VIROLOGICA SINICA

Electronic Supplementary Material

An Assessment of Amplicon-Sequencing Based Method for Viral Intrahost Analysis

Ming Ni[1], Chen Chen[2], Di Liu[3]✉

1. Beijing Institute of Radiation Medicine, Beijing 100850, China
2. Institute of Infectious Diseases, Beijing Ditan Hospital, Capital Medical University, Beijing 100015, China
3. Wuhan Institute of Virology, Chinese Academy of Sciences, Wuhan 430071, China

Supporting information to DOI: 10.1007/s12250-018-0052-z

Supplementary Table S1. The mutant sequences of EBOV genome. The mutant sequences from 1,100 bp to 3,600 bp of EBOV genome (C15, GenBank accession: KJ660346.2). Sequences were synthesized, validated by Sanger sequencing, and cloned in plasmid pUC57.

Mutant sequences
AATTTCTTCTAATACACCAAGGGATGCACATGGTTGCCGGACATGATGCCAAGGATGCTGTGATTTCAAATTCAGTGGCTCAAGCTCGTTTTTCAGGTCTA
TTGATTGTCAAAACAGTACTTGATCATCTCCTACAAAAGACAGAACGAGGAGTTCGTCTCCATCCTCTTGCAAGGACCGCCAAGGTAAAAAATGAGGTGA
ACTCCTTCAAGGCTGCACTCAGCTCCCTGGCCAAGCATGGAGAGTATGCTCCTTTCGCCCGACTTTTGAACCTTTCTGGAGTAAATAATCCTGAGCATGGT
CTTTTCCCTCAACTGTCGGCAATTACACTCGGAGTCGCCACGGGTCACGGGAGTACCCTCGCAGGAGTAAATGTTGGAGAACAGTATCAACAGCTCAGAG
AGGCAGCCACTGAGGCCGAGAAGCAACTCCAACAATATGCGGAGTCTCGTGAACTTGACCATCTTGGACTTGATGATCAGGAAAAGAAAATTCTTATGAA
CTTCCATCAGAAAAAAAGCGAAATCAGCTTCCAGCAAACACACGCGATGGTAACTCTAAGAAAAGAGCGTCTGGCCAAGCTGACAGAAGCTATCACTGCT
GCATCACTGCCCAAAACAAGTGGACATTACGATGATGATGACGACATTCCCTTCCCAGGACCCATCAATGATGACGACAATCCTTGCCATCAAGATGATG
ATCCGACTTACTCACAGGATACGACCATTCTCGAAGTGGTAGTTGATCCCTATAATGGAGGCTACAGCGAATACCAAAGTTACTCGGAAAACGGCATGAG
TGCACCAGATGACCTGGTCCTATTCGATCTAGACGAGGACGACGAGGACACCAAGACAGTGTCTAACAGATCGACCAAGGGTGGACAACAGAGAAACAG
TCAAAAGAGCCAGCATACAGAGGGCAGACAGACACAATCCAGGCCAACTCAAAACGTCGCAGGCCCTCGCAGAACAATCCACCATGCCATTGCTCCACTC
ACGGACAATGACAGAAGAAACGAACCCCCCGGCTCAACCAGCCATCGCATACTGACCCCAATCAACGAATAGGCAGACCCACCGGACGATGCCGACGACG
AGACGTCTAGCCTTCTGCCCTTAGAGTCAGATGATGAAGAACAGGACAGGGATGGAACTTCTAACCGCACACCCACTGTCGCACCACCGGCTCCCGTATAC
AGAGATCACTCCGAAAGGAAAGAACTCCCGCAAGATGAACAACAAGATCAGGACCACATTCAAGAGGCCAGGAACCAAGACAGTGACAACACCCAGCCA
GAACATTCTTTTGAGGAGATGTATCGCCACATTCTAAGATCACAGGGGCCATTTGATGCCGTTTTGTATTATCATATGATGAAGGATGAGCCTGTAGTTTT
CAGTACCAGTGATGGTAAAGAGTACACGTATCCGGACTCCCTTGAAGAGGAATATCCACCATGGCTCACTAAAAAAGAGGCCGTGAATGATGAGAATAA
ATTTGTTACACTGGATAGCCAACAATTTCATTGGTCAGCAATGAATCACAGGAATAAATTCATAGCAATCCTGCAACATCATCAGCGAATGAGCATGCAA
CAATGGGATGATTTAATCGACAAATAGCTAACATTAAATAGTCAAGGAACGCAAACAGGCAGAATTTTTGATGTCTAAGGTGTGAATTATTATCACAATA
AAAGTGATTCTTAGTTTTGAATTTAAAGCTAGCCTATTATTACTAGCCGTTCTTCAAAGTTCAATTTGAGTCTTAATGCAAATAAGAGTTAAGCCACAGTT
ATAGCCATAATGGTAACTCAATATCTTAGCCAGCGATTTATCTAAATTAAATTACATTATGCTTTTATAACTTACCTACTAGCGTGCCTAACATTTACACG
ATCACTTCATGATTAAGAAAAAACTAATGATGAAGATTAAAACCTTCATCATCCTTACGTCAATTGAATTCTCTAGCACTAGAAGCTTATTGTCTTCAATG
TAAAAGAAAAACTGGCCTAACAAGATGACAACTAGAATAAAGGGCAGGGGCCATACTGTGGCCACGACTCAAAACGACAGAATGCCAGGCCCTGAGCTT
TCGGGCTGGATCTCTGAGCAGCTAATGACCGGAAGGATTCCTGTAAACAACATCTTCTGTGATATTGAGAACAATCCAGGATTATGCTACGCATCCCAAA
TGCAACAAACGAAGCCAAACCCGAAGATGCGCAACAGTCAAACCCAAACGGACCCAATTTGCAATCATAGTTTTGAGGAGGTAGTACAAACATTGGCTTC
ACTGGCTACTGTTGTGCAACAACAAACCATCGCATCAGAATCATTAGAACAACGCGTTACGAGTCTTGAGAATGGTATAAAGCCAGTTTATGATATGGCA
AAAACAATCTCCTCATTGAACAGGGTTTGTGCTGAAATGGTTGCAAAATATGATCTTCTAGTGATGACAACCGGTCGAGCAACAACAACTGCTGCGG

Supplementary Table S2. Substitutions of the mutant sequences. ORF, Open Reading Frame.

#	EBOV site (bp, KJ660346.2)	ORF	Subsitution
1	1153	NP	C>G
2	1229	NP	A>C
3	1392	NP	T>C
4	1427	NP	G>A
5	1444	NP	A>G
6	1446	NP	C>G
7	1447	NP	C>T
8	1456	NP	C>T
9	1519	NP	T>C
10	1618	NP	G>A
11	1620	NP	A>G
12	1643	NP	A>C
13	1672	NP	C>T
14	1756	NP	T>C
15	1787	NP	G>T
16	1811	NP	G>T
17	1833	NP	C>T
18	1837	NP	T>A
19	1853	NP	G>T
20	1856	NP	G>A
21	1868	NP	G>A
22	1916	NP	T>C
23	1958	NP	C>A
24	1964	NP	C>T
25	1995	NP	A>G
26	2009	NP	G>A
27	2043	NP	C>G
28	2060	NP	A>G
29	2091	NP	G>T
30	2129	NP	T>C
31	2145	NP	C>A
32	2152	NP	G>A
33	2171	NP	G>T
34	2184	NP	T>C
35	2217	NP	C>T
36	2254	NP	C>T
37	2284	NP	C>A
38	2319	NP	A>G
39	2573	NP	G>A

40	2585	NP	A>G
41	2601	NP	G>A
42	2618	NP	G>A
43	2620	NP	T>C
44	2630	NP	T>C
45	2636	NP	C>T
46	2640	NP	T>C
47	2665	NP	G>A
48	2687	NP	T>C
49	2699	-	T>C
50	2702	-	T>C
51	2761	-	A>C
52	2835	-	T>C
53	2853	-	T>C
54	2888	-	C>A
55	2986	-	C>G
56	2991	-	C>T
57	3007	-	G>A
58	3008	-	T>C
59	3011	-	T>C
60	3014	-	A>G
61	3115	-	G>A
62	3142	VP35	C>T
63	3252	VP35	G>A
64	3405	VP35	T>C
65	3459	VP35	A>G
66	3480	VP35	C>A
67	3539	VP35	G>A
68	3563	VP35	G>A
69	3581	VP35	G>A
70	3588	VP35	G>A
71	3593	VP35	C>T

Supplementary Table S3. Summary of amplicon-based and direct sequencing of the mixed samples. Direct-seq, direct sequencing of plasmid DNA without viral specific amplification. Amplicon-seq, sequencing of viral specific PCR products. The mutant:wild-type ratios were the designed ratios. Reference genome for alignemnt was EBOV C15 (GenBank accession no. KJ660346.2).

Sample Id	Sequencing approach	mutant:wild-type ratio	viral content (copies/μL)	Clean data (Mbp)	Alignment ratio	Mean target site depth (X)
1	direct-seq	1:2	100000	99.1499	0.4566	17099
2	amplicon-seq	1:2	100	45.6132	0.9463	17718
3	amplicon-seq	1:2	100000	50.5606	0.9812	20008
4	direct-seq	1:4	100000	100.745	0.4619	17596
5	amplicon-seq	1:4	100	56.7952	0.9595	21166
6	amplicon-seq	1:4	100000	50.8051	0.9834	20423
7	direct-seq	1:8	100000	98.7527	0.4635	17233
8	amplicon-seq	1:8	100	53.8624	0.9443	21295
9	amplicon-seq	1:8	100000	49.9563	0.9837	20082
10	direct-seq	1:32	100000	98.0562	0.4659	17251
11	amplicon-seq	1:32	100	49.9313	0.9603	18698
12	amplicon-seq	1:32	100000	54.0076	0.983	20952

Sample Preparation

The wild-type and mutant sequences from position 1,100 to 3,600 of EBOV genome (GenBank accession: KJ660346.2) were synthesized, validated by Sanger sequencing, and cloned in plasmid pUC57 by Sangon Biotech Co., Ltd (Shanghai, China). The mutant had 71 substitutions corresponding to iSNV events occurred in this region among patients (Supplementary Table S1). Plasmid DNA was extracted by using Tiangen Plasmid DNA Mini Kit (Tiangen, Beijing, China) and quantified by using Qubit 2.0 Fluorometer (Invitrogen, USA). The DNA samples from mutant and wild-type were mixed with a mutant:wild-type DNA amount ratio of 1:2, 1:4, 1:8, and 1:32, respectively. Aliquots of the four mixture samples were directly sequenced as baselines. Other aliquots of the mixture samples were diluted to ~10^5 copies/μL (~ a 25 *Ct* value of EBOV viral load) and ~ 10^2 copies/μL (~ a 35 *Ct* value), respectively. Namely, there were eight samples with two viral loads at four mutant:wild-type ratios. Each diluted sample was added with 100 ng of human cDNA from A549 cell-line.

EBOV-Specific Amplification

Amplifications were performed with two pairs of EBOV-specific primers. Primer pair 1: 5′-CCTACAAAAGACAGAACGAGGA-3′ (forward primer) and 3′-TACAAAACGGCATCAAATGGC-5′ (reverse primer). Primer pair 2: 5′-GGAACTTCTAACCGCACACC-3′ (forward primer) and 3′-TTCTAATGATTCTGATGCGATG-5′ (reverse primer). PCR amplification was performed with NEB Phusion High-Fidelity PCR Master Mix with HF Buff (New England Biolabs, USA). The regimen of thermal cycling: 3 min at 95 °C; 25 cycles (30 s at 95 °C, 30 s at 60 °C, 45 s at 72 °C), 5 min at 72 °C. The two PCR products of the two primer pairs for each sample were pooled and cleaned with QIAquick PCR Purification Kit (Qiagen, Germany) according to the manufacturer's instructions.

Next-Generation Sequencing

The pooled PCR products were prepared multiplex NGS library by using Nextera XT Sample Preparation Kit (Illumina Inc., USA) according to the manufacturer's instructions. Illumina MiSeq platform was employed to generate 2 ×150-bp pair-ended reads.

Bioinformatics of iSNV Calling

We implemented quality control and error correction according to Schirmer et al. investigations on amplicon-seq error patterns generated by Illumina's MiSeq and Nextera XT Sample Preparation Kit (Schirmer *et al.*, 2015). Because nucleotide-specific substitution errors are likely to enrich at both ends of reads, the first 10 bp of each read were trimmed and Sickle v1.3.3 (Joshi *et al.*, 2011) was employed to trim the low quality bases at the end of reads with a threshold of Q20 and a requirement of 100 bp minimum read length. Following, Bayeshammer (implemented in SPAdes v3.5.0) (Nikolenko *et al.*, 2013) was used for error correlation. Reads without their corresponding paired reads were disregarded. The remaining paired reads were used as clean reads.

Clean reads were pair-ended aligned to the reference EBOV genome (GenBank accession: KJ660346.2) by using Bowtie2 v2.2.5 (Langmead and Salzberg, 2012) with default parameters. SAMtools v1.2 (Li *et al.*, 2009) was employed to generate 'mpileup' files with no limit of the maximum site depth. Homemade PERL scripts (available at http://github.com/generality/iSNV-calling/) were developed for iSNV calling using the mpileup files as input. The calling processes are as follows. Firstly, for each site of EBOV genome, the aligned low quality bases (< Q20) and indels were excluded to reduce possible false positive, and the site depth and strand bias were re-calculated. Then, a series of criteria were used to call iSNVs: 1) Minor allele frequency ≥ 0.1% to ≥ 1% (see text); 2) Depth of the minor allele ≥ 5; and 3) The strand bias of the minor allele was less than 10-fold. Moreover, the iSNV sites either located within the EBOV-specific primers, or the 30-bp downstream the 5′-primer, or 30-bp upstream the 3′-primer were also discarded.

References

Joshi NA and Fass JN (2011) Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software]. Available at https://github.com/najoshi/sickle. Accessed June 5th 2016.

Langmead B and Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nature Methods 9:357-359.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078-2079.

Nikolenko SI, Korobeynikov AI, and Alekseyev MA (2013) BayesHammer: Bayesian clustering for error correction in single-cell sequencing. BMC Genomics, 14: Suppl 1, S7.

Schirmer M, Ijaz UZ, D'Amore R, Hall N, Sloan WT, Quince C (2015) Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. Nucleic Acids Res 43:e37