Yang, Z. (2000). Estimation of primate speciation dates using local molecular clocks. Mol Biol Evol *17*, 1081-1090.

329 Materials Table:

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Chemicals, Peptides, and Recombinant Proteins		
Dnase I, Rnase-free	Roche	4716728001
Hybridase, Rnase H	Epicentre	H39100
Qiagen Rnase-free Dnase I	Qiagen	79254
Superase-in	Ambion	AM1694
Agencourt RNAclean XP SPRI clean up beads	Beckman Coulter Genomics	41105518
Dulbecco's modified Eagle's medium high glucose (DMEM)	Invitrogen	11960-044
HyClone fetal bovine serum	Thermo Scientific	SH30070.03
100x Non-essential amino acids	Invitrogen	11140050
10,000 U/mL Penicillin-streptomycin	Invitrogen	15140122
GlutaMAX	Invitrogen	35050061
FluoroBrite FluorBrite Dulbecco's modified Eagle's medium (DMEM)	Invitrogen	A1886701
TriPure Isolation reagent	Roche	11667165001
Critical Commercial Assays		
MagMax Pathogen RNA/DNA Isolation Kit	Invitrogen	4462359
Direct-zop-96 MagBead RNA Isolation Kit	Zymo Research	
TruSeq RNA Access Library Prep kit	Illumina	RS-301-2001
MiSeq Reagent Kit vs (300 cycle)	Illumina	MS-102-2002
HiSeq Rapid PE Cluster Kit v2	Illumina	PE-402-4002
TruSeq Rapid Duo cBot v1 Sample loading kit	Illumina	CT-402-4001
Moxi Z cell counter, M cassettes	Orflo Technologies	MXC001
Deposited Data		
EBOV genomcs from clinical specimens	This study	KY401638-KY401675, KY805810-2

Experimental Models: Cell Lines		
Huh7 cells	Albarino, C.G., Guerrero, L.W., Chakrabarti, A.K., Kainulainen, M.H., Whitmer, S.L., Welch, S.R., and Nichol, S.T. (2016). Virus fitness differences observed between two naturally occurring isolates of Ebola virus Makona variant using a reverse genetics approach. Virology 496, 237-243.	
Experimental Models: Organisms/Strains		
rEBOV-L2014/ZsG	Albarino, C.G., Guerrero, L.W., Chakrabarti, A.K., Kainulainen, M.H., Whitmer, S.L., Welch, S.R., and Nichol, S.T. (2016). Virus fitness differences observed between two naturally occurring isolates of Ebola virus Makona variant using a reverse genetics approach. Virology 496, 237-243.	
Sequence-Based Reagents		
rRNA-specific DNA probes	Adiconis X, Borges-Rivera D, Satija R, DeLuca DS, Busby MA, Berlin AM, Sivachenko A, Thompson DA, Wysoker A, Fennell T, Gnirke A, Pochet N, Regev A, Levin JZ: Comparative analysis of RNA sequencing methods for degraded or low-input samples. Nat Methods. 2013, 10: 623-629. 10.1038/nmeth.2483.	
Oligo dT, 40 nucleotides	idtdna.com	
Ebola virus-specific custom hybridization probes	Mate, S.E., Kugelman, J.R., Nyenswah, T.G., Ladner, J.T., Wiley, M.R., Cordier-Lassalle, T., Christie, A., Schroth, G.P., Gross, S.M., Davies-Wayne, G.J., et al. (2015). Molecular Evidence of Sexual Transmission of Ebola Virus. The New England journal of medicine 373, 2448-2454.	
Positive-sense NP synthetic RNA	This study.	Available upon request.
Negative-sense NP synthetic RNA	This study.	Available upon request.
NP forward_tagged strand-spcific primer	This study.	Available upon request.
NP reverse_tagged strand-specific primer	This study.	Available upon request.
Forward tagged primer	This study.	Available upon request.
Reverse tagged primer	This study.	Available upon request.
NP-specific probe	This study.	Available upon request.
Software and Algorithms		
Cutadapt v1.21	Martin, M. (2011). Cutadapt removes adaptor sequences from high-throughput sequences reads. EMBnetjournal 17, 10-12.	

Prinseq-lite v0.20.4	Schmeider, R.a.E., R (2011). Quality Control and preprocessing of metagenomics datasets. Bioinformatics 27, 863-864.	
Bowtie2	Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Methods 9, 357-359.	
Picard	broadinstitute.github.io/picard	
Samtools v0.1.18	Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078-2079.	
Path-O-Gen (now called TempEst)	Rambaut, A., Lam, T.T., Carvalho, L.M., Pybus, O.G. (2016). Exploring the temporal structure of heterchronus sequences using TempEst (formerly Path-O-Gen). Virus Evolution, vew007.	
R v3.3.1	R Core Team RFfSC: R: A Language and Environment for Statistical Computing. In. Vienna, Austria; 2012.	
Custom Python scripts	Park, D.J., Dudas, G., Wohl, S., Goba, A., Whitmer, S.L., Andersen, K.G., Sealfon, R.S., Ladner, J.T., Kugelman, J.R., Matranga, C.B., et al. (2015). Ebola Virus Epidemiology, Transmission, and Evolution during Seven Months in Sierra Leone. Cell 161, 1516-1526.	
Beast v1.8.2	Drummond, A.J., Suchard, M.A., Xie, D., and Rambaut, A. (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol 29, 1969-1973.	
Tracer v1.6	http://tree.bio.ed.ac.uk/software/tracer/	
CLC Genomics v9.0		
PopArt v1.7.2	http://popart.otago.ac.nz/index.shtml	
FreeBayes v1.0.2	Garrison, E.a.M., G. (2012). Haplotype-based variant detection from short-read sequencing. arXiv 1207.3907 [q-bio.GN]	
IRMA v0.6.5	Shepard, S.S., Meno, S., Bahl, J., Wilson, M.M., Barnes, J., and Neuhaus, E. (2016). Viral deep sequencing needs an adaptive approach: IRMA, the iterative refinement meta-assembler. BMC Genomics 17, 708.	
Tableau v10.0		
paml v4.5	http://abacus.gene.ucl.ac.uk/software/paml.html	
Custom strandedness, ADAR, and samogitia.py scripts	https://github.com/jtladner/ and https://github.com/evk3/EBOV_semen_sequencing	
Samogitia.py scripts	https://github.com/blab/baltic, by Gytis Dudas	

330

Sierra Leone Ebola Virus Persistence Study Group Affiliations:

- **Sierra Leone Ministry of Health and Sanitation:** Gibrilla Fadlu Deen (principle investigator), James Bangura, Amara Jambai, Faustine James, Alie H. Wurie, Francis Yamba
- **Sierra Leone Ministry of Defense:** Foday Sahr, Foday R. Sesay, Thomas A. Massaquoi
- **Sierra Leone Ministry of Social Welfare, Gender, Children's Affair:** Tina Davies
- **World Health Organization (WHO):** Nathalie Broutet (principle investigator), Pierre Formenty, Anna E. Thorsen, Archchun Ariyarajah, Marylin Carino, Antoine Coursier, Kara N. Durski, Ndema Habib, Philippe Gaillard, Sihem Landoulski, Margaret O. Lamunu, Jacyln E. Marrinan, Suzanna L.R. McDonald, Dhamari Naidoo
- **United States Centers for Disease Control and Prevention (US-CDC):** Barbara Knust (principle investigator), Neetu Abad, Kyle T. Bernstein, Elizabeth Ervin, John D. Klena, Tasneem Malik, Oliver Morgan, Stuart T. Nichol, Christine Ross, Ute Ströher
- **Chinese Center for Disease Control and Prevention (China-CDC):** Wnbo Xu (principle investigator), Hongtu Liu, William Jun Liu, Yong Xhang, Guizhen Wu, Mifang Liang
- **Joint United Nations Programme on HIV/AIDS (UNAIDS):** Patricia Ongpin