Table S1. Mock bacterial community details [a]

	Sequence name	Taxonomy (Phyla or class)	Source	Insert length (nt)[b]
1	Acidobacteria	Acidobacteria	Drinking water	1359
2	Actinobacteria	Actinobacteria	Wastewater reactor	1392
3	Bacteroidetes clone 1	Bacteroidetes	Wastewater reactor	1355
4	Bacteroidetes clone 2	Bacteroidetes	Drinking water	1352
5	*Caldisericum exile*	OP5	DSMZ culture collection-13637	1426
6	Chlorobi	Chlorobi	Surface water	1374
7	Cyanobacteria	Cyanobacteria	Surface water	1324
8	*Deferribacter desulfuricans*	Deferribacteres	DSMZ culture collection-14783	1410
9	*Deinococcus indicus*	Deinococcus-Thermus	DSMZ culture collection-1537	1366
10	*Desulfurispirillum alkaliphilum*	Chrysiogenetes	DSMZ culture collection-1827	1375
11	*Dictyoglomus thermophilum*	Dictyoglomi	DSMZ culture collection-396	1415
12	*Fibrobacter succinogenes* S85	Fibrobacteres	Donated by Isaac Cann, University of Illinois-Urbana Champaign	1372
13	Gemmatimonadetes	Gemmatimonadetes	Wastewater reactor	1360
14	Leptotrichia hofstadii	Fusobacteria	DSMZ culture collection-21561	1367
15	*Mycoplasma orale*	Firmicutes	DSMZ culture collection-1915	1375
16	Nitrospira	Nitrospirae	Wastewater reactor	1376
17	*Persephonella hydrogeniphiia* H3	Aquificae	Donated by Anne Louise Reysenbach, Portland State University	1389
18	Planctomycetes	Planctomycetes	Wastewater reactor	1376
19	*Protochlamydia amoebophilia*	Chlamydiae	Donated by Mathias Horn, University of Vienna	1360
20	Spirochaetes	Spirochaetae	Surface water	1396
21	*Sulfurihydrogenibium yellowstonense*	Aquificae	Donated by Anne Louise Reysenbach, Portland State University	1378
22	Synergistetes	Synergistetes	Surface water	1355
23	*Syntrophobacter fumaroxidans*	Deltaproteobacteria	Donated by Syed Hashsham, Michigan State University (DSMZ# 117)	1415
24	*Syntrophococcus sucromutans*	Firmicutes	Donated by Syed Hashsham, Michigan State University (ATCC# 43584)	1380
25	*Syntrophomonas bryantii*	Firmicutes	Donated by Syed Hashsham, Michigan State University (DSMZ# 314A)	1412
26	*Syntrophothermus lipocalidus*	Firmicutes	Donated by Syed Hashsham, Michigan State University (DSMZ# 1268)	1500
27	*Syntrophus buswellii*	Deltaproteobacteria	Donated by Syed Hashsham, Michigan State University (DSMZ# 2612A)	1413
28	*Syntrophus gentianae*	Deltaproteobacteria	Donated by Syed Hashsham, Michigan State University (DMZ# 8423)	1412
29	*Thermodesulfobacterium commune*	Thermodesulfobacteria	DSMZ culture collection-2178	1422
30	*Thermomicrobium roseum*	Chloroflexi	DSMZ culture collection-5159	1371
31	*Thermotoga neapolitana*	Thermotogae.	Donated by Claire Vielle, Michigan State University	1412
32	Verrucomicrobia	Verrucomicrobia	Surface water	1379
33	*Victivallis vadensis*	Lentisphaerae	DSMZ culture collection-8748	1360

[a] The mock community was a gift from Dr. Lutgarde Raskin, Department of Civil and Environmental Engineering, University of Michigan, United States of America.

[b] The insertions are near full length 16S rDNA sequences.

Table S3. Technical Aspect Comparison of PAS and Other Phasing Methods

Methods	Strategy	Frame shift	Spacer position	Complementary Design	Number of Bases Sacrificed from the Sequence (excluding primers)	PCR, Number of cycles	Extra PCR Bias	Chimera Potential	Systematic Evaluation[c]	Reference
PAS	Adding spacers	1-7 bases	After sequencing primers, in both forward and reverse primers	yes	7 (spacer)	Two-step, (10, 20)	no	Less, lower cycle number at each step	Yes	This paper
Improved Dual-Index	Adding spacers	1-7 bases	After barcode, in both forward and reverse primers	No[b]	≥31 (barcode, spacer)	One-step (30)	yes	More, higher cycle number	No	(Fadrosh, Ma et al. 2014)
Dual-Index	N/A[a]	N/A	N/A	N/A	24 (barcode)	One- step (30)	yes	More, higher cycle number	No	(Kozich, Westcott et al. 2013)
Random Molecular tag	Adding spacers	1-5 bases	In the middle of random tags	No[b]	≥30 (tag, spacer, barcode)	Tagging, One step PCR (30)	yes	More, higher cycle number	Partial[d]	(Lundberg, Yourstone et al. 2013)
Staggered barcode	Staggered barcode	1-3 bases	Before the target gene primer of the forward primer	N/A	1-3 (barcode)	One step (30)	Yes	More, higher cycle number	No	(Hummelen, Macklaim et al. 2011)

[a] N/A. not applicable

[b] Not mentioned in the manuscript

[c] A systematic evaluation includes evaluation of sequence output, reads quality, error rate, etc.

[d] Systematic evaluation for data output, and quality

Fadrosh, D. W., B. Ma, P. Gajer, N. Sengamalay, S. Ott, R. M. Brotman and J. Ravel (2014). "An improved dual-indexing approach for multiplexed 16S rRNA gene sequencing on the Illumina MiSeq platform." Microbiome **2**(6).

Hummelen, R., J. M. Macklaim, J. E. Bisanz, J.-A. Hammond, A. McMillan, R. Vongsa, D. Koenig, G. B. Gloor and G. Reid (2011). "Vaginal Microbiome and Epithelial Gene Array in Post-Menopausal Women with Moderate to Severe Dryness." Plos One **6**(11).

Kozich, J. J., S. L. Westcott, N. T. Baxter, S. K. Highlander and P. D. Schloss (2013). "Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform." Applied and Environmental Microbiology: 7846-7849.

Lundberg, D. S., S. Yourstone, P. Mieczkowski, C. D. Jones and J. L. Dangl (2013). "Practical innovations for high-throughput amplicon sequencing." Nature Methods **10**(10): 999-+.

Table S4. Estimated Base Frequencies of V4 Region 16S rDNA Amplicons after PCR with Target Primers [a]

Position	Base Frequencies （%）							
	After 515F				After 806R			
	A	T	C	G	A	T	C	G
1	1.6	**95.48**	1.4	1.51	0.01	0.36	**99.54**	0.01
2	**99.78**	0.07	0.04	0.1	0.02	0.33	**99.59**	0.02
3	0.07	0.05	**99.82**	0.04	0.06	**92.81**	5.33	1.76
4	6.29	0.2	1.45	**92.02**	1.63	1.06	0.68	**96.6**
5	5.8	67.47	0.14	26.58	0.08	**99.29**	0.24	0.35
6	**96.72**	0.1	0.02	3.14	0.04	**99.54**	0.17	0.09
7	3.6	2.96	1.69	**91.57**	0.03	69.72	29.93	0.28
8	0.99	0.31	0.3	**98.37**	0.2	0.36	0.24	99.17
9	6.5	31.37	0.9	61.21	19.54	0.49	79.54	0.4
10	3.68	47.28	11.84	37.19	0.09	**98.67**	0.6	0.61
11	2.31	1.4	12.73	83.55	21.94	0.45	77.43	0.12
12	0.13	0.67	**98.8**	0.37	0.1	0.1	**99.74**	0.02
13	54.66	5.85	0.53	38.88	0.06	0.09	99.79	0.02
14	**97.89**	0.05	0.07	1.97	0.3	0.43	**99.21**	0.03
15	1.38	0.03	0.21	**98.37**	**90.44**	5.47	2.25	1.81
16	0.06	1.47	**97.97**	0.49	5.03	12.53	79.7	2.69
17	0.17	0.21	0.35	**99.24**	6.58	0.1	1.13	**92.17**
18	0.24	**98.15**	0.21	1.4	0.08	0.44	**98.62**	0.82
19	0.67	**99**	0.15	0.17	0.04	**97.27**	2.56	0.09
20	55.66	0.63	0.36	43.32	0.04	**98.86**	0.65	0.41
21	18.85	77.09	3.12	0.92	0.03	**99.32**	0.53	0.07
22	0.14	29.02	70.29	0.53	0.06	0.97	**98.49**	0.42
23	0.08	0.3	**98.98**	0.63	0.63	0.28	0.99	**98.05**
24	0.76	0.11	0.69	**98.42**	20.92	22.43	54.79	1.83
25	0.66	0.48	0.68	**98.16**	29.87	19.8	5.57	44.74
26	**97.76**	0.94	0.04	1.23	0.72	4.78	**93.84**	0.62
27	64.12	35	0.28	0.59	24.43	3.32	72.02	0.17
28	0.59	**99.21**	0.12	0.05	0.29	**98.16**	1.34	0.18
29	1.32	**90.67**	6.3	1.7	0.8	1.95	83.5	13.71
30	**98.8**	0.87	0.21	0.1	**98.09**	0.18	0.83	0.87
31	0.62	52.44	46.85	0.08	7.37	0.18	2.06	**90.32**
32	0.11	**99.09**	0.35	0.43	0.27	13	85.3	1.37
33	0.18	0.65	0.07	**99.08**	0.12	0.66	0.84	**98.35**
34	0.18	0.16	0.11	**99.52**	0.12	**98.12**	0.73	1
35	0.09	0.13	0.36	**99.41**	0.51	1.06	**98.19**	0.21
36	0.06	17.7	81.04	1.2	**96.81**	0.33	0.99	1.84
37	0.33	10	1.71	87.95	5.65	0.61	0.29	**93.41**
38	0.48	**98.57**	0.21	0.74	3.63	86.76	0.83	8.74
39	**98.83**	0.82	0.07	0.22	34.74	50.17	7.19	7.85
40	**99.35**	0.19	0.15	0.29	58.45	23.17	8.95	9.38
41	**98.87**	0.07	0.11	0.94	14.93	23.64	57.4	3.99
42	1.16	0.05	0.4	**98.37**	45.68	17.5	7.29	29.49
43	12.03	0.19	49	38.77	10.24	9.3	3.7	76.71
44	0.81	0.19	0.85	**98.14**	22.82	27.42	35.42	14.3
45	20.86	22.56	54.46	2.11	0.85	4.82	**93.38**	0.91
46	9.73	19.9	5.6	64.76	0.77	7.26	**91.73**	0.21
47	0.69	8.27	89.51	1.52	**93.35**	0.51	2.19	3.92
48	0.23	0.63	1.6	**97.53**	3.32	0.32	0.41	**95.9**
49	0.72	67.73	30.4	1.15	49.18	23	6.97	20.82
50	**97.79**	0.72	0.4	1.06	47.27	7.38	3.37	41.95
51	1.54	0.27	0.49	**97.51**	71.22	2.64	1.53	24.57
52	6.47	0.5	2.54	**90.48**	13.4	2.34	14.17	70.05
53	0.16	24.88	72.32	2.63	1.43	27.88	67.46	3.2
54	0.17	1.77	1.94	**96.11**	0.21	30.93	67.81	1
55	0.18	1.4	0.34	**98.06**	0.27	0.53	1.97	**97.19**
56	10.16	65.62	16.77	7.43	0.18	0.77	**97.26**	1.73
57	8.27	70.88	15	5.83	0.06	9.16	**90.28**	0.33
58	18.5	55.99	10.71	14.78	0.06	**97.85**	1.89	0.16

59	18.53	32.87	10.11	38.46	0.62	**97.58**	1.41	0.34
60	15.81	59.07	3.26	21.84	0.08	1.96	**97.05**	0.85
61	3.91	61.72	33.34	1.01	0.54	0.54	2.96	**95.91**
62	80.02	2.12	2.18	15.65	0.22	0.42	**97.65**	1.67
63	78.57	0.94	17.3	3.17	9.13	1.88	88.17	0.79
64	1.5	2.07	0.55	**95.86**	**93.32**	2.78	2.55	1.33
65	4.94	86.74	6.9	1.4	1.71	31.27	66.65	0.33
66	2.56	11.63	80.95	4.83	1.62	52.87	43.37	2.1
67	29.01	38.52	3.95	28.5	0.58	1.18	1.07	**97.14**
68	5.56	9.71	1.85	82.85	2.63	1.1	0.3	**95.95**
69	60.5	14.56	12.45	12.47	4.44	**90.58**	1.61	3.34
70	17.66	60.3	5.87	16.15	13.35	2.65	0.48	83.5
71	0.79	3.79	0.38	**95.02**	0.42	**96.33**	0.43	2.79
72	0.71	**95.25**	0.32	3.71	0.13	**98.21**	1.19	0.44
73	4.49	9.63	1.41	84.46	0.06	2.62	**97.1**	0.18
74	**90.52**	6.78	0.18	2.5	0.38	20.48	78.94	0.17
75	**98.72**	0.29	0.2	0.77	2.61	**90.98**	3.22	3.16
76	**96.61**	0.94	0.43	2	2.61	9.98	85.59	1.79
77	7.48	40.78	2.94	48.77	1.97	2.2	87.55	8.24
78	4.16	11.88	71.78	12.17	38.71	39.56	13.07	8.63
79	0.83	17.42	78.78	2.94	22.56	26.88	11.99	38.55
80	7.64	22.04	66.26	4.04	**95.43**	1.48	0.66	2.41
81	25.58	9.85	24.05	40.49	2.63	**95.25**	0.64	1.46
82	5.59	10.71	29.49	54.19	66.75	2.91	30.14	0.18
83	11.89	1.93	4.35	81.82	1.61	**96.72**	1.58	0.07
84	0.67	0.88	2.81	**95.62**	0.62	4.14	**95.04**	0.14
85	0.24	3.11	**94.59**	2.04	2.79	**94.84**	2.16	0.18
86	0.64	**94.27**	4.72	0.36	63.05	2.25	1.98	32.7
87	3.1	38.27	58.42	0.18	7.01	5.46	86.19	1.3
88	**96.73**	1.01	2.02	0.21	1.68	0.41	2.55	**95.33**
89	**95.78**	0.5	3.38	0.31	7.25	0.68	87.83	4.22
90	2.16	1.08	**96.13**	0.62	**96.32**	1.21	1.85	0.6
91	1.6	20.82	72.54	5.03	2.01	**97.08**	0.71	0.18
92	9.7	28.6	30.95	30.73	0.6	**98.53**	0.64	0.19
93	7.03	41.28	23.94	27.72	0.37	83.65	15.87	0.09
94	14.48	7.99	2.89	74.57	1.08	3.89	**94.93**	0.07
95	15.75	3.94	1.08	79.2	**95.88**	0.81	3	0.29
96	58.28	9.43	3.89	28.36	1.98	0.49	**97.33**	0.18
97	53.99	5.93	17.25	22.8	0.52	8.4	90.22	0.84
98	5.53	16.92	42.33	35.2	1.16	6.6	5.54	86.68
99	1.65	61.1	2.47	34.75	0.34	1.51	**96.36**	1.74
100	0.89	20.12	12.6	66.38	1.06	**95.97**	2.51	0.43
Percentage of positions with dominant base of base frequencies >90%	49				63			
Percentage of positions with dominant base of base frequencies >75%	63				79			

[a] Base frequencies were calculated based on aligned prokaryotic 16S rRNA sequences (current_prokMSA_aligned.fasta.gz) downloaded from the Greengenes database. Regions between the primer sets (515F and 806R) were extracted and subjected to frequency calculation from both directions.

Table S5. Method Differences and Advantages and Disadvantages of the PAS and Other Phasing Methods

Method	Features	
PAS method	Strategy	1. Spacers are located before target primers to provide sequence position frame shifts among samples to be pooled and sequenced in one run; 2. Barcode is sequenced separately; 3. Uses complementally spacers in forward and reverse primers; 4. Uses a two-step PCR to avoid extra PCR biases potentially resulting from different affinities of upstream and downstream regions of the DNA template to the various primer components (i.e., Illumina adapter, spacer, barcode, etc.); 5. Uses lower PCR cycles in both steps PCR to minimize chimeras.
	Method evaluation	Systematically evaluated and compared to other methods, to assess data output (highest cluster density and the percentage of clusters passing filter), read quality (average length of effective read sequences and the number of effective reads after quality trimming, error rates and bias).
	Advantages	1. Use of 1-7 base spacers in both forward and reverse primers provides a sufficient sequence position frame shift among samples; 2. Minimum base sacrifice since barcodes are sequenced in separate reads; 3. Uses the minimum number of bases necessary for base frame shift via a complementary primer design to achieve both the maximum amplicon length and the minimum level of bias in read quality; 4. No extra PCR bias from added primer components using two step PCR, and lower potential of chimeras using lower number of cycles in both steps of PCR.
	Disadvantages	A larger number of barcoded primers are needed for large numbers of multiplex amplicon sequencing.
	Conditions to use	Can be used with version two or higher Miseq sequencing kits.
	Reference	This manuscript
Improved dual-Indexing	Strategy	1. Uses a dual-index to decrease the cost of primers; barcodes of 12 bp are in both forward and reverse primers; 2. Places spacers between the index and target gene primers to insert a frame shift among template DNA; sequencing starts at the index.
	Method evaluation	No systematic evaluation.
	Advantages	1. Provides a sequence shift; 2. Fewer primers are needed to pool a larger number of samples.
	Disadvantages	1. The 16S amplicon selected (v3-v4) is 469bp, too long for most of paired end reads to be merged with overlap after quality trimming; 2. Use of a one-step PCR could add extra PCR bias, and use of 30 cycles in PCR amplification could introduce more chimeras; 3. Both forward and reverse reads include barcodes which sacrifices 24 bp of sequence at the 5' end, which has the best base quality; 4. Lack of systematic evaluation, so not ready to be adapted.
	Conditions to use	This method sacrifices more bases; so, particular caution must be paid for amplicon selection.
	Reference	(Fadrosh, Ma et al. 2014)
Dual-Indexing	Strategy	Use of dual-indexing to save primer costs, place 12 bp barcodes on both forward and reverse primers, which does not provide a frame shift.
	Method evaluation	Evaluated Miseq amplicon sequencing using the v3-4, v4, and v4-5 regions of the 16S rRNA genes in term of primer pair selection; Compared the Miseq amplicon sequencing to 454 amplicon sequencing.
	Advantages	Fewer primers are needed to pool a large number of samples using dual-indexing.
	Disadvantages	1. No frame shift, no solution to the low sample sequence diversity issue; 2. A one-step PCR could introduce extra PCR bias caused; 3. High PCR cycles potentially introduce more chimeras; 4. 12 bp barcodes are included in both forward and reverse reads, sacrificing 24 bases of the paired end reads at the 5’ end.
	Conditions to use	Not a good option for any kind of Miseq amplicon sequencing.

	Reference	(Kozich, Westcott et al. 2013)
Molecular random tagging	Strategy	1. Uses random tags (8 random bases at the 5' end of both forward and reverse primers) to label each template DNA before PCR to get a consensus sequence to correct errors; 2. Spacers are added within the random tags (1-5 bases) in both forward and reverse primers and barcodes (9 bp) are added within random tags in the reverse primer, sequencing starts at the random tag.
	Method evaluation	Compared error rates with and without consensus sequences; Tested with one cloned 16S template and pooled soil and root samples; Limited comparisons were made between runs with and without the frame shift.
	Advantages	Lowers sequence errors and reduces the number of artificial OTUs, and increases sequence quality.
	Disadvantages	1. Method has not been evaluated using mock communities with high complexity; 2. The spacers do not achieve sufficient frame shift for template DNA; 3. One-step PCR could introduce additional PCR bias; 4. Sacrifices 30 bases in addition to the target primers at the 5' end of the paired end reads; 5. High PCR cycles potentially introduce more chimeras.
	Conditions to use	The method needs more systematic evaluation, and is not ready for general application.
	Reference	(Lundberg, Yourstone et al. 2013)
Staggered barcodes	Strategy	Uses barcodes of 3-6 bases in the forward primers to make a frame shift of 1-3 bases only for the forward reads.
	Method evaluation	No systematic evaluation or comparison to other methods.
	Advantages	1. With a 1-3 base frame shift, the low diversity sample issue is mitigated to some extent; 2. Sacrifices only additional 3-6 bases of the read sequence in addition to the target gene primers.
	Disadvantages	1. The frame shift is not sufficient for forward reads, and it is lack for reverse reads; 2. Use of a one-step PCR could introduce additional PCR bias; 3. Use of 30 PCR cycles could increase chimera formation; 4. Lack of systematic evaluation, not ready for adaptation.
	Conditions to use	Not a good option for any kind of Miseq amplicon sequencing.
	Reference	(Hummelen, Macklaim et al. 2011)

Fadrosh, D. W., B. Ma, P. Gajer, N. Sengamalay, S. Ott, R. M. Brotman and J. Ravel (2014). "An improved dual-indexing approach for multiplexed 16S rRNA gene sequencing on the Illumina MiSeq platform." Microbiome **2**(6).

Hummelen, R., J. M. Macklaim, J. E. Bisanz, J.-A. Hammond, A. McMillan, R. Vongsa, D. Koenig, G. B. Gloor and G. Reid (2011). "Vaginal Microbiome and Epithelial Gene Array in Post-Menopausal Women with Moderate to Severe Dryness." Plos One **6**(11).

Kozich, J. J., S. L. Westcott, N. T. Baxter, S. K. Highlander and P. D. Schloss (2013). "Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform." Applied and Environmental Microbiology: 7846-7849.

Lundberg, D. S., S. Yourstone, P. Mieczkowski, C. D. Jones and J. L. Dangl (2013). "Practical innovations for high-throughput amplicon sequencing." Nature Methods **10**(10): 999-+.

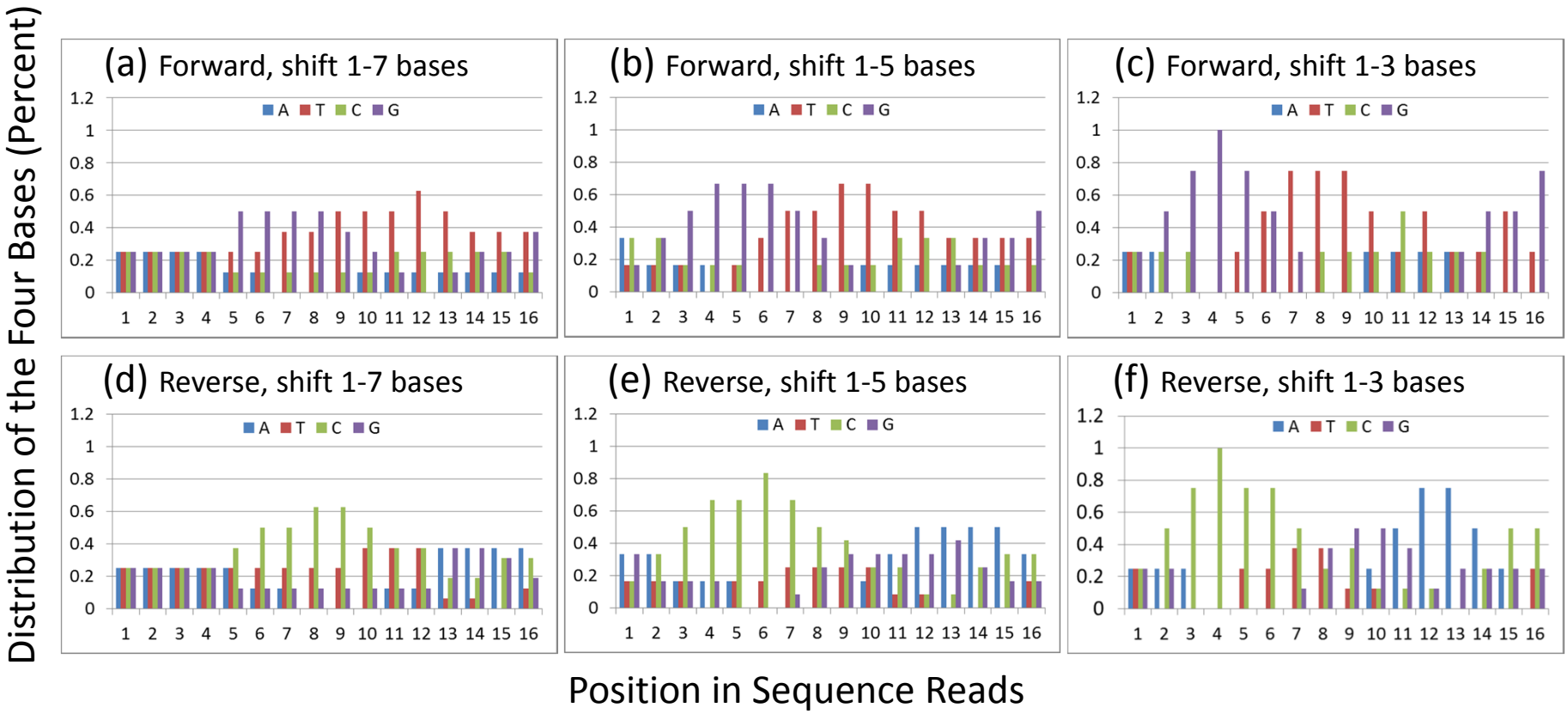


Figure S1. Theoretical base distribution of the first 16 positions in both the forward and reverse reads of the amoA gene amplicon amplified by the primer pair amoA F1 (5'-GGGGTTTCTACTGGTGGT) and R2 (5'-CCCCTCKGSAAAGCCTTCTTC) with different frame shift lengths. Distribution of the four bases, A, T, C, G, in the forward (a, b, c) and reverse reads (d, e, f) with frame shifts of 1-7 (a, d), 1-5 (b, e), and 1-3 (c, f) bases.

Figure S2. Theoretical base distribution of the first 16 positions of both the forward and reverse reads of the 16S rRNA gene amplified by the primer pair of 515F (5'-GTGCCAGCMGCCGCGGTAA) and 806R (5'-GGACTACHVGGGTWTCTAAT) (v4 region) with frame shifts of different lengths. Distribution of the four bases, A, T, C, G, in the forward (a, b, c) and reverse reads (d, e, f) with frame shifts of 1-7 (a, d), 1-5 (b, e), and 1-3 (c, f) bases.

Overview of phasing amplicon sequencing strategy

A. **16S rDNA target only primers** Used in the first step PCR

515F 806R

5' GTGCCAGCMGCCGCGGTAA3' 3'T A A T C T WT GGGVHCATCAG5'

B. **16S rDNA template**

C. **First step PCR** Using target only primers

D. **First step PCR products**

G TGCCAGCMGCCGCGGTAATAC G G A G G G T G C . . . . . GGAGCAAACAGGATTAGATACCCTGGTAGTC

C ACGGTCGTCGGCGCCATTATG C C T C C C A C G . . . . . CCTCGTTTGTCCTAATCTWTGGGVHCATCAG

E. **16S rDNA phasing primers** with Illumina adapters, sequencing primers, spacers of 0-7 bases, and barcodes for reverse primers, which used in the second step PCR

Forward	Illumina adapter	Forward read sequencing primer	Spacer	Target gene primer
515F-set 1	5' AATGATACGGCGA CCACCGAGATCTACACTCTTT C C C T A C ACGACGCT	CTTCCGATCTTCGCATAGTGCCAGCMGCCGCGGTAA3'		
515F-set 2	5' AATGATACGGCG ACCACCGAGATCTACACTCTT T C C C T A C ACGACGCT	CTTCCGATCTCGCATAGTGCCAGCMGCCGCGGTAA3'		
515F-set 3	5' AATGATACGGC GACCACCGAGATCTACACTCT T T C C C T A C ACGACGCT	CTTCCGATCTGCATAGTGCCAGCMGCCGCGGTAA3'		
515F-set 4	5' AATGATACGG CGACCACCGAGATCTACACTC T T T C C C T A C ACGACGCT	CTTCCGATCTCATAGTGCCAGCMGCCGCGGTAA3'		
515F-set 5	5' AATGATACG GCGACCACCGAGATCTACACT C T T T C C C T A C ACGACGCT	CTTCCGATCTATAGTGCCAGCMGCCGCGGTAA3'		
515F-set 6	5' AATGATAC GGCGACCACCGAGATCTACAC T C T T T C C C T A C ACGACGCT	CTTCCGATCTTAGTGCCAGCMGCCGCGGTAA3'		
515F-set 7	5' AATGATA CGGCGACCACCGAGATCTACA C T C T T T C C C T A C ACGACGCT	CTTCCGATCTAGTGCCAGCMGCCGCGGTAA3'		
515F-set 8	5' AATGAT ACGGCGACCACCGAGATCTAC A C T C T T T C C C T A C ACGACGCT	CTTCCGATCTGTGCCAGCMGCCGCGGTAA3'		

Reverse	Target gene primer	Reverse read sequencing primer	Barcode	Illumina adapter
806R-Set 1	3' TAATCTWTGGGVH CATCAGTCTAGCCTTCTCGTG T G C AGACTTGA	GGTCAGTG	NNNNNNNNNNNNT	AGAGCATACGGCAGAAG AC GAAC 5'
806R-Set 2	3' TAATCTWTGGGVH CATCAGTTCTAGCCTTCTCGT G T G C AGACTTGA	GGTCAGTG	NNNNNNNNNNNNT	AGAGCATACGGCAGAAG GA CGAAC 5'
806R-Set 3	3' TAATCTWTGGGVH CATCAGTCTCTAGCCTTCTCG T G T G C AGACTTGA	GGTCAGTG	NNNNNNNNNNNNT	AGAGCATACGGCAGAAG AG ACGAAC 5'
806R-Set 4	3' TAATCTWTGGGVH CATCAGTCATCTAGCCTTCTC G T G T G C AGACTTGA	GGTCAGTG	NNNNNNNNNNNNT	AGAGCATACGGCAGAAG GA CGAAC 5'
806R-Set 5	3' TAATCTWTGGGVH CATCAGTCACTCTAGCCTTCT C G T G T G C AGACTTGA	GGTCAGTG	NNNNNNNNNNNNT	AGAGCATACGGCAGAAG GA AGACGAAC 5'
806R-Set 6	3' TAATCTWTGGGVH CATCAGTCACTTCTAGCCTTC T C G T G T G C AGACTTGA	GGTCAGTG	NNNNNNNNNNNNT	AGAGCATACGGC AG AAGACGAA C 5'
806R-Set 7	3' TAATCTWTGGGVH CATCAGTCACTATCTAGCCTT C T C G T G T G C AGACTTGA	GGTCAGTG	NNNNNNNNNNNNT	AGAGCATACGG CA GAAGACGA AC 5'
806R-Set 8	3' TAATCTWTGGGVH CATCAGTCACTAGTCTAGCCT T C T C G T G T G C AGACTTGA	GGTCAGTG	NNNNNNNNNNNNT	AGAGCATACG GC AGAAGACG AAC 5'

Spacers

F. **Second step PCR** using 16S rDNA phasing primers (Each conponents shorted due to space limitation)

G. **Products from the second step PCR**

H. **Miseq sequencing** with 2x250bp paired ends reads (500 cycle kits, version 2)

I. **Sequence reads** (frame-shifted among samples due to the shifts of template DNA in the pooled PCR library, which amplified by different sets of primers with spacers of different number of bases

	Forward Reads	Reverse Reads		Index Reads
Forward-Set 1	5' TCGCATAGTGCC A GCMGCCGCGGTAATACGGAGG G T G C A 3'	3' CCTCGTTTGTCCTAATCTWTGGGVHCATCAG	Reverse-Set 1	5' NNNNNNNNNNNN 3'
Forward-Set 2	5' CGCATAGTGCCA G CMGCCGCGGTAATACGGAGG G T G C A 3'	3' CCTCGTTTGTCCTAATCTWTGGGVHCATCAGT	Reverse-Set 2	5' NNNNNNNNNNNN 3'
Forward-Set 3	5' GCATAGTGCCAG C MGCCGCGGTAATACGGAGGG T G C A 3'	3' C CTCGTTTGTCCTAATCTWTGGGVHCATCAGTC	Reverse-Set 3	5' NNNNNNNNNNNN 3'
Forward-Set 4	5' CATAGTGCCAGC M GCCGCGGTAATACGGAGGGT G C A 3'	3' C C TCGTTTGTCCTAATCTWTGGGVHCATCAGTCA	Reverse-Set 4	5' NNNNNNNNNNNN 3'
Forward-Set 5	5' ATAGTGCCAGCM G CCGCGGTAATACGGAGGGTGC A 3'	3' C C T CGTTTGTCCTAATCTWTGGGVHCATCAGTCAC	Reverse-Set 5	5' NNNNNNNNNNNN 3'
Forward-Set 6	5' TAGTGCCAGCMG C CGCGGTAATACGGAGGGTGC A 3'	3' C C T C GTTTGTCCTAATCTWTGGGVHCATCAGTCACT	Reverse-Set 6	5' NNNNNNNNNNNN 3'
Forward-Set 7	5' AGTGCCAGCMGC C GCGGTAATACGGAGGGTGCA 3'	3' C C T C G TTTGTCCTAATCTWTGGGVHCATCAGTCACTA	Reverse-Set 7	5' NNNNNNNNNNNN 3'
Forward-Set 8	5' GTGCCAGCMGCC G CGGTAATACGGAGGGTGCA 3'	3' C C T C G T TTGTCCTAATCTWTGGGVHCATCAGTCACTAG	Reverse-Set 8	5' NNNNNNNNNNNN 3'

Figure S3. Overview of phasing amplicon sequencing strategy. (a) Standard primer sequences used for amplifying the 16S rRNA genes, yielding an amplicon of 253 bp (not including the primers); (b) The V4 region of the 16S rRNA genes displaying the target primer binding sites and the upstream sequences; (c) First step PCR amplification with standard primers; (d) First step PCR products without upstream sequences; (e) Phasing primers used in the second step PCR amplification of DNA templates for MiSeq sequencing. Spacers of 0-7 bases were added to the 5' end of each of the 8 forward and reverse primers sets. In general, 96 barcoded primers are synthesized for reverse primers, and they are typically divided into 12 groups, each with 8 primer sets. To ensure that the total length of the amplified sequence region does not vary with the primer set used, the forward and reverse primers with spacers were combined in a complementary fashion so that all of the extended primer sets have exactly 7 extra bases. (f) Second step PCR amplification with phasing primers. (g) Products from the second step PCR; (h) MiSeq sequencing, showing sequencing primers of the forward, reverse, and index reads binding to the template DNA generated using phasing primer set 4. Sequencing of the forward and reverse reads starts from the first base of the spacer or the first base of the target primer if there is no spacer. (i) Sequencing reads of all (eight) primer sets (one set for one sample), showing the sequence phasing shift among samples prepared by the different primer sets.

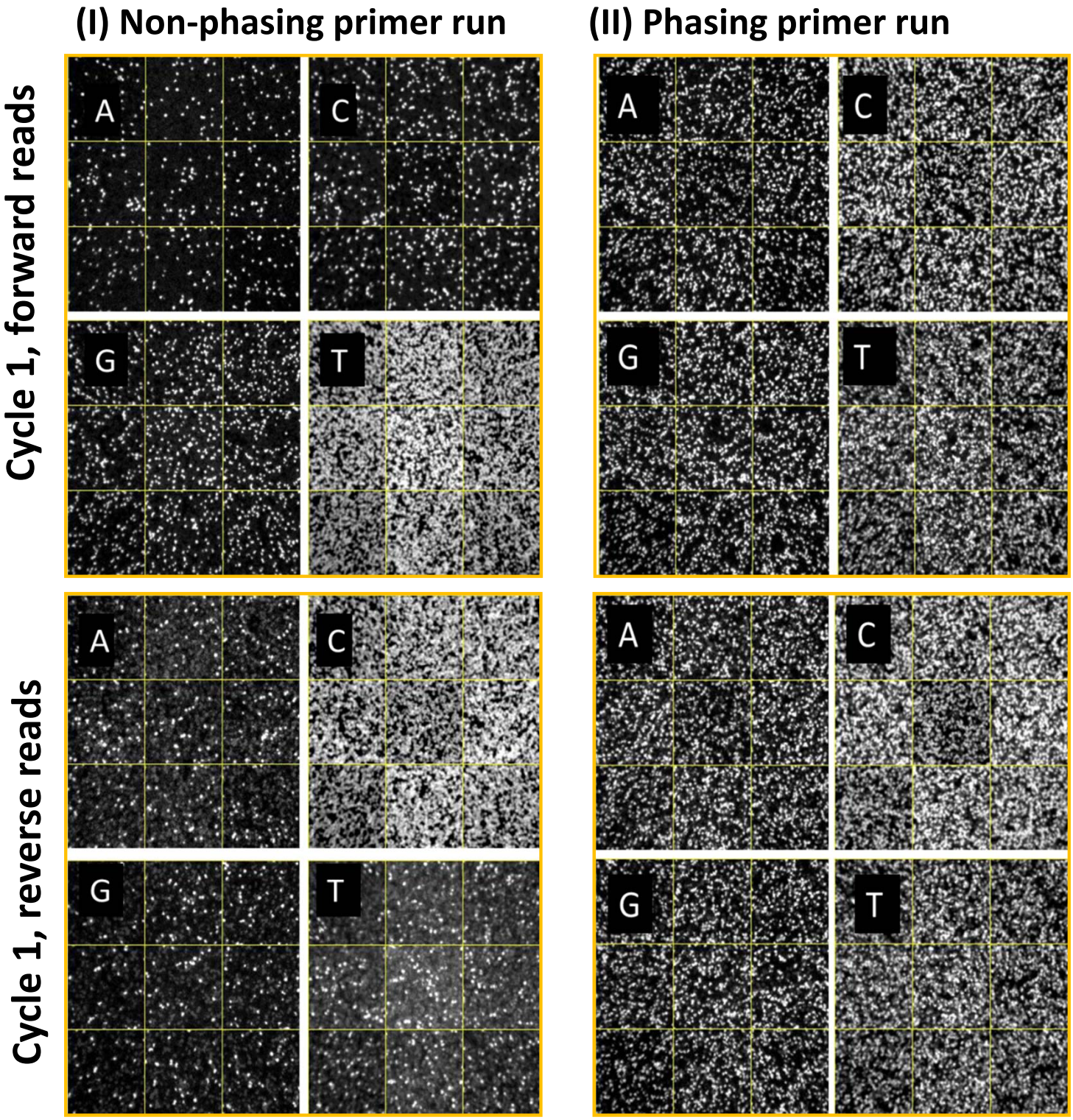


Figure S4. Thumbnail images showing fluorescence signal. Thumbnail images were compared between a non-phase primer run (I) and a phasing primer (II) run. Both runs had moderate cluster densities (~800k/mm^2) and similar amount of spiked PhiX (~15%). (I), Images from the first cycle of the forward reads for the non-phasing run showed that "T" was the dominant base as indicated by the large number of lighted clusters (white dots), while "C" was the dominant base in the first cycle of the reverse reads; (II), images from the first cycle of both forward and reverse reads for the phasing run showing a much more even distribution of fluorescent signal among the four bases (A, T, C, G).

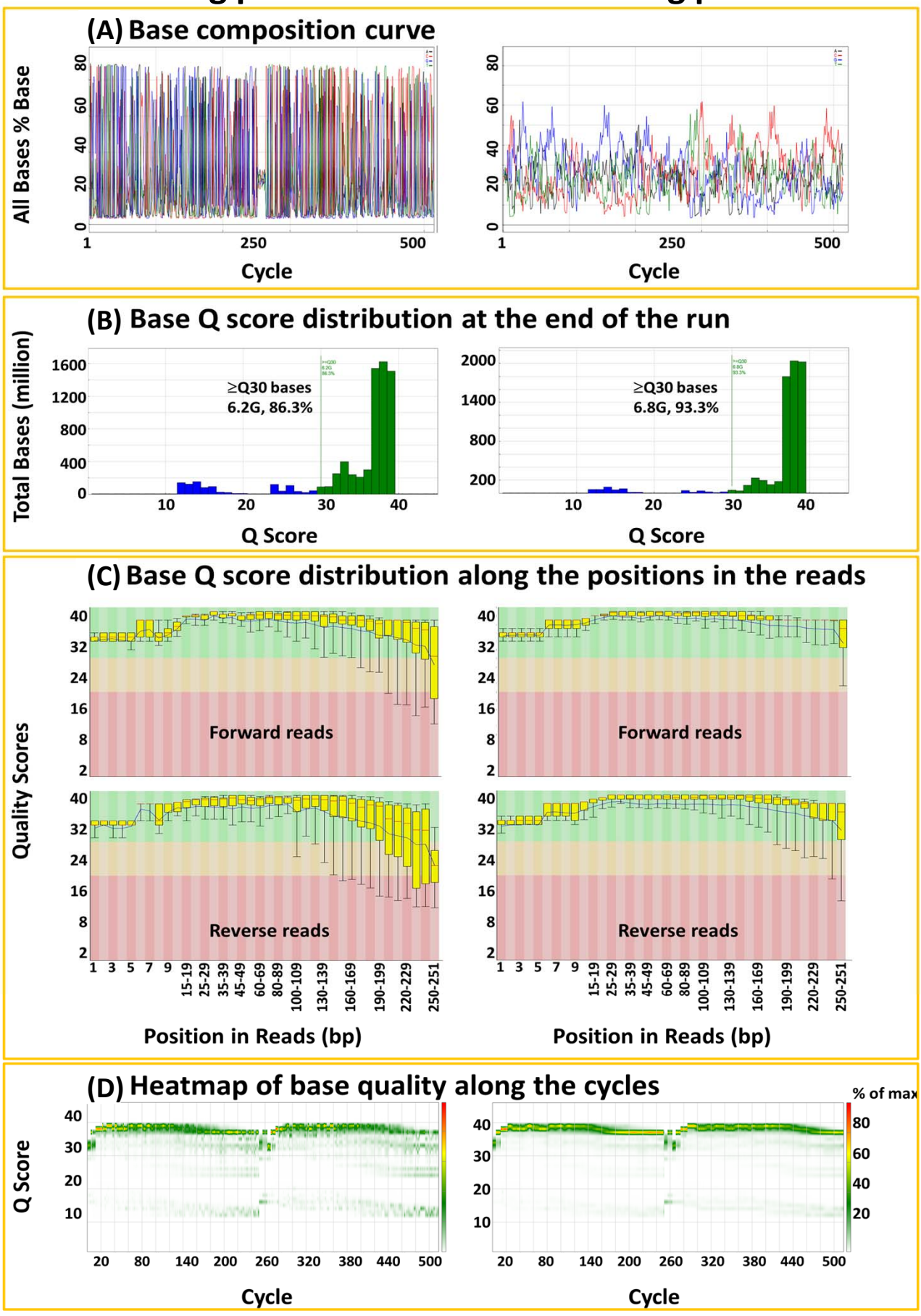

Non-Phasing primer run
Phasing primer run
(A) Base composition curve
All Bases % Base
Cycle
(B) Base Q score distribution at the end of the run
Total Bases (million)
≥Q30 bases
6.2G, 86.3%
≥Q30 bases
6.8G, 93.3%
Q Score
(C) Base Q score distribution along the positions in the reads
Quality Scores
Forward reads
Reverse reads
Position in Reads (bp)
(D) Heatmap of base quality along the cycles
Q Score
% of max
Cycle

Figure S5. Performance comparison between a non-phasing run (left) and a phasing run (right), both with moderate cluster densities (~800k/mm^2) and similar amount of spiked PhiX (~15%). (a) Base distribution curves showing base diversity along all positions of both forward and reverse reads: the observed base distribution for the non-phasing run was uneven while that of the phasing run was more even; (b) base quality distribution at the end of phasing and non-phasing runs: 93.3% of the bases (forward + reverse) in the phasing run had quality scores ≥30, while only 86.3% did in the non-phasing run; (c) base quality distribution as evaluated by FastQC. The green, yellow, and red shaded areas indicate high, acceptable, and low quality. The red line indicated the median quality score for each base position(s), the blue line represents the mean quality score for each base position(s), the yellow boxes represent the inter-quartile range (25-75%) of the quality score for each base position(s), and the upper and lower whiskers represent the 10% and 90% points, respectively. For the phasing run (right), the mean quality score for both forward and reverse reads was well above 32. The 90% interval for the forward reads was above Q30 until the last two bases while the reverse reads were above Q30 until the 170th base position. For the non-phasing run (left), the mean quality scores of the forward reads were above 30 for all positions, but the reverse reads dropped below Q30 at the 200th base position. The 90% interval for the forward reads dropped below Q30 at the 120th base position, while the reverse reads dropped below Q30 at the 100th base position. (d) Heatmaps of base quality over the runs. The phasing primer run had more bases with quality scores ≥30 than the non-phasing primer run.

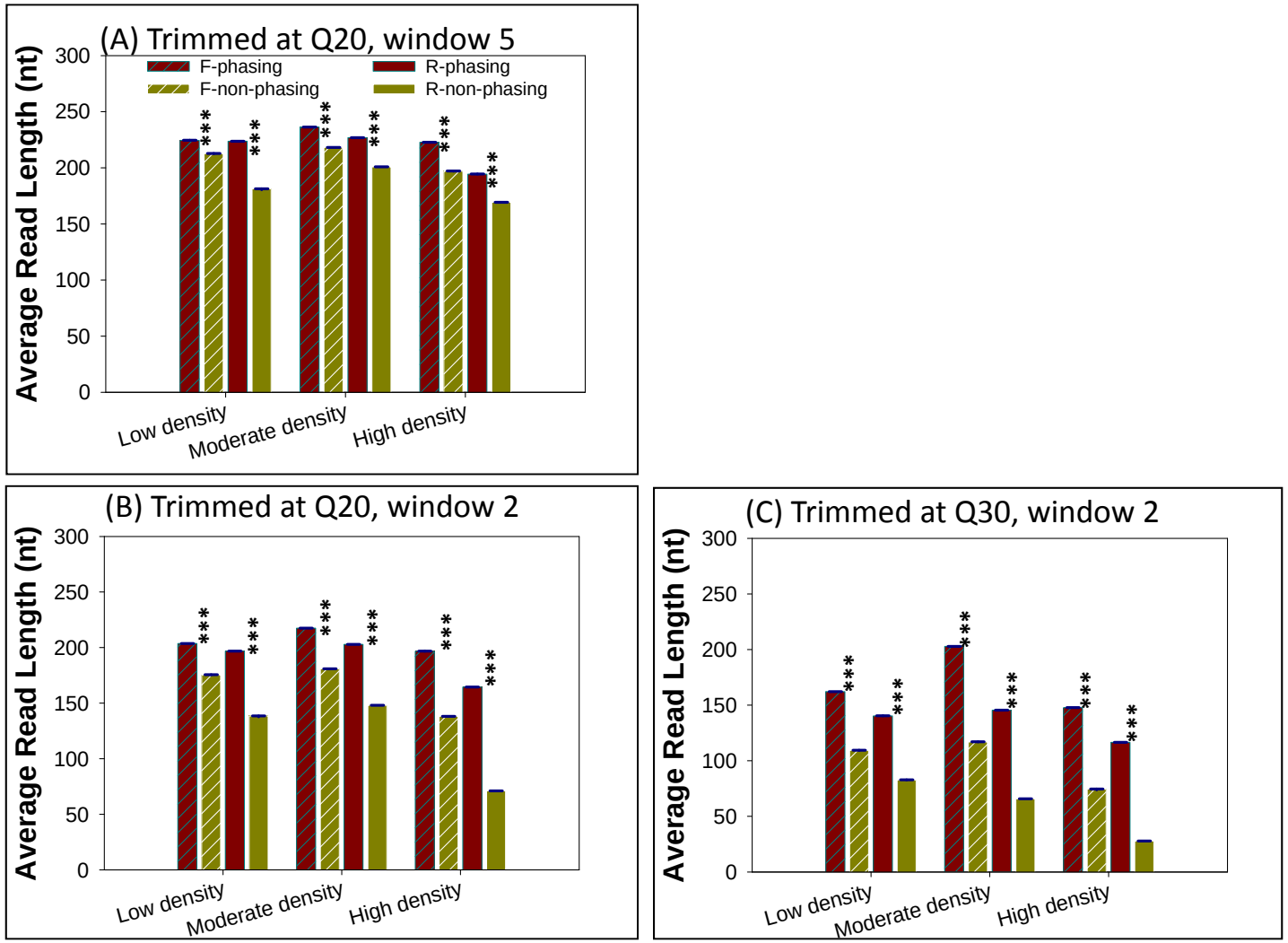


Figure S6. Effects of quality trimming on sequence read length. The average length of all reads after quality trimming were compared between phasing and non-phasing runs at different cluster densities: low, ~400k /mm^2; moderate, ~800k / mm^2; and high, >1000k / mm^2, under different quality trimming stringencies. (a) Q20, trimming window 5; (b) Q20, trimming window 2; (c) Q30, trimming window 2. The Student t test was used to test significance: *, $p<0.05$; **, $p<0.01$; ***, $p<0.001$. F, forward; R, reverse; C, combined.

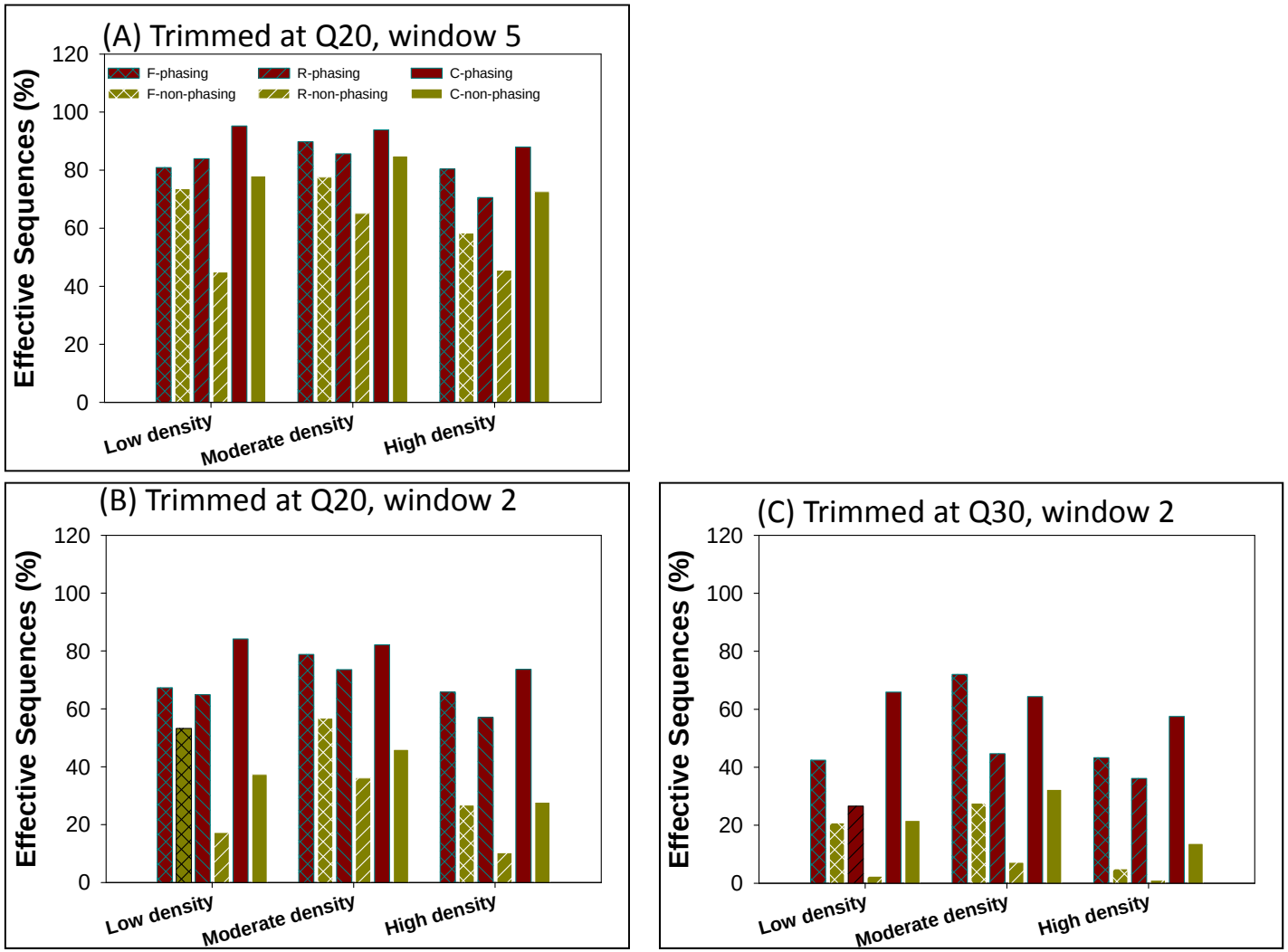


Figure S7. Effects of quality trimming on effective read sequences. The effective read sequences refer to those sequences in which all bases have quality scores >Q30 for at least 80% of the theoretical length (e.g. 200 bp for 2×250 bp paired ends reads). The percentage of effective sequences were compared between phasing and non-phasing runs at different cluster densities (low, ~400k /mm2; moderate, ~800k / mm2; and high, >1000k / mm2) with different quality trimming stringencies: (a) Q20, trimming window 5; (b) Q20, trimming window 2; (c) Q30, trimming window 2. F, forward; R, reverse; C, combined.

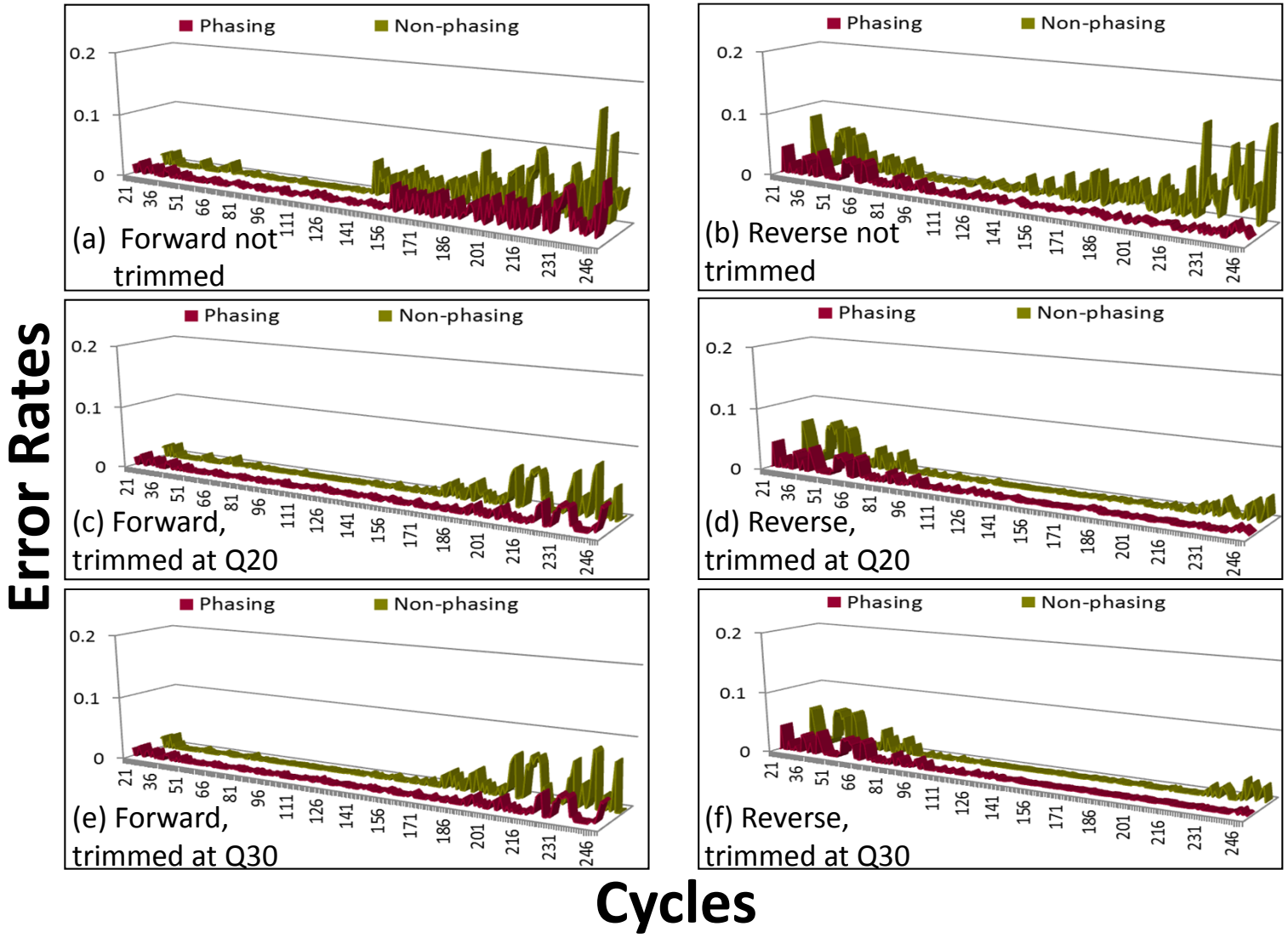


Figure S8. Sequence error rates at each sequence position for the phasing (dark red) and non-phasing (dark yellow) runs. The error rates were determined based on a mock community, which contains full length 16S rDNA sequences of 33 bacteria from different phyla or classes. (a) Raw sequences without quality trimming, forward reads; (b) Raw sequences without quality trimming, reverse reads; (c) Quality trimming with bases below Q20, forward reads; (d) Quality trimming with bases below Q20, reverse reads; (e) Quality trimming with bases below Q30, forward reads; (f) Quality trimming bases with below Q30.

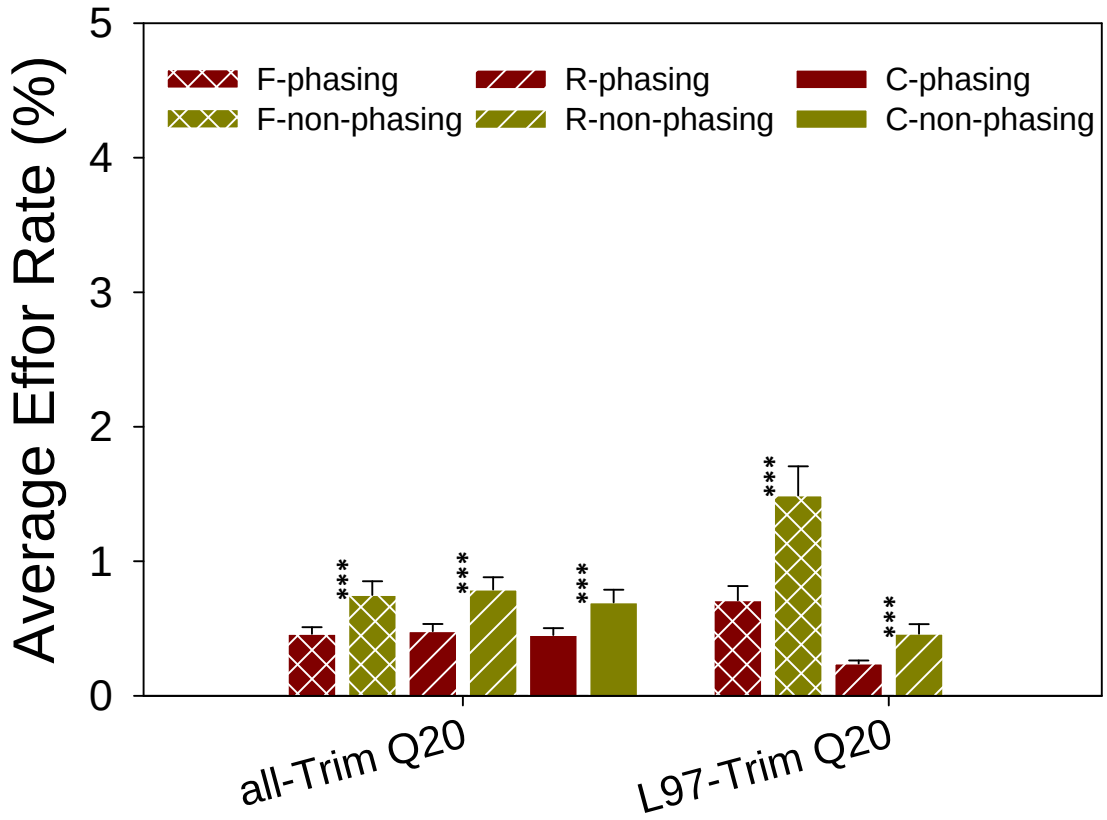


Figure S9. Sequence error rates from mock community sequencing at lower quality trimming standards. The average sequence error rate of the forward and reverse sequence reads with Q20 and window size of 5 were lower for PAS (0.46% for forward, 0.48% for reverse) than non-PAS (0.75% for forward, 0.79% for reverse). The quality differences are more noticeable for the last 97 bases of the forward and reverse reads. The Student t test was used to test significance: *, $p<0.05$; **, $p<0.01$; ***, $p<0.001$. F, forward; R, reverse; C, combined.

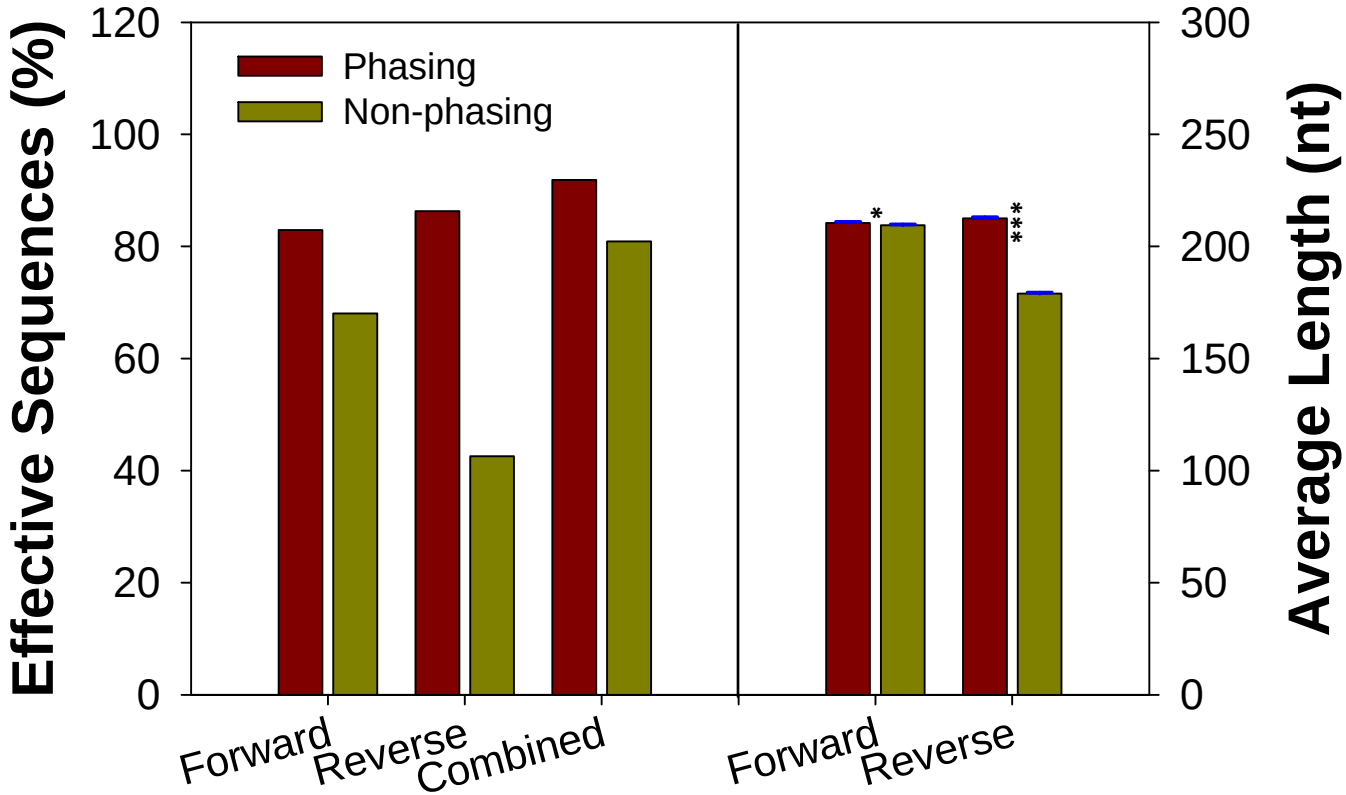


Figure S10. Impact of PAS on the number of effective sequences and the read length at lower quality trimming standards (Q20, window size of 5). Left panel: percentage of effective sequences for forward and reverse reads, and combined sequences. Right panel: average sequence length forward and reverse reads. The Student t test was used to test significance: *, $p<0.05$; **, $p<0.01$; ***, $p<0.001$.

Altering the `MiSeqConfiguration.xml` file to hardcode matrix and phasing

This method requires manual editing of the `MiSeqConfiguration.xml` file used to store RTA options for the MiSeq. It will supply constants for matrix and phasing values for all runs on that MiSeq.

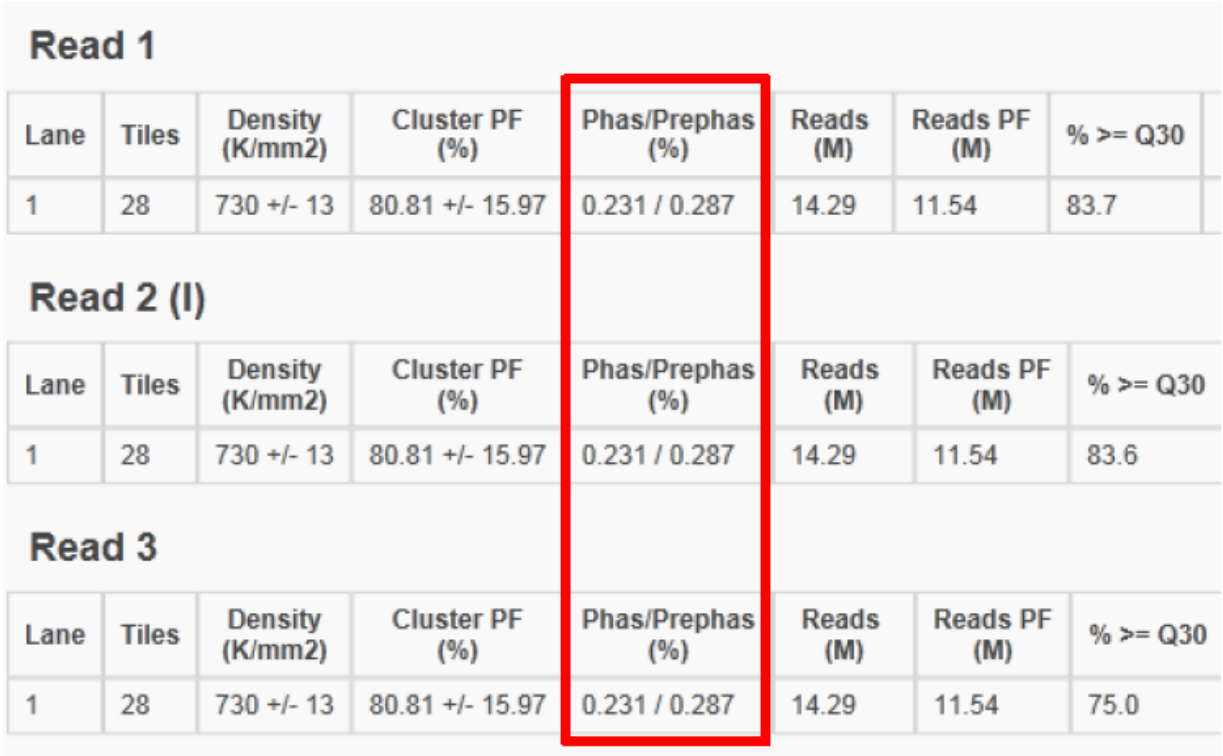
In practice, the best way to use this strategy is to have two versions of `MiSeqConfiguration.xml`, one for normal runs and one for hardcoded runs, and switch between them as necessary. The template files for normal and low-diversity runs are attached, and these instructions will assume that the template file for low diversity is being used. This will avoid errors when editing the XML files. Small mistakes like not using the correct case or leaving out a slash or angle bracket can make this file unusable. If this happens, RTA will not be able to start, and no note of this will be made in the run logs.

When doing these edits, use a text editor like Notepad or Wordpad. Save the files with the XML extension. Do not open using a word processor.

These instructions assume that the run used to generate matrix and phasing files is a paired-end run without an index read. This should be from the same instrument and reagent version as the low diversity run. The two versions of MiSeq reagents can have different phasing and prephasing rates. If major service is done on a MiSeq, a new PhiX run should be done to generate new matrix and phasing files.

If this run is indexed, references to the matrix and phasing files from the second read (normally `s_2_matrix.txt` and `s_2_phasing.txt`) should be read as referring to the last read (`s_3_matrix.txt` and `s_3_phasing.txt` for a single-index read). When entering the values for hardcoding, the index read will be set to the same values as the first read, so an index read on the PhiX run is not necessary.

There is an incompatibility with SAV 1.8.11 (and presumably other versions of SAV) where the phasing values for the last read are displayed for all three reads. The phasing values that are actually used are listed in the run folder, and these do reflect what is in `MiSeqConfiguration.xml`. Please note this display bug.



Read 1

Lane	Tiles	Density (K/mm2)	Cluster PF (%)	Phas/Prephas (%)	Reads (M)	Reads PF (M)	% >= Q30
1	28	730 +/- 13	80.81 +/- 15.97	0.231 / 0.287	14.29	11.54	83.7

Read 2 (I)

Lane	Tiles	Density (K/mm2)	Cluster PF (%)	Phas/Prephas (%)	Reads (M)	Reads PF (M)	% >= Q30
1	28	730 +/- 13	80.81 +/- 15.97	0.231 / 0.287	14.29	11.54	83.6

Read 3

Lane	Tiles	Density (K/mm2)	Cluster PF (%)	Phas/Prephas (%)	Reads (M)	Reads PF (M)	% >= Q30
1	28	730 +/- 13	80.81 +/- 15.97	0.231 / 0.287	14.29	11.54	75.0

You will need:
- RTA 1.14.23 or 1.16.18 installed
- The template `lowdiversity_MiSeqConfiguration.xml` file.
- The run folder from a successful PhiX run.

Proceed as follows:

1) Go `to C:\Illumina\RTA\Configs.`

2) Make a copy of `MiSeqConfiguration.xml` and rename it `standard_MiSeqConfiguration.xml`. This is the backup and the file version for normal runs.

3) Delete `MiSeqConfiguration.xml`.

4) Copy the `lowdiversity_MiSeqConfiguration.xml` template into `C:\Illumina\RTA\Config.`

5) Open `lowdiversity_MiSeqConfiguration.xml` with a text editor.

6) In the PhiX run folder, go to `Data\Intensities\Basecalls\Matrix` and locate the `s_N_matrix.txt` files. For most PhiX runs, there will be two files with file names of `s_1_matrix.txt` and `s_2_matrix.txt` for the two reads.

   a. Indexed runs will have additional files for the additional index reads – i.e. a single-indexed run will have an `s_2_matrix.txt` file for the index read, and the second data read would be `s_3_matrix.txt`. If this is the case, use `s_3_matrix.txt` in place of `s_2_matrix.txt`.

   b. There are many other matrix files that are specific to lane (`s_1_1_matrix.txt`) or tile (`s_1_1_1101_matrix.txt`). You want the lane aggregate files.

7) Open each of the `s_N_matrix.txt` files, one per read of the PhiX run.

   a. These files have 16 numbers in them, each on its own line. These are the 16 elements of a 4x4 matrix of the four color channels.

8) Copy the numbers from `s_1_matrix.txt` into the first `<ArrayOfFloat>` block in the `lowdiversity_MiSeqConfiguration.xml` file. Keep the numbers in the same order, one per line, and be careful to replace the NNN but nothing else. The numbers can be truncated to three significant digits.

9) Copy these same numbers from `s_1_matrix.txt` into the second `<ArrayOfFloat>` block for the index read.

10) Copy the numbers from `s_2_matrix.txt` into the third `<ArrayOfFloat>` block for the second data read.

11) Close the `s_N_matrix.txt` files.

12) Go into `Data\Intensities\Basecalls\Phasing` and locate the `s_N_phasing.txt` files. Similar to the matrix files, there will be lane- and tile-specific phasing files, but you want the read aggregate files.

13) Open each of the `s_N_phasing.txt` files, one per read of the PhiX run.

a. These files have two numbers in them, each on its own line. The top number is the phasing rate for that read, and the bottom number is the prephasing rate. The numbers are raw rates – to convert to the more familiar percentages, multiply by 100 in your head.

14) Copy the first number from `s_1_phasing.txt` into the first line of the `<HardCodedPhasing>` block of `lowdiversity_MiSeqConfiguration.xml`, replacing the NNN but nothing else. Three significant digits are sufficient. This is phasing for the first read.

15) Copy that same number into the second line of the `<HardCodedPhasing>` block. This is phasing for the index read.

16) Copy the first number from `s_2_phasing.txt` into the third line of the `<HardCodedPhasing>` block. This is phasing for the second data read.

17) Copy the second number from `s_1_phasing.txt` into the first line of the `<HardCodedPrePhasing>` block, and again for the second line of the `<HardCodedPrePhasing>` block. This is prephasing for the first read and index read.

18) Copy the second number from `s_2_phasing.txt` into the third line of the `<HardCodedPrePhasing>` block. This is prephasing for the second data read.

19) Check the file to make sure all of the XML elements are intact.

20) Save `lowdiversity_MiSeqConfiguration.xml`.

Once this is done, one or the other of the two `MiSeqConfiguration.xml` files (either `standard_MiSeqConfiguration.xml` or `lowdiversity_MiSeqConfiguration.xml`) can be copied and renamed to `MiSeqConfiguration.xml`. This should be done before the start of the run.