Supplementary material

KMC3 and CHTKC: Best Scenarios, Deficiencies, and Challenges in High-Throughput Sequencing Data Analysis

Table S1. Comparison of KMC3 and CHTKC with various memory limit for H.1 and H.2 (k=32).

Algorithm	Dataset	Memory (GB)	Running Time (s)	Memory accesses (Billion)	CPU utilization rate	Peak disk (GB)	Branch prediction miss rate	Cache miss rate
CHTKC	H1	16	3478.14	32.8	31.6%	56.6	3.5%	59.4%
		24	2555.36	23.0	31.7%	55.2	3.6%	64.1%
		32	2114.49	18.3	31.9%	53.7	3.7%	67.6%
		48	1703.67	14.0	31.9%	51.8	3.7%	71.5%
		64	1513.68	12.0	32.0%	47.7	3.8%	73.5%
		96	1264.34	9.8	32.1%	40.0	3.8%	76.8%
		128	1201.34	9.8	32.1%	40.0	3.5%	75.8%
		192	1204.74	9.9	31.8%	40.0	3.2%	72.1%
		256	1208.25	9.9	31.6%	40.0	3.2%	71.0%
	H2	16	2332.62	24.2	31.4%	63.4	3.6%	60.8%
		24	1769.22	17.5	31.8%	60.8	3.8%	65.2%
		32	1530.84	14.3	31.9%	56.7	3.9%	68.4%
		48	1278.91	11.4	32.5%	47.9	4.0%	72.5%
		64	1167.65	9.9	32.6%	40.0	4.1%	75.3%
		96	1093.66	9.2	32.6%	34.7	4.0%	77.9%
		128	1095.06	9.2	32.5%	34.7	3.8%	77.1%
		192	1119.54	9.4	32.3%	34.7	3.7%	73.5%
		256	1163.09	9.4	31.3%	34.7	3.6%	71.3%
KMC3	H.1	2	1470.64	15.9	7.7%	87.5	/	/
		4	1299.27	15.3	8.4%	87.5	/	/
		8	1008.82	15.2	9.0%	87.5	/	/
		16	846.64	15.2	9.8%	87.5	2.9%	77.4%
		24	828.66	15.2	9.8%	87.5	2.9%	78.0%
		32	777.04	15.2	10.4%	87.5	2.9%	77.6%
		48	801.99	15.3	10.5%	87.5	2.9%	77.5%
		64	793.09	15.3	11.4%	87.5	2.9%	78.5%
		96	887.16	15.3	10.3%	87.5	2.9%	77.3%
		128	1016.62	15.4	9.1%	87.5	2.9%	76.7%
		192	1079.05	15.4	9.0%	87.5	2.9%	75.0%
		256	1318.80	15.4	5.9%	87.5	2.9%	75.4%
	H2	2	1140.75	16.2	10.4%	93.4	/	/
		4	1028.20	17.4	11.7%	93.4	/	/
		8	730.82	17.4	13.7%	93.4	/	/
		16	576.15	17.5	16.5%	93.4	2.9%	73.0%

		24	561.49	17.5	16.5%	93.4	2.9%	78.3%
		32	545.38	17.5	18.4%	93.4	2.9%	78.8%
		48	556.49	17.5	17.1%	93.4	2.9%	78.1%
		64	556.61	17.5	18.1%	93.4	2.9%	78.9%
		96	619.27	17.5	16.2%	93.4	2.9%	77.7%
		128	667.86	17.6	16.3%	93.4	2.9%	77.7%
		192	773.55	17.7	13.6%	93.4	2.9%	75.9%
		256	1105.92	17.6	8.0%	93.4	2.9%	76.8%

Table S2. CHTKC's correlation between the running time and the ratio of the number of distinct k-mers to the number of total k-mers. The total k-mers is set to 1 billion.

Distinct k-mers	Batches	Time(s)
110M	1	3.91
124M	1	4.57
142M	1	4.84
165M	2	6.63
199M	2	7.87
248M	2	10.38
250M	2	10.28
251M	2	10.48
252M	2	11.34
254M	2	11.71
255M	2	10.74
331M	3	14.19
364M	3	16.16
397M	3	18.23
430M	3	17.75
463M	3	23.34
496M	4	23.02

Table S3. Comparison of KMC3 and KMC3-R with various memory limit (k=32).

Algorithm	Dataset	Memory (GB)	Actual used memory (GB)	Running time (s)	Stage 1 time (s)	Stage 2 time (s)
KMC3	H1	4	3.9	1385.19	558.40	751.58
		8	7.6	1049.2	552.77	488.14
		16	15.0	880.76	540.27	346.13
		32	29.9	804.55	552.49	251.78
		64	59.6	822.86	568.05	254.32
		128	119.3	1058.16	809.64	232.23
		256	236.0	1392.27	792.86	437.61
	H2	4	3.9	1069.14	338.51	796.10

		8	7.6	718.5	214.34	494.82
		16	15.0	585.48	236.63	334.66
		32	29.9	537.71	263.98	264.23
		64	59.9	573.91	318.25	260.63
		128	118.7	667.12	416.07	286.10
		256	238.2	1148.45	487.26	419.52
KMC3-R	H1	4	92.0	1347.7	544.86	755.18
		8	92.2	1034.51	542.82	493.76
		16	99.2	885.97	543.10	327.18
		32	112.3	843.11	552.45	254.64
		64	139.4	803.63	562.79	244.93
		128	174.7	922.64	646.32	260.59
		256	274.4	914.85	679.33	282.49
	H2	4	98.6	953.44	211.69	784.47
		8	100.9	667.44	134.54	508.17
		16	106.2	485.16	158.68	344.99
		32	118.6	533.3	213.21	287.83
		64	148.5	533.0	231.01	260.20
		128	180.4	551.53	242.78	255.22
		256	296.9	618.52	260.78	253.85

Table S4. Comparison of KMC3 and CHTKC using various GC-ratio datasets. α is the ratio of the number of distinct k-mers to the number of total k-mers (k=32).

Dataset	α	Algorithm	Time per 32-mer (ns)
S-GC21	0.132	KMC3	14.61
S-GC21	0.132	CHTKC	18.80
A-GC50	0.128	KMC3	14.54
A-GC50	0.128	CHTKC	19.79
B-GC25	0.124	KMC3	11.84
B-GC25	0.124	CHTKC	12.77
T-GC45	0.123	KMC3	11.64
T-GC45	0.123	CHTKC	12.49
T-GC30	0.110	KMC3	11.41
T-GC30	0.110	CHTKC	11.86
D-GC36	0.110	KMC3	11.36
D-GC36	0.110	CHTKC	11.84

Table S5. Comparison of KMC3 and CHTKC with various sequencing depth.

k	Depth	Algorithm	Cache-reference	Cache-miss-rate	Branch-miss-rate	Time (s)	Time per k-mer (ns)
16	5	CHTKC	54.04G	63.7%	3.5%	161.12	11.6
		KMC3	19.08G	67.4%	6.3%	139.27	10.0

	10	CHTKC	106.45G	64.1%	3.5%	316.87	11.4
		KMC3	38.96G	65.9%	6.0%	304.13	10.9
	30	CHTKC	314.35G	66.4%	3.6%	941.85	11.3
		KMC3	117.78G	66.8%	5.7%	1021.20	12.2
	60	CHTKC	633.61G	67.4%	3.6%	1913.96	11.5
		KMC3	231.43G	68.5%	5.7%	2433.99	14.6
32	5	CHTKC	50.46G	73.2%	3.5%	167.00	14.8
		KMC3	19.84G	72.2%	4.6%	134.60	11.9
	10	CHTKC	100.08G	73.4%	3.8%	310.40	13.7
		KMC3	41.32G	70.8%	4.1%	262.80	11.6
	30	CHTKC	356.63G	75.6%	3.7%	1050.82	**15.5**
		KMC3	118.41G	76.5%	3.3%	819.47	12.1
	60	CHTKC	924.86G	75.5%	3.5%	2619.28	**19.3**
		KMC3	240.66G	78.0%	2.6%	1775.03	13.1
64	5	CHTKC	35.15G	74.5%	4.0%	119.79	19.8
		KMC3	19.92G	74.8%	5.6%	130.72	21.6
	10	CHTKC	79.77G	71.7%	4.2%	253.21	20.9
		KMC3	40.96G	75.2%	5.3%	263.12	21.7
	30	CHTKC	264.05G	76.8%	4.1%	841.40	**23.2**
		KMC3	116.58G	80.9%	4.6%	778.74	21.4
	60	CHTKC	724.36G	75.8%	3.8%	2507.24	**34.5**
		KMC3	239.91G	81.8%	4.2%	1573.78	21.7

Table S6. Comparison of KMC3 and CHTKC with various sequencing errors.

Algorithm & k	Dataset	Running time (s)	Cache references (Billion)
CHTKC (k = 16)	H-69	954.68	212.0
CHTKC (k = 16)	H-75	937.41	208.9
CHTKC (k = 16)	H-80	920.92	204.9
CHTKC (k = 16)	H-86	905.27	201.7
CHTKC (k = 16)	H-91	882.72	196.5
CHTKC (k = 16)	H-96	889.46	194.1
CHTKC (k = 16)	H-100	861.04	189.7
KMC3 (k = 16)	H-69	1079.52	80.2
KMC3 (k = 16)	H-75	1064.36	79.9
KMC3 (k = 16)	H-80	1061.47	78.7
KMC3 (k = 16)	H-86	1126.09	77.7
KMC3 (k = 16)	H-91	1025.92	78.2
KMC3 (k = 16)	H-96	979.87	79.2
KMC3 (k = 16)	H-100	798.56	77.9
CHTKC (k = 32)	H-69	1232.47	305.1
CHTKC (k = 32)	H-75	1082.86	267.1
CHTKC (k = 32)	H-80	1002.08	245.6

CHTKC (k = 32)	H-86	857.60	212.0
CHTKC (k = 32)	H-91	793.07	191.4
CHTKC (k = 32)	H-96	770.82	186.6
CHTKC (k = 32)	H-100	751.93	182.6
KMC3 (k = 32)	H-69	865.00	93.2
KMC3 (k = 32)	H-75	853.41	93.2
KMC3 (k = 32)	H-80	906.28	92.5
KMC3 (k = 32)	H-86	844.56	92.9
KMC3 (k = 32)	H-91	844.82	93.0
KMC3 (k = 32)	H-96	805.96	92.0
KMC3 (k = 32)	H-100	573.17	91.7
CHTKC (k = 64)	H-69	995.05	237.2
CHTKC (k = 64)	H-75	861.46	204.5
CHTKC (k = 64)	H-80	802.44	184.9
CHTKC (k = 64)	H-86	675.22	157.7
CHTKC (k = 64)	H-91	595.54	135.7
CHTKC (k = 64)	H-96	569.21	129.0
CHTKC (k = 64)	H-100	506.85	123.7
KMC3 (k = 64)	H-69	893.45	96.1
KMC3 (k = 64)	H-75	851.36	96.3
KMC3 (k = 64)	H-80	811.57	97.7
KMC3 (k = 64)	H-86	885.08	95.5
KMC3 (k = 64)	H-91	804.03	98.3
KMC3 (k = 64)	H-96	729.10	97.3
KMC3 (k = 64)	H-100	512.74	96.2

Table S7. Comparison of the time usage of different stages of KMC3 with various sequencing errors.

k	dataset	Stage 1 (s)	Stage 2 (s)
16	H-69	921.249	155.047
	H-75	909.624	151.100
	H-80	905.916	152.268
	H-86	900.539	221.955
	H-91	869.927	151.414
	H-96	836.725	139.119
	H-100	647.830	147.026
32	H-69	677.999	179.735
	H-75	668.138	188.980
	H-80	669.551	191.800
	H-86	653.806	193.173
	H-91	646.541	195.335
	H-96	603.211	204.101
	H-100	364.219	181.669
64	H-69	610.041	282.416

	H-75	608.527	241.974
	H-80	603.483	207.470
	H-86	594.733	289.499
	H-91	575.173	228.260
	H-96	541.705	186.638
	H-100	325.719	186.442

Table S8. Strong scalability comparison of KMC3 and CHTKC.

Dataset	Algorithm & k	Threads	Speedup	Clocks per instruction	Cache miss rate	Branch prediction miss rate
H.1	CHTKC (k = 16)	4	1	5.05	72.6%	2.9%
H.1	CHTKC (k = 16)	8	3.26	5.44	63.7%	3.4%
H.1	CHTKC (k = 16)	16	7.95	5.17	66.5%	3.3%
H.1	CHTKC (k = 16)	32	16.92	5.05	67.1%	3.2%
H.1	CHTKC (k = 32)	4	1	4.61	83.2%	2.8%
H.1	CHTKC (k = 32)	8	3.52	5.17	70.9%	3.5%
H.1	CHTKC (k = 32)	16	8.45	4.82	72.9%	3.5%
H.1	CHTKC (k = 32)	32	18.38	4.58	76.4%	3.5%
H.1	CHTKC (k = 64)	4	1	4.24	84.6%	2.9%
H.1	CHTKC (k = 64)	8	2.64	6.07	55.4%	3.5%
H.1	CHTKC (k = 64)	16	8.20	4.31	72.6%	3.6%
H.1	CHTKC (k = 64)	32	17.45	4.20	76.4%	3.6%
H.1	KMC3 (k = 16)	4	1	0.71	67.0%	5.5%
H.1	KMC3 (k = 16)	8	2.72	0.71	68.1%	5.5%
H.1	KMC3 (k = 16)	16	3.32	0.73	67.7%	5.5%
H.1	KMC3 (k = 16)	32	3.38	0.76	67.5%	5.5%
H.1	KMC3 (k = 32)	4	1	0.61	75.1%	2.9%
H.1	KMC3 (k = 32)	8	2.33	0.61	76.2%	2.9%
H.1	KMC3 (k = 32)	16	3.02	0.64	77.2%	2.9%
H.1	KMC3 (k = 32)	32	3.18	0.80	76.2%	2.9%
H.1	KMC3 (k = 64)	4	1	0.72	81.8%	4.0%
H.1	KMC3 (k = 64)	8	2.35	0.72	82.6%	4.1%
H.1	KMC3 (k = 64)	16	3.10	0.82	83.1%	4.1%
H.1	KMC3 (k = 64)	32	3.23	1.11	82.8%	4.1%
H.2	CHTKC (k = 16)	4	1	4.93	72.5%	3.1%
H.2	CHTKC (k = 16)	8	3.23	5.26	64.8%	3.7%
H.2	CHTKC (k = 16)	16	7.98	4.92	65.9%	3.6%
H.2	CHTKC (k = 16)	32	16.91	4.82	66.2%	3.4%
H.2	CHTKC (k = 32)	4	1	4.48	83.3%	3.1%
H.2	CHTKC (k = 32)	8	3.48	4.89	73.7%	3.8%
H.2	CHTKC (k = 32)	16	9.23	4.39	76.9%	3.8%
H.2	CHTKC (k = 32)	32	19.24	4.32	77.4%	3.8%

H.2	CHTKC (k = 64)	4	1	3.47	85.5%	3.5%
H.2	CHTKC (k = 64)	8	2.85	4.37	56.9%	4.1%
H.2	CHTKC (k = 64)	16	8.71	3.23	76.0%	4.2%
H.2	CHTKC (k = 64)	32	17.57	3.11	78.5%	4.1%
H.2	KMC3 (k = 16)	4	1	0.72	68.2%	5.6%
H.2	KMC3 (k = 16)	8	2.51	0.72	69.4%	5.6%
H.2	KMC3 (k = 16)	16	3.57	0.73	68.7%	5.7%
H.2	KMC3 (k = 16)	32	3.79	0.74	69.0%	5.7%
H.2	KMC3 (k = 32)	4	1	0.60	75.7%	2.9%
H.2	KMC3 (k = 32)	8	2.52	0.60	77.9%	2.9%
H.2	KMC3 (k = 32)	16	4.29	0.64	78.4%	2.9%
H.2	KMC3 (k = 32)	32	5.68	0.81	76.3%	2.9%
H.2	KMC3 (k = 64)	4	1	0.71	80.5%	4.3%
H.2	KMC3 (k = 64)	8	2.67	0.72	80.7%	4.3%
H.2	KMC3 (k = 64)	16	4.77	0.81	82.0%	4.3%
H.2	KMC3 (k = 64)	32	5.48	0.92	81.4%	4.3%

Table S9. Weak scalability comparison of KMC3 and CHTKC.

Dataset	Threads	Algorithm & k	Time (s)	Clocks per instruction	Cache miss rate	Branch prediction miss rate
H (86%) * 1	4	CHTKC (k = 16)	15630.55	4.99	71.1%	3.3%
H (86%) * 1	4	CHTKC (k = 32)	15488.00	4.58	81.7%	3.4%
H (86%) * 1	4	CHTKC (k = 64)	10670.50	3.61	84.4%	3.9%
H (86%) * 1	4	KMC3 (k = 16)	3382.17	0.74	67.9%	5.6%
H (86%) * 1	4	KMC3 (k = 32)	2365.18	0.63	74.8%	3.3%
H (86%) * 1	4	KMC3 (k = 64)	2206.46	0.73	78.8%	4.4%
H (86%) * 2	8	CHTKC (k = 16)	9878.45	5.34	64.7%	3.9%
H (86%) * 2	8	CHTKC (k = 32)	8699.36	4.94	73.5%	4.2%
H (86%) * 2	8	CHTKC (k = 64)	7990.86	4.77	66.3%	4.6%
H (86%) * 2	8	KMC3 (k = 16)	2955.87	0.72	69.5%	5.7%
H (86%) * 2	8	KMC3 (k = 32)	2417.72	0.59	79.2%	2.5%
H (86%) * 2	8	KMC3 (k = 64)	2071.99	0.68	82.5%	3.9%
H (86%) * 4	16	CHTKC (k = 16)	7982.25	5.01	66.0%	3.7%
H (86%) * 4	16	CHTKC (k = 32)	6857.55	4.63	77.0%	4.2%
H (86%) * 4	16	CHTKC (k = 64)	4784.17	3.38	77.2%	4.7%
H (86%) * 4	16	KMC3 (k = 16)	4748.59	0.73	70.8%	5.7%
H (86%) * 4	16	KMC3 (k = 32)	3331.70	0.64	79.7%	2.8%
H (86%) * 4	16	KMC3 (k = 64)	2490.05	0.67	84.2%	3.2%
H (86%) * 8	32	CHTKC (k = 16)	7478.97	4.93	68.2%	3.7%
H (86%) * 8	32	CHTKC (k = 32)	6307.99	4.51	80.0%	4.2%
H (86%) * 8	32	CHTKC (k = 64)	4529.93	3.37	81.4%	4.8%
H (86%) * 8	32	KMC3 (k = 16)	9230.04	0.73	70.4%	5.8%

H (86%) * 8	32	KMC3 (k = 32)	7080.60	0.78	80.1%	3.0%
H (86%) * 8	32	KMC3 (k = 64)	5772.04	0.89	83.4%	3.5%

Table S10. Comparison of the dumper of KMC3, CHTKC and our PBKMC with various sequencing errors.

Dataset	k	Algorithm	Decode time (s)
H-69	32	CHTKC	2311.15
H-69	32	KMC3	2030.98
H-69	32	Revised KMC3	2000.1
H-75	32	CHTKC	1688.83
H-75	32	KMC3	1589.97
H-75	32	Revised KMC3	1499.37
H-80	32	CHTKC	1350.53
H-80	32	KMC3	1222.06
H-80	32	Revised KMC3	1150.47
H-86	32	CHTKC	1027.35
H-86	32	KMC3	823.11
H-86	32	Revised KMC3	752.18
H-69	64	CHTKC	5157.64
H-69	64	KMC3	4452.98
H-69	64	Revised KMC3	4342.47
H-75	64	CHTKC	4314.38
H-75	64	KMC3	3546.47
H-75	64	Revised KMC3	3501.45
H-80	64	CHTKC	3587.74
H-80	64	KMC3	2991.38
H-80	64	Revised KMC3	2907.8
H-86	64	CHTKC	2353.41
H-86	64	KMC3	2226.65
H-86	64	Revised KMC3	2083.01