Supplementary Information for "KmerCo: A lightweight K-mer counting technique with a tiny memory footprint"

Sabuzima Nayak and Ripon Patgiri

1. RESULTS

We have conducted some additional experiments on KmerCo to evaluate its performance based on the different countBF counter lengths in bits and data structure size. These experiments are performed on the largest dataset, i.e., the Balaenoptera dataset.

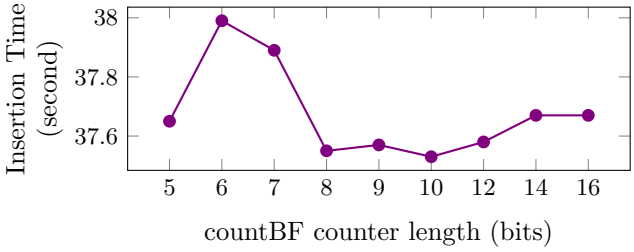
1.1 Varying countBF counter length

Counter length (bits)	#Counter	Wasted bits per cell
5	12	4
6	10	4
7	9	1
8	8	0
9	7	1
10	6	4
12	5	4
14	4	8
16	4	0

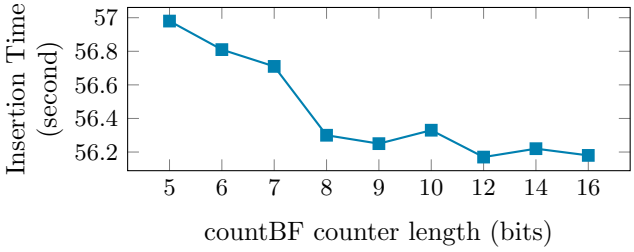
Table 1: **Details of the number of counters and the wasted bits in each cell of countBF in KmerCo. #Counter: Number of counters in each cell.**

In experimentation, we have considered the countBF counter length from 5 up to 16. The counter length of less than 5 is not considered because a 4-bit counter can store a maximum frequency of 15 which is a low value. Hence, the experiment is conducted by considering the counter length of more than 4 bits. Similarly, a counter length of more than 16 bits has less than 4 counters in each cell and more wasted bits. Table 1 highlights the number of counters and bits wasted per cell by varying the counter length of countBF in KmerCo. The ideal counter length is 8 and 16 with zero wasted bits. On the contrary, the counter length 14 has the highest wasted bits (with reference to Table 1). The KmerCo have the same size but different counter lengths within a cell. All KmerCo inserted the same number of K-mers, and the inserted-to-ignored K-mer ratio is zero. Hence, in this section the comparison is presented based on insertion time, number of insertions per second, and trustworthy rate.

sabuzima_rs@cse.nits.ac.in (Sabuzima Nayak), ripon@cse.nits.ac.in (Ripon Patgiri)
Department of Computer Science & Engineering, National Institute of Technology Silchar
Silchar, 788010, Assam, India



(a) 28-mers



(b) 55-mers

time wrt counter length

Figure 1: **Comparison of insertion time in second among various KmerCo having different counter lengths of countBF using (a) 28-mers and (b) 55-mers Balaenoptera dataset. Lower is better.**

Figure 1 illustrates the comparison of KmerCo having varying counter lengths based on the insertion time in seconds using the 28-mer (Figure 1a) and 55-mer (Figure 1b) Balaenoptera dataset. KmerCo having countBF with a counter length of less than 8 bits has more insertion time compared to countBF with a counter length of more than 8 bits. In most of the cases, the insertion time increases with an increase in the counter length of countBF with exception of counter lengths 7, 8 and 10. In these cases, the insertion time decreases from the previous counter length countBF. KmerCo having countBF with a 10-bit counter has the least time followed by countBF with an 8-bit counter. But a countBF with a 10-bit counter has 4 wasted bits per

cell; hence, the countBF with an 8-bit counter is better. In the case of the 55-mer Balaenoptera dataset, the insertion time decreases with an increase in the counter length of countBF. The least time is taken by countBF with a 12-bit counter while the highest time is taken by countBF with a 5-bit counter.

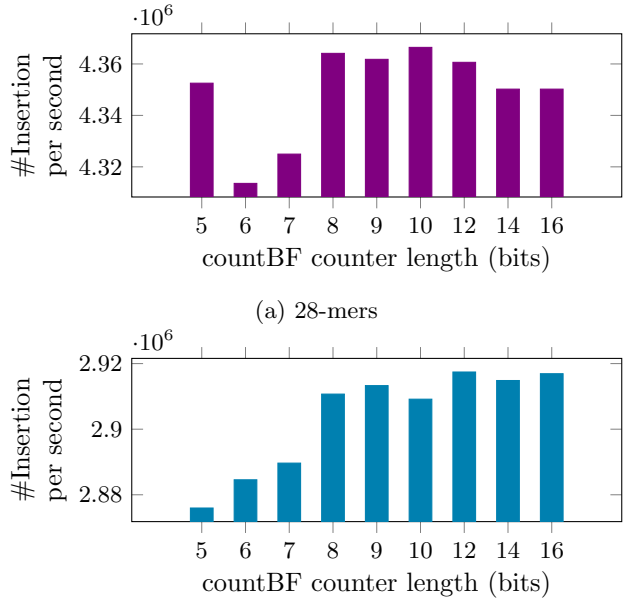


Figure 2: **Comparison of the number of (a) 28-mers and (b) 55-mers inserted per second among various KmerCo having different counter lengths of countBF using Balaenoptera dataset. Higher is better. #Insertion: Number of insertions.**

Figure 2 illuminates the comparison among the KmerCo having different counter lengths based on the number of insertions per second using the 28-mer (Figure 2a) and 55-mer (Figure 2b) Balaenoptera dataset. The countBF with counter lengths less than 8 bits gives an unusual pattern as they have more counters but less number of insertions per second compared to countBF with counter lengths more than 8 bits. The countBF with a 10-bit counter exhibits the highest number of insertions per second followed by countBF with an 8-bit counter. Therefore, countBF with an 8-bit counter is better because a countBF with a 10-bit counter has 4 wasted bits per cell. In the case of 55-mer Balaenoptera dataset, the number of insertions per second increases with an increase in the counter length of countBF. The countBF with a 12-bit counter has the highest number of insertions per second. The number of insertions per second of countBF with an 8-bit counter is close to countBF with a 12-bit counter. Moreover, both countBF with 10 bits and 12 bits counters have 4 wasted bits per cell; hence countBF with an 8-bit counter is better.

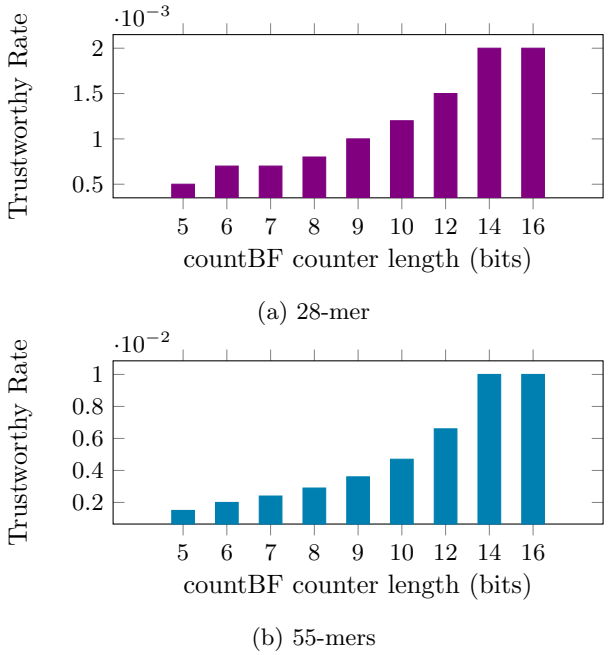


Figure 3: **Comparison of the trustworthy rate among various KmerCo having different counter lengths of countBF using 55-mer Balaenoptera dataset. Positive and close to zero is better.**

Figure 3 highlights the comparison among various KmerCo having different counter lengths based on trustworthy rate using 28-mer (Figure 3a) and 55-mer (Figure 3b) Balaenoptera dataset. The trustworthy rate increases with the increase in the counter length in both cases, i.e., 28-mer and 55-mer. All KmerCo has a positive trustworthy rate. The trustworthy rate is better if it is closer to zero. Hence, the countBF with 5-bit has the better trustworthy rate whereas the countBF with 16-bit has the least trustworthy rate.

Overall, KmerCo having countBF with counter length 8, 10 and 12 bits showcase better performance based on insertion time, and number of insertions per second. However, both countBF with counter lengths 10 and 12 bits waste 4 bits per cell which is a huge waste considering the whole data structure. Considering the trustworthy rate, 8-bit countBF has better performance than 10-bit or 12-bit countBF. Thus, we have considered countBF with an 8-bit counter for evaluating the performance of KmerCo using real DNA sequence datasets.

	28-mers	55-mers	CountBF size
Total K-mers / 8	20484059	20484055	9.1
Total K-mers / 4	40968118	40968111	18
Total K-mers / 2	81936236	81936222	36.19
Total K-mers	163872472	163872445	70.92
Total K-mers * 2	327744944	327744890	142.2

Table 2: **Details of the number of K-mers and countBF size in Megabytes by changing the number of total K-mers for construction of countBF of KmerCo using Balaenoptera dataset.**

1.2 Varying countBF data structure size

The KmerCo's Bloom Filter, i.e., countBF data structure size depends on the number of total K-mers present in the input datasets. We have conducted a few experiments to observe the performance of KmerCo by increasing and reducing the total number of K-mers. The counter length of countBF is 8 bits. Table 2 provides details regarding the number of K-mers considered for the construction of countBF and countBF size in megabytes. All KmerCo having different countBF size inserts the same number of K-mers and the inserted-to-ignored K-mer ratio is zero. Therefore, this section presents the comparison among various KmerCo having different countBF sizes based on the insertion time, number of insertions per second, and trustworthy rate.

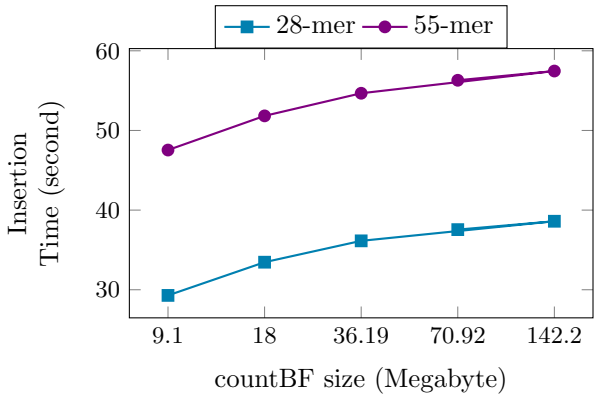
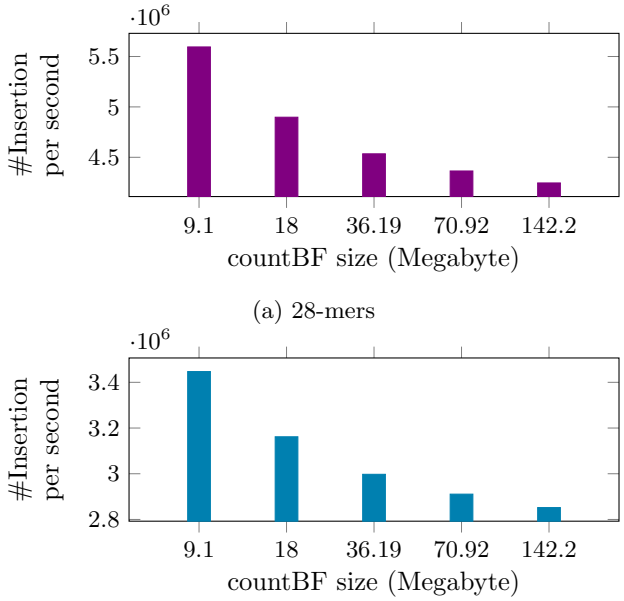


Figure 4: **Comparison of insertion time in second among various KmerCo having different countBF size using 28-mer and 55-mer Balaenoptera dataset. Lower is better.**

Figure 4 elucidates the comparison of various KmerCo with different countBF sizes based on insertion time in seconds using 28-mer and 55-mer Balaenoptera datasets. The figure clearly highlights that the insertion time decreases with a decrease in countBF size. The 142.2 MB countBF took 38.6 sec and 57.46 sec for 28-mer and 55-mer Balaenoptera datasets, respectively. Similarly, 9.1 MB countBF took 29.29 sec and 47.54 sec for the 28-mer and 55-mer Balaenoptera datasets, respectively. On average, the insertion time increases by 2.33 sec and 2.48 sec in the 28-mer and 55-mer Balaenoptera datasets, respectively.



(a) 28-mers

(b) 55-mers

Figure 5: **Comparison of the number of (a) 28-mers and (b) 55-mers inserted per second among various KmerCo having different countBF size using Balaenoptera dataset. Higher is better. #Insertion: Number of insertions.**

Figure 5 illuminate the performance of various KmerCo having different countBF size based on the number of insertions per second using 28-mer (Figure 5a) and 55-mer (Figure 5b) Balaenoptera dataset. The number of insertions per second increases with a decrease in KmerCo's countBF size. All KmerCo inserted the same K-mers but insertion time decreases with decreases in KmerCo's countBF size. On average, the number of insertions per second increases by 337356.4503 and 148775.9785 with the decrease in the KmerCo's countBF size in the 28-mer and 55-mer Balaenoptera datasets, respectively.

Figure 6 expound on the performance of various KmerCo having different countBF sizes based on the trustworthy rate using the 28-mer and 55-mer Balaenoptera

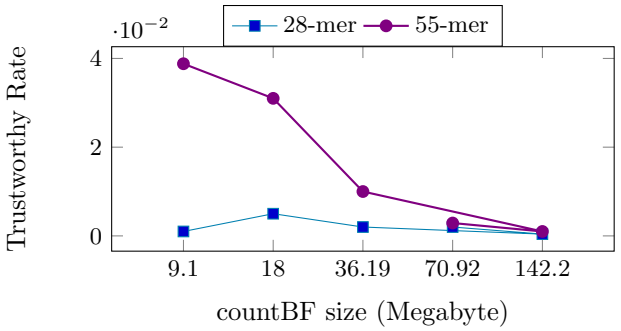


Figure 6: **Comparison of the trustworthy rate among various KmerCo having different countBF size using 28-mers and 55-mer Balaenoptera dataset. Positive and close to zero is better.**

datasets. The figure adduces the issues of having a tiny-sized Bloom Filter. The insertion time decreases and the number of insertions per second increases with a decrease in KmerCo's countBF size; however, small-sized countBF has more collisions. All KmerCo has a positive trustworthy rate. The KmerCo with 142.2 MB and 70.92 MB countBF has close to zero trustworthy rates. Others have close to zero trustworthy rates in the case of the 28-mer Balaenoptera dataset but in the case of the 55-mer Balaenoptera dataset, the trustworthy curve deviated upward indicating more erroneous K-mers considered as trustworthy K-mers.

Overall, the KmerCo with 142.2 MB and 70.92 MB countBF showcased better performance based on insertion time, number of insertions per second, and trustworthy rate. However, KmerCo with 142.2 MB countBF is twice the size of KmerCo with 70.92 MB countBF. Along with better performance, the K-mer counting technique should maintain a small memory footprint. Thus, we considered the KmerCo with 70.92 MB countBF for evaluating the performance of KmerCo using real datasets.