ntCard: A streaming algorithm for cardinality estimation in genomics data

Hamid Mohamadi, Hamza Khan, Inanc Birol
Canada's Michael Smith Genome Sciences Centre, British Columbia Cancer Agency, Vancouver, BC, Canada
University of British Columbia, Vancouver, BC, V6T 1Z4, Canada

Supplementary Data

Software comparison

We have compared ntCard with KmerGenie, KmerStream, Khmer and DSK algorithms. We have used KmerGenie version 1.7016, KmerStream version 1.1, Khmer version 2.0, DSK version 2.1.0, and ntCard version 1.0.0 in our experiments.

KmerGenie is first proposed method for estimating the full *k*-mer coverage frequencies histograms. It samples *k*-mers to approximate the frequency histogram of *k*-mer occurrences. It also offers a fast heuristic for putative *k* values to estimate the best possible value of *k*.

KmerStream is an algorithm for estimating the number of distinct *k*-mers, F_0, as well as the number of singletons, f_1, present in high-throughput sequencing data. The runtime of KmerStream is linear with the size of the input data and the space requirement is logarithmic in the size of the input.

Khmer is an open implementation of the HyperLogLog cardinality estimation sketch for *k*-mers implemented in C++ with a Python interface, and is distributed as part of the khmer software package. Khmer gives only the estimated number of distinct *k*-mers, F_0, in a dataset.

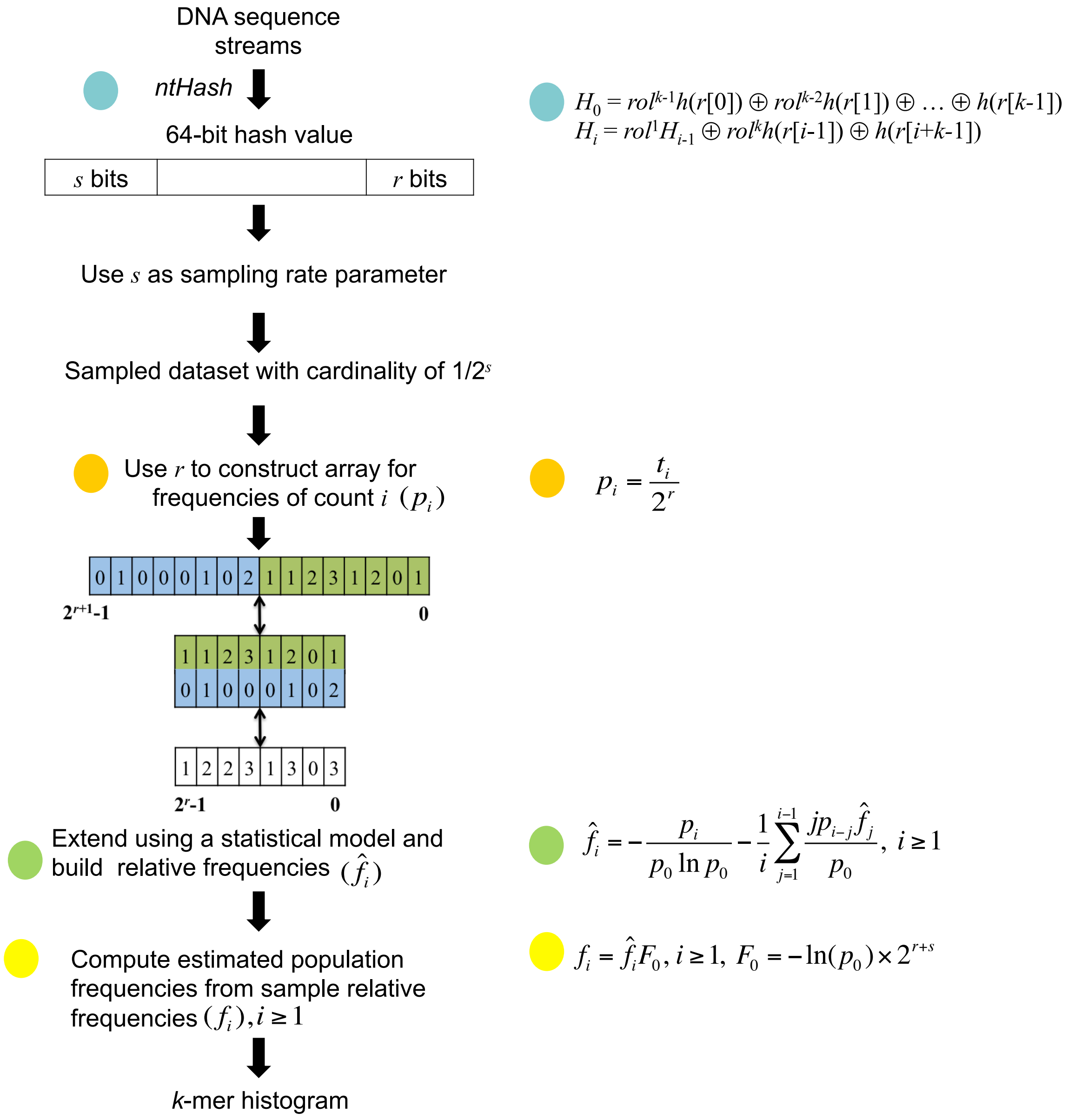
DSK is a disk-based streaming algorithm for *k*-mer counting which works by first partitioning and storing the multi-set of all *k*-mers present in the reads, and then loading and counting each partition separately.

All four tools are run with their default parameters, and the parameters related to the resource usage are set in a way to utilize the maximum capacity on each computing node as described in Supplementary Data. For example, all tools are run in multi-threaded mode with the maximum number of threads.

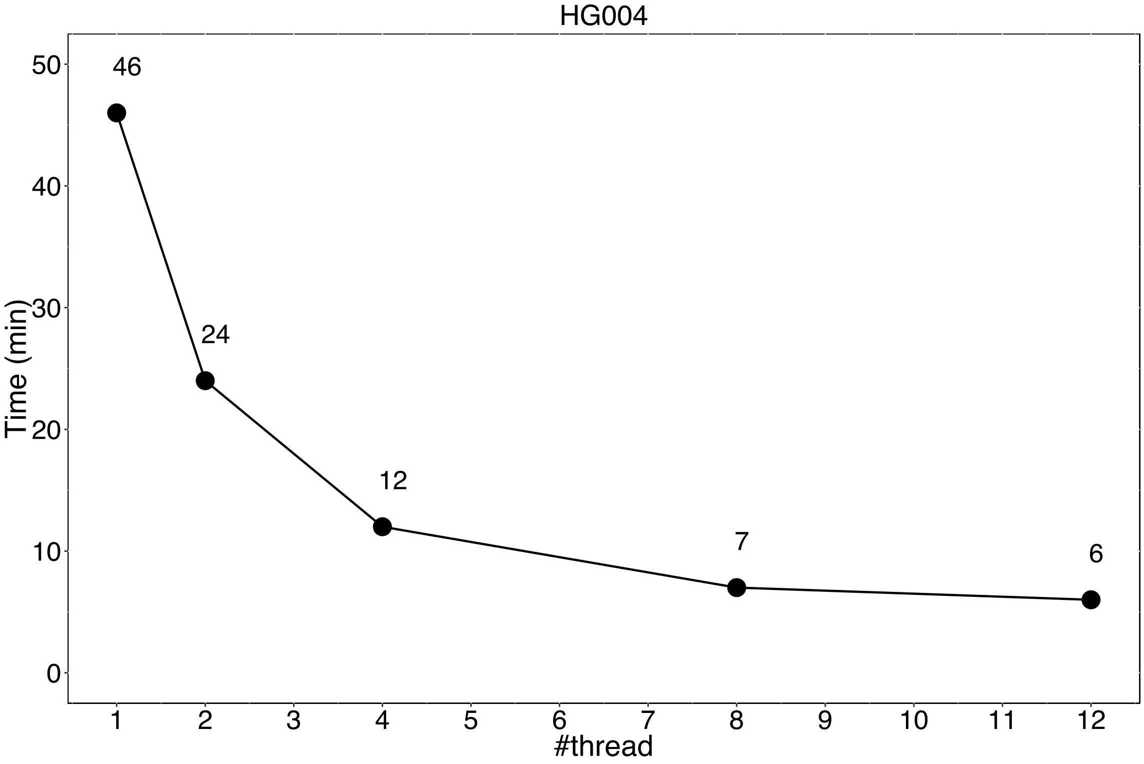
To run each tool, we used the following command:

- KmerGenie http://kmergenie.bx.psu.edu/
```
/usr/bin/time -v specialk <inputs> -k 128 –l 32 –s 32 –t 12 // multi k
/usr/bin/time -v specialk <inputs> -k <int> –l <int> // single k
```
- Khmer https://github.com/dib-lab/khmer
```
/usr/bin/time -v unique-kmers.py -k <int> <inputs>
-k: size of k-mer. We have used k={32,64,96,128}
for the number of threads, we have set OMP_NUM_THREADS=12
```
- KmerStream: https://github.com/pmelsted/KmerStream
```
/usr/bin/time -v KmerStream -k <int> -t <int> -o <output> <inputs>
-k: size of k-mer. We have used k={32,64,96,128}
-t: number of threads to use. We have used t=12
```
- DSK: http://minia.genouest.org/dsk/
```
/usr/bin/time -v dsk -nb-cores <t> -kmer-size <int> -file <inputs> -out $output
-kmer-size: size of k-mer. We have used k={32,64,96,128}
-nb-cores: number of cores. We have used t=12
```
- ntCard: https://github.com/bcgsc/ntCard
```
/usr/bin/time -v ntcard -k <int> -t <int> <inputs>
-k: size of k-mer. We have used k={32,64,96,128}
-t: number of threads to use. We have used t=12
```

All details for using and running the ntCard tool have been explained in the github page:
https://github.com/bcgsc/ntCard


DNA sequence streams
ntHash
64-bit hash value
s bits
r bits
Use s as sampling rate parameter
Sampled dataset with cardinality of $1/2^s$
Use r to construct array for frequencies of count i (p_i)
0 1 0 0 0 1 0 2 1 1 2 3 1 2 0 1
2^{r+1}-1
0
1 1 2 3 1 2 0 1
0 1 0 0 0 1 0 2
1 2 2 3 1 3 0 3
2^r-1
0
Extend using a statistical model and build relative frequencies $(\hat{f}_i)$
Compute estimated population frequencies from sample relative frequencies $(f_i), i \geq 1$
k-mer histogram
$H_0 = rol^{k-1}h(r[0]) \oplus rol^{k-2}h(r[1]) \oplus \ldots \oplus h(r[k-1])$
$H_i = rol^1 H_{i-1} \oplus rol^k h(r[i-1]) \oplus h(r[i+k-1])$
$p_i = \frac{t_i}{2^r}$
$\hat{f}_i = -\frac{p_i}{p_0 \ln p_0} - \frac{1}{i}\sum_{j=1}^{i-1}\frac{jp_{i-j}\hat{f}_j}{p_0}, \ i \geq 1$
$f_i = \hat{f}_i F_0, i \geq 1, \ F_0 = -\ln(p_0) \times 2^{r+s}$

Supplementary Figure 1. The workflow of ntCard algorithm for estimating the *k*-mer coverage frequencies and the total number of distinct *k*-mers in DNA sequence streams.



Supplementary Figure 2. The runtime of ntCard algorithm with different number of threads for HG004 dataset and *k*=64.

Supp. Table 1. F_0 and f_1 of different algorithms for HG004.

k		**DSK**	**ntCard**	**err%**	**KmerGenie**	**err%**	**KmerStream**	**err%**	**Khmer**	**err%**
32	f_1	13,319,957,567	13,319,064,830	0.01	13,449,549,282	0.97	12,382,028,923	7.04	-	-
	F_0	16,539,753,749	16,536,020,926	0.02	16,645,716,258	0.64	15,692,266,031	5.12	16,650,898,471	0.67
64	f_1	17,898,672,342	17,902,301,273	0.02	17,961,523,300	0.35	17,767,155,271	0.73	-	-
	F_0	21,343,659,785	21,343,785,902	0.00	21,391,489,196	0.22	21,203,285,348	0.66	21,375,310,744	0.15
96	f_1	18,827,062,018	18,759,470,914	0.36	18,990,183,057	0.87	18,827,278,899	0.00	-	-
	F_0	22,313,944,415	22,260,048,925	0.24	22,467,161,829	0.69	22,324,811,221	0.05	22,244,672,164	0.31
128	f_1	18,091,241,186	18,026,093,555	0.36	18,228,874,975	0.76	18,018,922,859	0.40	-	-
	F_0	21,555,678,676	21,501,437,324	0.25	21,688,723,775	0.62	21,513,520,195	0.20	21,620,187,510	0.30

Supp. Table 2. F_0 and f_1 of different algorithms for NA19238.

k		**DSK**	**ntCard**	**err%**	**KmerGenie**	**err%**	**KmerStream**	**err%**	**Khmer**	**err%**
32	f_1	14,881,561,565	14,881,680,570	0.00	14,960,406,900	0.53	13,934,516,684	6.36	-	-
	F_0	18,091,801,391	18,091,827,603	0.00	18,163,642,050	0.40	17,251,954,617	4.64	18,421,615,787	1.82
64	f_1	19,074,667,480	19,078,850,494	0.02	19,217,432,098	0.75	18,945,178,784	0.68	-	-
	F_0	22,527,419,136	22,530,412,581	0.01	22,700,537,032	0.77	22,381,144,183	0.65	22,802,557,178	1.22
96	f_1	19,420,503,673	19,376,931,559	0.22	19,548,508,320	0.66	19,437,405,253	0.09	-	-
	F_0	22,932,238,161	22,896,217,215	0.16	23,082,721,560	0.66	22,915,649,020	0.07	23,038,122,134	0.46
128	f_1	17,902,027,438	17,864,030,547	0.21	18,053,843,452	0.85	17,935,913,335	0.19	-	-
	F_0	21,421,517,759	21,393,328,674	0.13	21,583,356,391	0.76	21,414,422,860	0.03	21,646,674,001	1.05

Supp. Table 3. F_0 and f_1 of different algorithms for PG29.

k		**DSK**	**ntCard**	**err%**	**KmerGenie**	**err%**	**KmerStream**	**err%**	**Khmer**	**err%**
32	f_1	27,430,910,938	27,426,448,310	0.02	31,637,221,856	15.33	24,850,928,409	9.41	-	-
	F_0	42,642,198,777	42,637,044,318	0.01	47,339,924,096	11.02	39,500,595,213	7.37	46,420,237,663	8.86
64	f_1	44,344,130,469	44,359,962,143	0.04	51,598,993,212	16.36	43,186,071,050	2.61	-	-
	F_0	67,800,291,613	67,816,101,981	0.02	75,351,916,016	11.14	66,626,955,212	1.73	75,382,499,931	11.18
96	f_1	43,300,244,443	43,015,754,808	0.66	50,880,746,585	17.51	42,983,262,871	0.73	-	-
	F_0	69,855,690,006	69,535,745,007	0.46	77,627,984,210	11.13	69,455,435,699	0.57	76,394,432,529	9.36
128	f_1	32,089,613,024	31,961,397,892	0.40	36,846,185,246	14.82	32,069,861,651	0.06	-	-
	F_0	58,195,246,941	58,022,167,722	0.30	63,055,303,509	8.35	58,038,850,303	0.27	62,498,531,497	7.39

Supp. Table 4. The *k*-mer frequencies of DSK, ntCard, and KmerGenie for *k*=32 for HG004.

f	**DSK**	**ntCard**	**Error%**	**KmerGenie**	**Error%**
1	13,319,957,567	13,319,064,830	0.01	13,449,549,282	0.97
2	467,744,995	465,412,740	0.50	469,104,456	0.29
3	73,943,396	74,147,705	0.28	70,899,240	4.12
4	29,741,793	29,885,180	0.48	29,748,444	0.02
5	16,256,725	16,291,021	0.21	16,299,516	0.26
6	10,375,102	10,463,788	0.85	9,684,690	6.65
7	7,311,138	7,294,592	0.23	7,491,930	2.47
8	5,577,512	5,602,832	0.45	5,299,170	4.99
9	4,528,486	4,494,946	0.74	4,056,606	10.42
10	3,900,724	3,783,827	3.00	3,873,876	0.69
11	3,554,869	3,619,298	1.81	3,033,318	14.67
12	3,391,536	3,420,913	0.87	3,252,594	4.10
13	3,388,829	3,486,498	2.88	4,129,698	21.86
14	3,524,600	3,602,339	2.21	3,508,416	0.46
15	3,786,332	3,687,941	2.60	3,325,686	12.17
16	4,173,083	4,124,531	1.16	3,837,330	8.05
17	4,653,875	4,809,006	3.33	4,166,244	10.48
18	5,258,640	5,314,146	1.06	5,956,998	13.28
19	5,954,594	5,955,551	0.02	6,541,734	9.86
20	6,719,630	6,734,972	0.23	6,505,188	3.19
21	7,526,843	7,583,123	0.75	7,053,378	6.29
22	8,349,288	8,476,467	1.52	7,272,654	12.89
23	9,174,841	9,094,423	0.88	9,392,322	2.37
24	9,954,313	9,991,480	0.37	10,488,702	5.37
25	10,672,410	10,848,587	1.65	11,438,898	7.18
26	11,327,187	11,163,564	1.44	10,744,524	5.14
27	11,904,008	11,884,563	0.16	13,266,198	11.44
28	12,377,312	12,336,112	0.33	13,375,836	8.07
29	12,803,476	12,730,166	0.57	12,352,548	3.52
30	13,189,709	13,249,488	0.45	12,498,732	5.24
31	13,576,611	13,681,116	0.77	12,316,002	9.29
32	13,984,009	14,019,853	0.26	13,120,014	6.18
33	14,515,075	14,392,451	0.84	15,093,498	3.98
34	15,187,602	15,214,149	0.17	15,897,510	4.67
35	16,049,070	16,090,314	0.26	15,751,326	1.86
36	17,191,798	17,368,703	1.03	18,053,724	5.01
37	18,594,400	18,664,946	0.38	18,784,644	1.02
38	20,346,073	20,405,263	0.29	19,698,294	3.18
39	22,451,820	22,334,024	0.52	22,402,698	0.22
40	24,930,053	24,912,887	0.07	23,206,710	6.91
41	27,788,584	27,590,644	0.71	27,336,408	1.63
42	31,043,301	31,237,138	0.62	28,798,248	7.23
43	34,643,415	34,467,024	0.51	35,778,534	3.28
44	38,542,699	38,703,775	0.42	38,848,398	0.79
45	42,743,785	42,473,213	0.63	43,745,562	2.34
46	47,106,429	47,048,948	0.12	45,755,592	2.87
47	51,651,062	51,243,087	0.79	49,885,290	3.42
48	56,215,439	56,442,476	0.40	57,742,680	2.72
49	60,709,080	61,022,233	0.52	62,822,574	3.48
50	65,085,342	65,234,680	0.23	63,041,850	3.14
51	69,158,889	69,237,094	0.11	66,769,542	3.45
52	72,909,204	72,811,430	0.13	74,590,386	2.31
53	76,249,148	77,282,685	1.36	77,440,974	1.56
54	79,031,938	79,060,515	0.04	78,756,630	0.35
55	81,237,214	81,228,764	0.01	80,328,108	1.12
56	82,745,999	82,655,892	0.11	84,640,536	2.29
57	83,612,261	83,434,547	0.21	83,215,242	0.47
58	83,751,485	83,245,025	0.60	82,118,862	1.95
59	83,159,954	83,126,450	0.04	78,975,906	5.03
60	81,934,177	82,303,951	0.45	81,753,402	0.22
61	80,073,943	80,269,408	0.24	78,683,538	1.74
62	77,588,136	77,424,998	0.21	75,065,484	3.25
63	74,560,903	74,124,434	0.59	75,065,484	0.68
		AVG	0.71	AVG	4.67
		MAX	3.33	MAX	21.86
		STDEV	0.75	STDEV	4.25

Supp. Table 5. The *k*-mer frequencies of DSK, ntCard, and KmerGenie for *k*=64 for HG004.

f	DSK	ntCard	Error%	KmerGenie	Error%
1	17,898,672,342	17,902,301,273	0.02	17,961,523,300	0.35
2	425,210,274	423,017,210	0.52	418,128,760	1.67
3	45,477,794	45,746,222	0.59	46,452,686	2.14
4	19,144,838	18,937,917	1.08	17,958,794	6.20
5	11,958,180	11,903,262	0.46	11,899,748	0.49
6	9,121,360	9,155,191	0.37	8,242,486	9.64
7	7,846,906	8,055,976	2.66	8,788,346	12.00
8	7,333,695	7,247,456	1.18	7,642,040	4.20
9	7,284,623	7,186,888	1.34	6,604,906	9.33
10	7,597,916	7,443,433	2.03	8,078,728	6.33
11	8,235,423	8,418,635	2.22	7,205,352	12.51
12	9,158,205	9,058,918	1.08	9,497,964	3.71
13	10,384,899	10,144,290	2.32	9,825,480	5.39
14	11,867,027	12,119,538	2.13	12,118,092	2.12
15	13,600,342	13,766,981	1.23	14,574,462	7.16
16	15,523,649	15,589,840	0.43	15,284,080	1.54
17	17,487,930	17,585,082	0.56	16,867,074	3.55
18	19,502,937	19,850,141	1.78	18,777,584	3.72
19	21,402,313	21,507,263	0.49	20,360,578	4.87
20	23,143,255	22,973,404	0.73	22,980,706	0.70
21	24,631,535	24,695,847	0.26	25,109,560	1.94
22	25,844,323	25,720,200	0.48	23,854,082	7.70
23	26,691,911	26,745,395	0.20	26,965,484	1.02
24	27,310,634	27,106,050	0.75	25,600,834	6.26
25	27,771,164	27,795,475	0.09	26,637,968	4.08
26	28,133,388	28,083,788	0.18	28,875,994	2.64
27	28,536,498	28,854,172	1.11	26,801,726	6.08
28	29,134,718	28,756,361	1.30	28,111,790	3.51
29	30,061,533	30,147,800	0.29	29,258,096	2.67
30	31,403,533	31,267,409	0.43	31,168,606	0.75
31	33,290,402	33,395,082	0.31	34,061,664	2.32
32	35,800,114	35,929,984	0.36	36,518,034	2.01
33	38,939,624	38,708,653	0.59	39,738,608	2.05
34	42,727,141	42,798,669	0.17	41,212,430	3.55
35	47,162,523	47,113,651	0.10	49,072,814	4.05
36	52,090,544	51,963,573	0.24	50,601,222	2.86
37	57,451,529	57,294,360	0.27	57,151,542	0.52
38	63,135,037	62,801,854	0.53	63,374,346	0.38
39	68,967,803	68,581,102	0.56	70,907,214	2.81
40	74,824,311	75,333,697	0.68	75,492,438	0.89
41	80,454,765	80,009,672	0.55	78,167,152	2.84
42	85,662,236	85,757,383	0.11	85,536,262	0.15
43	90,292,834	89,775,629	0.57	87,774,288	2.79
44	94,229,665	94,029,842	0.21	95,907,602	1.78
45	97,307,760	97,248,143	0.06	95,634,672	1.72
46	99,425,946	99,245,980	0.18	95,197,984	4.25
47	100,486,631	100,610,511	0.12	99,401,106	1.08
48	100,487,959	100,516,183	0.03	101,584,546	1.09
49	99,453,087	99,857,057	0.41	100,165,310	0.72
50	97,414,816	97,512,598	0.10	97,654,354	0.25
51	94,420,033	94,608,349	0.20	91,486,136	3.11
52	90,524,915	89,962,073	0.62	88,210,976	2.56
53	85,901,801	85,560,132	0.40	88,702,250	3.26
54	80,768,953	81,226,974	0.57	82,370,274	1.98
55	75,198,438	74,912,463	0.38	75,929,126	0.97
56	69,309,095	69,741,037	0.62	72,271,864	4.27
57	63,249,041	63,141,770	0.17	64,793,582	2.44
58	57,205,752	57,283,831	0.14	57,042,370	0.29
59	51,271,898	51,236,860	0.07	53,385,108	4.12
60	45,558,129	45,107,723	0.99	44,542,176	2.23
61	40,116,358	40,337,323	0.55	40,884,914	1.92
62	35,019,734	34,782,012	0.68	34,880,454	0.40
63	30,323,191	30,735,262	1.36	29,913,128	1.35
		AVG	0.65	AVG	3.19
		MAX	2.66	MAX	12.51
		STDEV	0.61	STDEV	2.71

Supp. Table 6. The *k*-mer frequencies of DSK, ntCard, and KmerGenie for *k*=96 for HG004.

f	DSK	ntCard	Error%	KmerGenie	Error%
1	18,827,062,018	18,759,470,914	0.36	18,990,183,057	0.87
2	364,383,376	380,907,398	4.53	366,562,350	0.60
3	38,522,680	39,379,048	2.22	38,719,140	0.51
4	19,759,875	19,677,230	0.42	20,057,784	1.51
5	14,876,194	14,741,074	0.91	15,106,812	1.55
6	13,321,524	13,146,118	1.32	14,567,283	9.35
7	13,168,741	12,957,215	1.61	14,725,968	11.83
8	13,945,693	13,967,026	0.15	14,186,439	1.73
9	15,481,368	15,470,269	0.07	14,948,127	3.44
10	17,677,466	17,561,916	0.65	16,725,399	5.39
11	20,499,684	20,305,119	0.95	19,581,729	4.48
12	23,827,848	23,741,318	0.36	23,802,750	0.11
13	27,490,572	27,264,398	0.82	28,372,878	3.21
14	31,203,519	31,150,309	0.17	30,562,731	2.05
15	34,726,379	34,703,961	0.06	34,339,434	1.11
16	37,858,509	37,662,276	0.52	36,878,394	2.59
17	40,347,051	40,206,576	0.35	40,559,886	0.53
18	42,196,442	41,916,239	0.66	42,114,999	0.19
19	43,413,257	43,456,106	0.10	42,718,002	1.60
20	44,133,270	44,169,683	0.08	43,352,742	1.77
21	44,639,674	44,465,290	0.39	45,161,751	1.17
22	45,173,198	45,292,639	0.26	43,797,060	3.05
23	46,026,846	46,163,795	0.30	45,955,176	0.16
24	47,484,551	47,345,448	0.29	45,955,176	3.22
25	49,790,722	50,441,108	1.31	50,842,674	2.11
26	53,082,441	52,917,255	0.31	52,048,680	1.95
27	57,347,549	57,207,094	0.24	57,126,600	0.39
28	62,580,703	62,227,109	0.57	59,506,875	4.91
29	68,712,341	68,768,798	0.08	70,392,666	2.45
30	75,439,537	75,160,042	0.37	72,931,626	3.32
31	82,547,612	82,707,528	0.19	82,071,882	0.58
32	89,665,503	89,682,216	0.02	91,434,297	1.97
33	96,403,135	95,906,163	0.52	96,448,743	0.05
34	102,464,581	102,793,524	0.32	105,811,158	3.27
35	107,605,600	107,896,266	0.27	105,588,999	1.87
36	111,537,592	111,434,283	0.09	110,857,341	0.61
37	114,010,516	113,556,158	0.40	111,777,714	1.96
38	115,008,537	115,168,374	0.14	113,523,249	1.29
39	114,372,770	114,509,176	0.12	114,824,466	0.39
40	112,278,189	112,006,610	0.24	110,666,919	1.44
41	108,653,683	108,514,719	0.13	106,636,320	1.86
42	103,746,398	102,991,163	0.73	107,302,797	3.43
43	97,758,925	97,520,670	0.24	98,575,122	0.83
44	90,916,492	91,331,973	0.46	89,117,496	1.98
45	83,491,510	83,301,003	0.23	82,484,463	1.21
46	75,725,271	75,671,548	0.07	74,804,109	1.22
47	67,835,317	67,547,180	0.42	68,361,498	0.78
48	60,058,097	60,326,170	0.45	61,379,358	2.20
49	52,534,126	52,208,347	0.62	54,460,692	3.67
50	45,487,185	45,501,591	0.03	47,065,971	3.47
51	38,930,851	38,500,346	1.11	38,211,348	1.85
52	32,961,136	32,800,912	0.49	31,578,315	4.20
53	27,639,650	27,774,175	0.49	27,611,190	0.10
54	22,921,533	23,281,287	1.57	21,993,741	4.05
55	18,849,865	19,074,113	1.19	18,121,827	3.86
56	15,362,405	15,347,634	0.10	15,265,497	0.63
57	12,433,625	12,097,706	2.70	11,647,479	6.32
58	9,993,725	9,932,056	0.62	9,489,363	5.05
59	7,984,024	8,025,887	0.52	7,267,773	8.97
60	6,369,514	6,419,329	0.78	7,077,351	11.11
61	5,057,240	5,070,371	0.26	5,173,131	2.29
62	4,019,929	3,812,763	5.15	4,094,073	1.84
63	3,207,122	3,244,020	1.15	2,951,541	7.97
		AVG	0.67	AVG	2.69
		MAX	5.15	MAX	11.83
		STDEV	0.92	STDEV	2.57

Supp. Table 7. The *k*-mer frequencies of DSK, ntCard, and KmerGenie for *k*=128 for HG004.

f	DSK	ntCard	Error%	KmerGenie	Error%
1	18,091,241,186	18,026,093,555	0.36	18,228,874,975	0.76
2	286,591,196	300,475,705	4.84	286,456,325	0.05
3	38,287,043	38,004,181	0.74	39,383,700	2.86
4	25,321,209	25,348,406	0.11	26,718,050	5.52
5	23,084,968	22,952,712	0.57	25,192,625	9.13
6	24,295,945	23,890,280	1.67	27,041,625	11.30
7	27,512,908	27,433,445	0.29	28,428,375	3.33
8	32,339,046	32,021,521	0.98	32,496,175	0.49
9	38,278,476	38,310,428	0.08	37,950,725	0.86
10	44,842,419	44,381,312	1.03	44,976,925	0.30
11	51,209,265	51,241,319	0.06	53,805,900	5.07
12	56,903,226	56,821,766	0.14	56,902,975	0.00
13	61,337,926	61,013,338	0.53	60,878,325	0.75
14	64,516,172	64,194,305	0.50	63,328,250	1.84
15	66,576,770	66,078,399	0.75	65,500,825	1.62
16	67,955,844	67,699,264	0.38	67,164,925	1.16
17	69,254,289	68,852,350	0.58	71,417,625	3.12
18	71,112,303	70,571,406	0.76	69,245,050	2.63
19	73,996,495	74,244,115	0.33	74,745,825	1.01
20	78,307,378	78,493,159	0.24	75,716,550	3.31
21	84,072,159	83,546,762	0.62	84,268,175	0.23
22	91,249,272	91,632,330	0.42	88,798,225	2.69
23	99,258,260	99,789,513	0.54	100,770,500	1.52
24	107,661,409	107,328,674	0.31	109,460,800	1.67
25	115,775,221	115,776,962	0.00	115,423,825	0.30
26	123,010,263	122,813,398	0.16	126,471,600	2.81
27	128,703,377	129,343,339	0.50	127,627,225	0.84
28	132,397,532	132,851,710	0.34	135,670,375	2.47
29	133,785,647	133,545,333	0.18	132,342,175	1.08
30	132,717,572	133,133,364	0.31	130,262,050	1.85
31	129,296,123	129,219,388	0.06	129,337,550	0.03
32	123,666,497	123,208,303	0.37	120,369,900	2.67
33	116,190,833	116,180,879	0.01	115,238,925	0.82
34	107,240,930	107,317,833	0.07	109,876,825	2.46
35	97,269,590	97,037,747	0.24	96,055,550	1.25
36	86,763,354	86,637,489	0.15	86,440,750	0.37
37	76,124,708	75,705,337	0.55	75,023,175	1.45
38	65,719,489	65,206,114	0.78	64,391,425	2.02
39	55,867,931	56,475,697	1.09	52,095,575	6.75
40	46,833,006	47,146,479	0.67	46,733,475	0.21
41	38,675,198	38,560,855	0.30	37,950,725	1.87
42	31,528,100	31,421,169	0.34	29,907,575	5.14
43	25,365,737	25,138,120	0.90	26,209,575	3.33
44	20,177,895	20,045,382	0.66	21,078,600	4.46
45	15,875,304	15,734,925	0.88	14,653,325	7.70
46	12,371,042	12,495,347	1.00	12,896,775	4.25
47	9,573,864	9,599,577	0.27	8,782,750	8.26
48	7,366,628	7,445,455	1.07	6,979,975	5.25
49	5,655,990	5,595,614	1.07	7,026,200	24.23
50	4,347,667	4,492,919	3.34	4,345,150	0.06
51	3,360,066	3,375,011	0.44	3,050,850	9.20
52	2,625,573	2,554,379	2.71	2,357,475	10.21
53	2,083,598	2,116,757	1.59	2,126,350	2.05
54	1,681,123	1,750,398	4.12	2,033,900	20.98
55	1,395,577	1,429,632	2.44	1,248,075	10.57
56	1,191,564	1,171,118	1.72	1,155,625	3.02
57	1,040,954	1,076,580	3.42	832,050	20.07
58	933,777	896,995	3.94	878,275	5.94
59	851,983	928,788	9.01	647,150	24.04
60	792,057	859,154	8.47	508,475	35.80
61	746,048	776,320	4.06	739,600	0.86
62	706,560	684,781	3.08	832,050	17.76
63	674,729	672,624	0.31	647,150	4.09
		AVG	1.23	AVG	5.04
		MAX	9.01	MAX	35.80
		STDEV	1.78	STDEV	6.93

Supp. Table 8. The *k*-mer frequencies of DSK, ntCard, and KmerGenie for *k*=32 for NA19238.

f	DSK	ntCard	Error%	KmerGenie	Error%
1	14,881,561,570	14,881,680,570	0.00	14,960,406,900	0.53
2	405,294,833	405,870,945	0.14	408,437,250	0.78
3	87,649,181	88,094,633	0.51	87,238,650	0.47
4	36,397,950	36,302,248	0.26	37,624,800	3.37
5	19,625,684	19,430,042	1.00	19,776,150	0.77
6	12,250,635	12,445,638	1.59	12,181,800	0.56
7	8,467,205	8,457,218	0.12	9,444,750	11.55
8	6,332,362	6,326,466	0.09	5,898,150	6.86
9	5,095,252	5,157,660	1.22	5,050,050	0.89
10	4,367,422	4,268,113	2.27	3,585,150	17.91
11	3,968,846	4,005,773	0.93	4,086,300	2.96
12	3,816,920	3,730,057	2.28	3,893,550	2.01
13	3,836,458	3,819,939	0.43	3,816,450	0.52
14	4,034,019	4,127,634	2.32	3,970,650	1.57
15	4,341,913	4,278,647	1.46	4,510,350	3.88
16	4,770,550	4,673,142	2.04	4,510,350	5.45
17	5,307,180	5,197,594	2.06	5,859,600	10.41
18	5,947,594	5,971,017	0.39	5,898,150	0.83
19	6,680,507	6,815,319	2.02	6,630,600	0.75
20	7,467,442	7,552,547	1.14	8,288,250	10.99
21	8,300,384	8,291,009	0.11	8,018,400	3.40
22	9,169,621	9,197,086	0.30	9,830,250	7.20
23	10,087,635	10,114,223	0.26	9,444,750	6.37
24	11,004,000	10,991,278	0.12	11,989,050	8.95
25	11,939,170	11,916,737	0.19	12,104,700	1.39
26	12,867,280	12,916,637	0.38	12,798,600	0.53
27	13,791,966	13,959,434	1.21	13,261,200	3.85
28	14,693,209	14,920,920	1.55	13,608,150	7.38
29	15,594,017	15,585,832	0.05	16,229,550	4.08
30	16,488,472	16,423,628	0.39	17,308,950	4.98
31	17,384,080	17,682,208	1.71	17,501,700	0.68
32	18,278,596	18,057,692	1.21	18,619,650	1.87
33	19,197,614	19,224,634	0.14	18,928,050	1.40
34	20,161,215	20,025,782	0.67	18,696,750	7.26
35	21,196,987	21,264,560	0.32	21,626,550	2.03
36	22,278,480	22,285,582	0.03	20,971,200	5.87
37	23,476,956	23,459,698	0.07	22,783,050	2.96
38	24,743,059	24,868,625	0.51	24,633,450	0.44
39	26,133,764	25,889,915	0.93	23,708,250	9.28
40	27,615,937	27,508,061	0.39	26,445,300	4.24
41	29,202,727	28,699,007	1.72	29,991,900	2.70
42	30,922,553	30,969,643	0.15	32,497,650	5.09
43	32,742,127	32,901,019	0.49	33,153,000	1.25
44	34,670,276	34,882,053	0.61	34,386,600	0.82
45	36,703,076	36,436,589	0.73	38,010,300	3.56
46	38,780,964	38,755,184	0.07	39,937,800	2.98
47	40,945,757	40,472,483	1.16	40,708,800	0.58
48	43,126,601	43,795,024	1.55	42,597,750	1.23
49	45,302,861	45,340,910	0.08	46,915,350	3.56
50	47,501,101	47,866,690	0.77	49,806,600	4.85
51	49,620,568	49,664,717	0.09	47,223,750	4.83
52	51,658,523	51,501,630	0.30	52,890,600	2.39
53	53,607,218	53,605,008	0.00	55,087,950	2.76
54	55,498,575	55,934,352	0.79	54,933,750	1.02
55	57,237,515	57,218,316	0.03	56,552,850	1.20
56	58,763,529	58,715,871	0.08	56,784,150	3.37
57	60,091,994	59,777,359	0.52	60,831,900	1.23
58	61,252,888	61,345,356	0.15	61,872,750	1.01
59	62,163,354	61,928,962	0.38	63,453,300	2.08
60	62,820,706	62,744,747	0.12	62,952,150	0.21
61	63,233,386	63,176,714	0.09	61,448,700	2.82
62	63,351,164	62,998,453	0.56	63,993,000	1.01
63	63,257,370	63,300,660	0.07	61,487,250	2.80
		AVG	0.69	AVG	3.50
		MAX	2.32	MAX	17.91
		STDEV	0.68	STDEV	3.36

Supp. Table 9. The *k*-mer frequencies of DSK, ntCard, and KmerGenie for *k*=64 for NA19238.

f	**DSK**	**ntCard**	**Error%**	**KmerGenie**	**Error%**
1	19,074,667,480	19,078,850,494	0.02	19,217,432,098	0.75
2	342,561,366	342,437,615	0.04	351,669,549	2.66
3	64,952,380	64,784,528	0.26	66,739,131	2.75
4	27,321,610	27,557,094	0.86	26,617,078	2.58
5	15,891,476	16,102,311	1.33	16,451,515	3.52
6	11,319,361	11,286,869	0.29	10,656,653	5.85
7	9,305,101	9,117,019	2.02	8,790,511	5.53
8	8,510,212	8,397,315	1.33	8,201,203	3.63
9	8,416,781	8,107,858	3.67	7,955,658	5.48
10	8,809,969	9,115,518	3.47	8,741,402	0.78
11	9,584,737	9,488,282	1.01	9,723,582	1.45
12	10,708,661	10,848,282	1.30	10,361,999	3.24
13	12,101,678	12,004,761	0.80	13,062,994	7.94
14	13,746,102	13,658,048	0.64	13,013,885	5.33
15	15,601,843	15,898,677	1.90	15,813,098	1.35
16	17,630,231	17,700,620	0.40	18,072,112	2.51
17	19,748,954	19,543,540	1.04	19,005,183	3.77
18	21,896,697	22,113,776	0.99	23,523,211	7.43
19	24,048,896	24,222,050	0.72	24,603,609	2.31
20	26,149,656	25,835,371	1.20	26,469,751	1.22
21	28,133,218	28,274,756	0.50	29,514,509	4.91
22	29,991,329	29,897,535	0.31	31,085,997	3.65
23	31,662,621	31,565,061	0.31	30,447,580	3.84
24	33,134,302	33,196,630	0.19	32,706,594	1.29
25	34,484,948	34,324,225	0.47	35,653,134	3.39
26	35,683,829	35,865,975	0.51	35,211,153	1.32
27	36,708,886	36,479,197	0.63	35,554,916	3.14
28	37,712,821	37,917,686	0.54	36,340,660	3.64
29	38,667,537	38,349,917	0.82	39,385,418	1.86
30	39,622,538	39,921,286	0.75	40,122,053	1.26
31	40,681,340	40,469,864	0.52	40,220,271	1.13
32	41,815,839	42,064,382	0.59	40,956,906	2.05
33	43,162,870	42,982,999	0.42	44,247,209	2.51
34	44,672,088	44,580,734	0.20	44,296,318	0.84
35	46,312,643	46,521,714	0.45	46,506,223	0.42
36	48,193,234	47,930,629	0.54	45,131,171	6.35
37	50,287,987	50,339,566	0.10	53,332,374	6.05
38	52,548,661	52,451,177	0.19	51,220,687	2.53
39	54,941,358	55,203,790	0.48	59,176,345	7.71
40	57,366,017	57,323,028	0.07	61,484,468	7.18
41	59,821,745	59,856,382	0.06	59,470,999	0.59
42	62,277,598	61,915,307	0.58	62,908,629	1.01
43	64,664,115	64,373,108	0.45	68,457,946	5.87
44	66,957,853	66,616,430	0.51	70,569,633	5.39
45	69,053,986	69,225,031	0.25	71,355,377	3.33
46	70,927,509	71,267,199	0.48	68,163,292	3.90
47	72,532,991	72,512,334	0.03	71,257,159	1.76
48	73,832,022	73,836,744	0.01	71,699,140	2.89
49	74,723,642	74,730,511	0.01	75,038,552	0.42
50	75,273,548	74,965,966	0.41	78,967,272	4.91
51	75,412,773	75,435,276	0.03	79,703,907	5.69
52	75,196,502	75,215,547	0.03	76,904,694	2.27
53	74,501,383	74,446,137	0.07	74,056,372	0.60
54	73,426,681	73,035,826	0.53	74,105,481	0.92
55	71,936,960	71,867,550	0.10	71,748,249	0.26
56	70,033,372	69,749,760	0.40	70,422,306	0.56
57	67,744,755	68,170,403	0.63	67,132,003	0.90
58	65,127,802	64,932,460	0.30	66,591,804	2.25
59	62,253,568	61,611,400	1.03	62,024,667	0.37
60	59,102,284	59,479,334	0.64	58,979,909	0.21
61	55,694,408	55,633,440	0.11	54,560,099	2.04
62	52,137,686	52,441,944	0.58	52,104,649	0.06
63	48,488,414	48,672,104	0.38	60,660,444	25.10
		AVG	0.63	AVG	3.28
		MAX	3.67	MAX	25.10
		STDEV	0.68	STDEV	3.48

Supp. Table 10. The *k*-mer frequencies of DSK, ntCard, and KmerGenie for *k*=96 for NA19238.

f	**DSK**	**ntCard**	**Error%**	**KmerGenie**	**Error%**
1	19,420,503,674	19,376,931,559	0.22	19,548,508,320	0.66
2	282,308,899	291,190,700	3.15	283,863,150	0.55
3	54,459,477	55,909,399	2.66	56,103,930	3.02
4	25,663,077	26,128,584	1.81	26,647,695	3.84
5	17,553,513	17,430,907	0.70	17,586,810	0.19
6	14,990,090	15,102,395	0.75	16,182,540	7.95
7	14,742,442	14,722,612	0.13	13,507,740	8.38
8	15,778,761	15,879,006	0.64	17,519,940	11.03
9	17,739,166	17,859,196	0.68	18,288,945	3.10
10	20,382,977	20,331,193	0.25	21,130,920	3.67
11	23,601,055	23,269,814	1.40	22,602,060	4.23
12	27,227,221	27,052,211	0.64	29,088,450	6.84
13	31,098,848	30,740,455	1.15	31,997,295	2.89
14	35,104,518	34,852,985	0.72	33,903,090	3.42
15	39,077,407	38,927,151	0.38	39,453,300	0.96
16	42,812,263	42,877,707	0.15	43,833,285	2.38
17	46,194,642	46,051,650	0.31	48,112,965	4.15
18	49,170,008	49,294,875	0.25	49,450,365	0.57
19	51,617,530	51,453,440	0.32	51,356,160	0.51
20	53,663,081	53,984,578	0.60	51,924,555	3.24
21	55,162,191	55,270,987	0.20	55,000,575	0.29
22	56,355,953	56,166,599	0.34	56,939,805	1.04
23	57,198,414	57,213,778	0.03	57,909,420	1.24
24	57,888,908	57,878,576	0.02	58,176,900	0.50
25	58,565,438	58,763,699	0.34	57,842,550	1.23
26	59,320,768	59,222,567	0.17	61,052,310	2.92
27	60,307,479	60,562,454	0.42	62,389,710	3.45
28	61,550,317	61,492,156	0.09	64,362,375	4.57
29	63,115,317	63,603,518	0.77	63,091,845	0.04
30	64,990,105	65,308,525	0.49	63,392,760	2.46
31	67,182,412	67,073,165	0.16	68,173,965	1.48
32	69,640,028	69,746,797	0.15	69,645,105	0.01
33	72,173,961	72,020,151	0.21	73,623,870	2.01
34	74,785,732	74,820,919	0.05	74,493,180	0.39
35	77,347,867	77,774,237	0.55	74,894,400	3.17
36	79,798,427	80,303,750	0.63	80,244,000	0.56
37	81,960,814	81,494,039	0.57	80,578,350	1.69
38	83,816,640	83,663,541	0.18	84,724,290	1.08
39	85,204,462	85,033,474	0.20	84,690,855	0.60
40	86,047,935	85,996,012	0.06	83,654,370	2.78
41	86,339,203	85,238,744	1.27	85,426,425	1.06
42	86,031,814	85,638,593	0.46	85,092,075	1.09
43	85,099,410	84,665,749	0.51	87,298,785	2.58
44	83,515,753	83,305,372	0.25	83,219,715	0.35
45	81,290,333	80,944,079	0.43	82,818,495	1.88
46	78,503,679	77,483,432	1.30	78,839,730	0.43
47	75,206,358	75,964,425	1.01	75,997,755	1.05
48	71,445,945	71,190,631	0.36	72,620,820	1.64
49	67,291,563	67,802,918	0.76	68,140,530	1.26
50	62,832,014	62,415,670	0.66	64,429,245	2.54
51	58,124,406	58,046,686	0.13	59,179,950	1.82
52	53,348,421	53,208,749	0.26	53,228,520	0.22
53	48,497,096	48,317,970	0.37	48,948,840	0.93
54	43,727,781	43,407,060	0.73	45,939,690	5.06
55	39,060,481	39,309,639	0.64	38,483,685	1.48
56	34,583,992	34,312,154	0.79	34,538,355	0.13
57	30,333,946	30,080,851	0.83	30,559,590	0.74
58	26,381,945	26,188,955	0.73	26,547,390	0.63
59	22,735,166	22,939,750	0.90	23,304,195	2.50
60	19,446,835	19,428,103	0.10	19,225,125	1.14
61	16,490,743	16,617,782	0.77	15,580,710	5.52
62	13,879,203	14,013,435	0.97	14,143,005	1.90
63	11,597,600	11,512,240	0.74	11,468,205	1.12
		AVG	0.60	AVG	2.23
		MAX	3.15	MAX	11.03
		STDEV	0.56	STDEV	2.15

Supp. Table 11. The *k*-mer frequencies of DSK, ntCard, and KmerGenie for *k*=128 for NA19238.

f	**DSK**	**ntCard**	**Error%**	**KmerGenie**	**Error%**
1	17,902,027,438	17,864,030,547	0.21	18,053,843,452	0.85
2	219,074,213	229,655,845	4.83	222,218,225	1.44
3	47,660,610	48,455,773	1.67	48,372,365	1.49
4	28,458,423	28,502,475	0.15	27,746,585	2.50
5	25,220,664	25,322,354	0.40	25,389,353	0.67
6	26,863,655	26,728,874	0.50	26,567,969	1.10
7	30,937,124	30,813,466	0.40	31,380,651	1.43
8	36,537,328	35,997,413	1.48	37,028,186	1.34
9	43,058,213	42,939,508	0.28	39,778,290	7.62
10	50,073,961	49,855,415	0.44	49,796,526	0.55
11	57,131,512	56,978,579	0.27	57,506,639	0.66
12	63,762,620	63,020,538	1.16	65,756,951	3.13
13	69,682,587	69,293,844	0.56	72,877,756	4.59
14	74,487,927	73,953,747	0.72	73,368,846	1.50
15	78,198,344	78,149,055	0.06	78,623,509	0.54
16	80,694,462	80,513,285	0.22	79,311,035	1.71
17	82,295,487	81,950,855	0.42	82,306,684	0.01
18	83,084,771	82,998,538	0.10	83,141,537	0.07
19	83,483,643	83,419,414	0.08	84,663,916	1.41
20	83,711,270	83,910,734	0.24	86,038,968	2.78
21	84,183,095	84,186,381	0.00	86,726,494	3.02
22	84,964,856	84,760,063	0.24	82,110,248	3.36
23	86,212,498	85,643,986	0.66	82,601,338	4.19
24	87,917,383	87,946,740	0.03	88,838,181	1.05
25	90,072,760	89,908,577	0.18	89,673,034	0.44
26	92,482,062	93,188,780	0.76	93,847,299	1.48
27	94,989,315	94,740,939	0.26	98,267,109	3.45
28	97,336,756	97,046,640	0.30	97,776,019	0.45
29	99,316,688	99,806,932	0.49	101,753,848	2.45
30	100,800,801	100,951,466	0.15	98,807,308	1.98
31	101,544,014	101,893,260	0.34	103,374,445	1.80
32	101,409,315	101,377,162	0.03	104,503,952	3.05
33	100,296,926	100,427,207	0.13	97,874,237	2.42
34	98,218,038	97,780,653	0.45	98,168,891	0.05
35	95,144,483	95,570,049	0.45	98,168,891	3.18
36	91,166,502	91,334,818	0.18	92,914,228	1.92
37	86,333,852	85,829,007	0.58	82,552,229	4.38
38	80,829,400	80,991,602	0.20	80,882,523	0.07
39	74,772,191	75,069,482	0.40	73,025,083	2.34
40	68,423,734	68,251,774	0.25	66,788,240	2.39
41	61,852,988	61,735,090	0.19	60,944,269	1.47
42	55,231,615	55,117,117	0.21	52,301,085	5.31
43	48,743,059	49,094,195	0.72	49,550,981	1.66
44	42,506,944	42,885,336	0.89	42,331,958	0.41
45	36,640,880	37,171,884	1.45	38,403,238	4.81
46	31,237,420	31,265,133	0.09	31,675,305	1.40
47	26,342,228	25,882,762	1.74	23,916,083	9.21
48	21,952,651	21,897,792	0.25	22,197,268	1.11
49	18,109,906	18,185,237	0.42	18,464,984	1.96
50	14,800,251	14,803,151	0.02	15,371,117	3.86
51	11,974,386	11,721,069	2.12	13,161,212	9.91
52	9,605,152	9,708,904	1.08	10,018,236	4.30
53	7,671,941	7,777,948	1.38	8,446,748	10.10
54	6,083,920	6,342,034	4.24	5,893,080	3.14
55	4,805,106	4,745,168	1.25	4,960,009	3.22
56	3,793,983	3,876,638	2.18	4,272,483	12.61
57	3,000,182	3,041,987	1.39	2,848,322	5.06
58	2,385,750	2,361,159	1.03	2,553,668	7.04
59	1,919,068	1,847,451	3.73	1,767,924	7.88
60	1,561,783	1,621,498	3.82	2,013,469	28.92
61	1,293,425	1,241,348	4.03	1,129,507	12.67
62	1,093,205	1,102,878	0.88	1,178,616	7.81
63	946,854	987,759	4.32	1,375,052	45.22
		AVG	0.92	AVG	4.25
		MAX	4.83	MAX	45.22
		STDEV	1.18	STDEV	6.80

Supp. Table 12. The *k*-mer frequencies of DSK, ntCard, and KmerGenie for *k*=32 for PG29.

f	DSK	ntCard	Error%	KmerGenie	Error%
1	27,430,910,949	27,426,448,310	0.02	31,637,221,856	15.33
2	1,387,809,246	1,389,451,001	0.12	1,691,533,120	21.89
3	350,390,148	349,714,416	0.19	425,758,688	21.51
4	171,711,802	172,892,861	0.69	203,856,512	18.72
5	116,050,402	116,055,438	0.00	133,656,896	15.17
6	99,156,313	99,268,914	0.11	109,067,200	10.00
7	99,121,826	98,670,654	0.46	99,746,912	0.63
8	109,752,818	109,416,541	0.31	104,307,904	4.96
9	128,437,274	128,480,787	0.03	144,365,312	12.40
10	154,072,012	153,474,253	0.39	161,022,848	4.51
11	185,369,757	185,748,995	0.20	181,249,856	2.22
12	221,102,222	220,672,412	0.19	228,247,904	3.23
13	259,377,565	259,868,740	0.19	258,191,808	0.46
14	298,066,994	299,110,382	0.35	297,059,392	0.34
15	335,024,898	336,029,735	0.30	336,720,192	0.51
16	368,147,603	366,904,043	0.34	377,570,816	2.56
17	395,765,430	395,084,465	0.17	406,523,200	2.72
18	416,598,835	416,666,618	0.02	415,645,184	0.23
19	429,772,109	430,464,573	0.16	418,223,136	2.69
20	435,392,488	435,566,519	0.04	421,990,912	3.08
21	433,825,841	433,488,148	0.08	422,189,216	2.68
22	425,873,062	425,590,688	0.07	416,438,400	2.22
23	412,724,311	412,131,175	0.14	410,290,976	0.59
24	395,426,021	395,825,922	0.10	386,494,496	2.26
25	375,336,933	374,103,194	0.33	385,701,280	2.76
26	353,602,476	354,641,223	0.29	359,525,152	1.67
27	331,198,158	331,619,395	0.13	346,437,088	4.60
28	309,025,096	309,165,243	0.05	313,716,928	1.52
29	287,724,851	289,309,919	0.55	286,747,584	0.34
30	267,671,913	266,967,790	0.26	265,727,360	0.73
31	249,191,152	249,083,627	0.04	268,305,312	7.67
32	232,311,547	231,847,348	0.20	229,041,120	1.41
33	217,260,210	218,177,312	0.42	223,290,304	2.78
34	203,859,537	202,776,544	0.53	209,409,024	2.72
35	192,020,454	192,590,567	0.30	186,009,152	3.13
36	181,556,134	180,358,602	0.66	193,346,400	6.49
37	172,225,210	172,692,530	0.27	173,714,304	0.86
38	163,885,099	163,233,828	0.40	161,617,760	1.38
39	156,398,560	157,010,116	0.39	150,909,344	3.51
40	149,589,130	149,331,444	0.17	138,614,496	7.34
41	143,302,761	144,161,261	0.60	140,795,840	1.75
42	137,445,696	137,216,320	0.17	132,467,072	3.62
43	131,887,301	132,254,754	0.28	131,872,160	0.01
44	126,598,354	127,215,314	0.49	124,336,608	1.79
45	121,458,616	120,810,803	0.53	117,792,576	3.02
46	116,516,109	116,190,475	0.28	112,041,760	3.84
47	111,646,342	111,992,636	0.31	110,455,328	1.07
48	106,943,071	107,136,764	0.18	111,645,152	4.40
49	102,288,148	101,821,828	0.46	107,084,160	4.69
50	97,773,975	97,541,728	0.24	99,746,912	2.02
51	93,349,409	94,018,838	0.72	96,970,656	3.88
52	89,052,632	89,072,634	0.02	92,211,360	3.55
53	84,864,906	84,409,044	0.54	81,899,552	3.49
54	80,830,407	81,257,866	0.53	79,916,512	1.13
55	76,945,977	76,993,685	0.06	79,321,600	3.09
56	73,165,096	73,054,356	0.15	74,165,696	1.37
57	69,534,960	69,693,107	0.23	70,397,920	1.24
58	66,048,290	65,619,336	0.65	66,630,144	0.88
59	62,727,490	62,433,288	0.47	63,655,584	1.48
60	59,582,910	59,270,336	0.52	62,267,456	4.51
61	56,576,498	56,264,043	0.55	51,360,736	9.22
62	53,711,573	53,694,153	0.03	59,292,896	10.39
63	51,013,336	50,971,744	0.08	52,352,256	2.62
		AVG	0.28	AVG	4.33
		MAX	0.72	MAX	21.89
		STDEV	0.20	STDEV	4.94

Supp. Table 13. The *k*-mer frequencies of DSK, ntCard, and KmerGenie for *k*=64 for PG29.

f	**DSK**	**ntCard**	**Error%**	**KmerGenie**	**Error%**
1	44,344,130,473	44,359,962,143	0.04	51,598,993,212	16.36
2	1,058,391,480	1,059,215,469	0.08	1,318,128,552	24.54
3	304,744,499	305,269,528	0.17	345,475,468	13.37
4	244,043,506	243,275,577	0.31	256,412,728	5.07
5	279,144,707	278,882,007	0.09	277,084,080	0.74
6	358,411,249	357,254,650	0.32	358,889,856	0.13
7	467,102,251	466,994,232	0.02	461,806,800	1.13
8	594,972,530	595,825,563	0.14	602,547,920	1.27
9	731,168,050	731,483,154	0.04	734,052,904	0.39
10	863,467,638	861,152,709	0.27	879,412,092	1.85
11	979,792,323	980,174,265	0.04	963,416,948	1.67
12	1,070,499,982	1,069,177,489	0.12	1,056,657,940	1.29
13	1,129,311,769	1,131,498,917	0.19	1,090,303,864	3.45
14	1,154,118,918	1,152,400,543	0.15	1,169,690,652	1.35
15	1,146,170,367	1,149,679,014	0.31	1,129,887,304	1.42
16	1,110,043,613	1,108,539,690	0.14	1,109,435,860	0.05
17	1,052,102,837	1,050,046,277	0.20	1,056,657,940	0.43
18	979,048,432	981,172,640	0.22	989,366,092	1.05
19	897,755,694	898,502,443	0.08	900,083,444	0.26
20	813,492,817	812,910,416	0.07	825,314,724	1.45
21	730,817,879	729,793,025	0.14	760,881,680	4.11
22	652,942,858	654,078,166	0.17	656,645,288	0.57
23	581,744,571	579,790,909	0.34	586,934,452	0.89
24	518,095,968	520,436,044	0.45	531,077,820	2.51
25	462,093,895	463,019,944	0.20	462,026,708	0.01
26	413,349,280	412,625,104	0.18	405,730,260	1.84
27	371,308,278	370,447,365	0.23	363,507,924	2.10
28	335,032,329	333,163,314	0.56	334,260,160	0.23
29	303,713,035	304,182,109	0.15	298,195,248	1.82
30	276,411,534	277,568,559	0.42	275,324,816	0.39
31	252,455,179	252,284,108	0.07	253,334,016	0.35
32	231,301,044	231,137,592	0.07	229,364,044	0.84
33	212,443,283	212,770,554	0.15	209,352,416	1.45
34	195,397,396	195,423,098	0.01	195,718,120	0.16
35	179,984,187	180,556,034	0.32	180,544,468	0.31
36	165,807,712	165,032,329	0.47	164,051,368	1.06
37	152,801,139	154,270,841	0.96	140,741,120	7.89
38	140,846,302	141,185,846	0.24	135,683,236	3.67
39	129,791,135	130,072,366	0.22	130,405,444	0.47
40	119,560,159	118,311,098	1.04	118,310,504	1.05
41	110,088,923	109,755,676	0.30	116,771,148	6.07
42	101,331,018	101,083,938	0.24	104,236,392	2.87
43	93,277,261	93,519,313	0.26	89,282,648	4.28
44	85,862,377	85,909,347	0.05	85,764,120	0.11
45	79,038,067	79,188,297	0.19	81,146,052	2.67
46	72,730,819	73,381,235	0.89	71,250,192	2.04
47	66,958,741	66,557,274	0.60	73,449,272	9.69
48	61,660,909	61,282,475	0.61	57,176,080	7.27
49	56,830,113	57,068,662	0.42	56,736,264	0.17
50	52,402,360	51,842,900	1.07	55,196,908	5.33
51	48,369,898	48,361,867	0.02	47,280,220	2.25
52	44,688,457	45,663,423	2.18	44,421,416	0.60
53	41,309,395	40,354,352	2.31	46,180,680	11.79
54	38,249,355	38,130,872	0.31	33,206,108	13.19
55	35,422,440	35,185,771	0.67	35,625,096	0.57
56	32,856,893	33,604,857	2.28	33,865,832	3.07
57	30,524,398	30,673,793	0.49	29,467,672	3.46
58	28,379,165	28,211,890	0.59	26,169,052	7.79
59	26,420,852	26,996,019	2.18	28,148,224	6.54
60	24,646,117	24,912,913	1.08	28,807,948	16.89
61	23,006,760	23,301,593	1.28	20,451,444	11.11
62	21,535,331	21,187,128	1.62	20,671,352	4.01
63	20,161,903	20,709,470	2.72	23,310,248	15.62
		AVG	0.50	AVG	3.91
		MAX	2.72	MAX	24.54
		STDEV	0.63	STDEV	5.07

Supp. Table 14. The *k*-mer frequencies of DSK, ntCard, and KmerGenie for *k*=96 for PG29.

f	**DSK**	**ntCard**	**Error%**	**KmerGenie**	**Error%**
1	43,300,244,443	43,036,616,081	0.61	50,880,746,585	17.51
2	898,031,149	932,941,144	3.89	1,066,153,205	18.72
3	707,016,398	706,351,268	0.09	732,501,615	3.60
4	965,800,373	961,568,021	0.44	956,679,615	0.94
5	1,291,696,378	1,283,755,561	0.61	1,273,331,040	1.42
6	1,606,040,417	1,597,494,114	0.53	1,582,696,680	1.45
7	1,853,981,273	1,844,250,584	0.52	1,834,149,670	1.07
8	2,001,239,303	1,993,152,433	0.40	1,993,502,865	0.39
9	2,037,849,212	2,029,572,623	0.41	2,027,690,010	0.50
10	1,973,342,921	1,963,162,359	0.52	1,947,546,375	1.31
11	1,832,321,839	1,824,183,587	0.44	1,824,435,290	0.43
12	1,643,575,419	1,636,762,494	0.41	1,651,444,600	0.48
13	1,434,671,059	1,429,904,970	0.33	1,432,684,235	0.14
14	1,226,867,295	1,221,175,808	0.46	1,256,144,060	2.39
15	1,034,419,205	1,032,492,988	0.19	1,042,614,515	0.79
16	864,953,818	862,316,459	0.30	874,107,385	1.06
17	720,753,222	720,427,454	0.05	727,457,610	0.93
18	601,253,972	600,639,316	0.10	606,401,490	0.86
19	503,334,855	502,868,040	0.09	505,334,575	0.40
20	423,827,756	423,347,225	0.11	427,619,535	0.89
21	359,196,771	358,663,860	0.15	354,574,870	1.29
22	306,266,415	305,851,779	0.14	320,761,355	4.73
23	262,638,051	262,319,477	0.12	274,618,050	4.56
24	226,304,457	227,361,944	0.47	225,859,335	0.20
25	195,644,114	195,681,121	0.02	196,342,565	0.36
26	169,676,672	169,231,732	0.26	170,188,465	0.30
27	147,420,488	146,262,619	0.79	155,056,450	5.18
28	128,282,598	129,012,537	0.57	130,396,870	1.65
29	111,768,049	111,630,842	0.12	112,836,260	0.96
30	97,496,043	97,720,828	0.23	101,814,175	4.43
31	85,133,870	85,044,445	0.11	86,495,345	1.60
32	74,443,436	75,299,775	1.15	76,220,520	2.39
33	65,183,093	65,630,422	0.69	61,462,135	5.71
34	57,172,299	57,273,331	0.18	63,330,285	10.77
35	50,254,833	50,453,857	0.40	51,934,570	3.34
36	44,225,682	44,697,157	1.07	44,275,155	0.11
37	39,030,772	39,077,854	0.12	44,088,340	12.96
38	34,551,288	34,871,851	0.93	33,066,255	4.30
39	30,662,154	30,756,149	0.31	27,835,435	9.22
40	27,302,004	27,745,656	1.62	28,769,510	5.38
41	24,397,583	24,325,429	0.30	22,604,615	7.35
42	21,858,560	22,017,184	0.73	22,417,800	2.56
43	19,662,239	19,783,303	0.62	20,549,650	4.51
44	17,738,982	18,031,501	1.65	19,615,575	10.58
45	16,062,223	16,403,515	2.12	17,560,610	9.33
46	14,594,560	14,836,392	1.66	10,835,270	25.76
47	13,301,488	13,651,239	2.63	15,879,275	19.38
48	12,162,628	12,550,440	3.19	14,758,385	21.34
49	11,144,319	11,318,751	1.57	11,022,085	1.10
50	10,254,185	10,362,347	1.05	9,527,565	7.09
51	9,472,392	9,451,002	0.23	9,340,750	1.39
52	8,760,594	8,916,220	1.78	6,538,525	25.36
53	8,136,102	8,332,961	2.42	10,648,455	30.88
54	7,573,126	7,577,645	0.06	8,219,860	8.54
55	7,060,989	7,276,106	3.05	5,604,450	20.63
56	6,608,684	6,730,716	1.85	6,912,155	4.59
57	6,199,642	6,373,294	2.80	6,538,525	5.47
58	5,826,148	5,911,328	1.46	4,483,560	23.04
59	5,492,975	5,711,879	3.99	6,912,155	25.84
60	5,176,400	5,229,073	1.02	5,230,820	1.05
61	4,897,209	5,011,077	2.33	6,164,895	25.89
62	4,639,376	4,729,919	1.95	4,109,930	11.41
63	4,400,250	4,516,981	2.65	4,109,930	6.60
		AVG	0.97	AVG	6.89
		MAX	3.99	MAX	30.88
		STDEV	1.00	STDEV	8.24

Supp. Table 15. The *k*-mer frequencies of DSK, ntCard, and KmerGenie for *k*=128 for PG29.

f	DSK	ntCard	Error%	KmerGenie	Error%
1	32,089,613,024	31,961,397,892	0.40	36,846,185,246	14.82
2	2,906,826,774	2,903,450,063	0.12	2,918,992,670	0.42
3	3,528,887,195	3,515,526,971	0.38	3,469,505,791	1.68
4	3,785,681,675	3,772,834,237	0.34	3,785,793,909	0.00
5	3,572,564,965	3,559,249,791	0.37	3,609,549,948	1.04
6	3,053,210,499	3,050,701,898	0.08	3,070,080,835	0.55
7	2,422,704,144	2,414,305,067	0.35	2,469,409,511	1.93
8	1,823,893,429	1,821,291,960	0.14	1,811,063,923	0.70
9	1,327,251,074	1,326,308,733	0.07	1,348,135,921	1.57
10	949,469,860	950,923,714	0.15	956,073,637	0.70
11	676,800,750	676,056,332	0.11	676,905,657	0.02
12	485,255,917	486,751,332	0.31	504,189,643	3.90
13	351,999,218	352,633,774	0.18	348,653,197	0.95
14	258,705,511	258,079,904	0.24	263,215,524	1.74
15	192,574,291	194,804,917	1.16	196,491,309	2.03
16	144,886,815	145,629,939	0.51	147,713,607	1.95
17	109,994,471	110,472,781	0.43	103,997,742	5.45
18	84,210,354	84,094,204	0.14	84,210,561	0.00
19	64,992,769	65,952,127	1.48	66,724,215	2.66
20	50,552,591	51,680,691	2.23	54,453,095	7.72
21	39,689,534	39,758,767	0.17	38,040,472	4.15
22	31,486,278	31,232,898	0.80	33,592,191	6.69
23	25,275,422	25,937,510	2.62	26,382,908	4.38
24	20,531,059	20,751,793	1.08	23,008,350	12.07
25	16,905,515	17,256,090	2.07	17,639,735	4.34
26	14,099,871	14,243,523	1.02	13,344,843	5.35
27	11,906,560	12,110,219	1.71	12,577,898	5.64
28	10,182,403	10,169,175	0.13	11,350,786	11.47
29	8,821,744	9,192,513	4.20	7,976,228	9.58
30	7,714,351	7,666,099	0.63	6,902,505	10.52
31	6,807,619	6,815,006	0.11	5,215,226	23.39
32	6,057,582	6,130,146	1.20	7,822,839	29.14
33	5,441,514	5,548,774	1.97	6,135,560	12.75
34	4,918,198	5,008,826	1.84	3,221,169	34.51
35	4,472,444	4,624,758	3.41	3,988,114	10.83
36	4,082,734	4,085,271	0.06	4,755,059	16.47
37	3,750,820	3,934,984	4.91	4,755,059	26.77
38	3,461,233	3,587,991	3.66	3,067,780	11.37
39	3,201,009	3,351,164	4.69	3,221,169	0.63
40	2,971,253	3,173,174	6.80	2,300,835	22.56
41	2,768,397	2,875,604	3.87	1,994,057	27.97
42	2,592,248	2,732,397	5.41	2,454,224	5.32
43	2,425,325	2,646,567	9.12	3,834,725	58.11
44	2,278,604	2,244,797	1.48	2,454,224	7.71
45	2,141,789	2,355,443	9.98	2,147,446	0.26
46	2,021,550	2,015,046	0.32	1,380,501	31.71
47	1,907,384	2,011,048	5.43	1,687,279	11.54
48	1,809,373	1,769,676	2.19	1,840,668	1.73
49	1,714,589	1,816,125	5.92	1,840,668	7.35
50	1,626,364	1,647,043	1.27	1,380,501	15.12
51	1,546,740	1,608,060	3.96	2,607,613	68.59
52	1,473,353	1,581,836	7.36	1,687,279	14.52
53	1,401,882	1,378,742	1.65	1,073,723	23.41
54	1,336,585	1,326,448	0.76	920,334	31.14
55	1,278,342	1,353,443	5.87	460,167	64.00
56	1,223,898	1,246,745	1.87	1,840,668	50.39
57	1,170,828	1,209,962	3.34	1,840,668	57.21
58	1,118,823	1,045,996	6.51	920,334	17.74
59	1,075,668	1,107,593	2.97	2,147,446	99.64
60	1,030,717	1,073,627	4.16	1,227,112	19.05
61	990,754	903,223	8.83	920,334	7.11
62	951,823	998,730	4.93	1,687,279	77.27
63	914,881	910,683	0.46	153,389	83.23
		AVG	2.38	AVG	17.34
		MAX	9.98	MAX	99.64
		STDEV	2.54	STDEV	22.58