Kollector Supplementary

Algorithm S1: Computing a progressive Bloom filter. In our algorithm r is depicted as an integer threshold but may actually also be set to be a value between 0 and 1 representing the a the score described in the BioBloom Tools publication (Chu et al. 2014) which is a score based on the the relative to the length of the read sequence.

Input: Parameters k and r, seed sequence q and set of reads pairs $P = \{(p_0,p'_0),...,(p_{|R|},p'_{|R|})\}$ with $|q| \geq k$

Output: Bloom filter of tagged k-mers from reads and seed sequence

Function TagKmer(q, P, k, r) **begin**

```
F ← ∅                                  //F is not a set, but a Bloom filter
for i ← 0 to |q| - k + 1 do            //initial seed of the filter
   F ← F ∪ kmer(q[i],...q[i+k])
for i ← 0 to |R| do                    //add seeds to filter
   x ← 0, y ← 0                        //initialize k-mer overlap counts to 0
   for j ← 0 to |p_i| - k + 1 do       //check if first read k-mers present
      if kmer(p_i[j],...p_i[j+k]) ∈ F  //increment if matches
         x ← x + 1
   for j ← 0 to |p'_i|- k + 1 do       //check of second read k-mers present
      if kmer(p'_i[j],...p'_i[j+k]) ∈ F //increment if matches
         y ← y + 1
   if x > r or y > r do                //if k-mer counts reach threshold
      for j ← 0 to |p_i| - k + 1 do    //insert k-mers of first read
         F ← F ∪ kmer(p_i[j],...p_i[j+k])
      for j ← 0 to |p'_i|- k + 1 do    //insert k-mers of second read
         F ← F ∪ kmer(p'_i[j],...p'_i[j+k])
return F
```

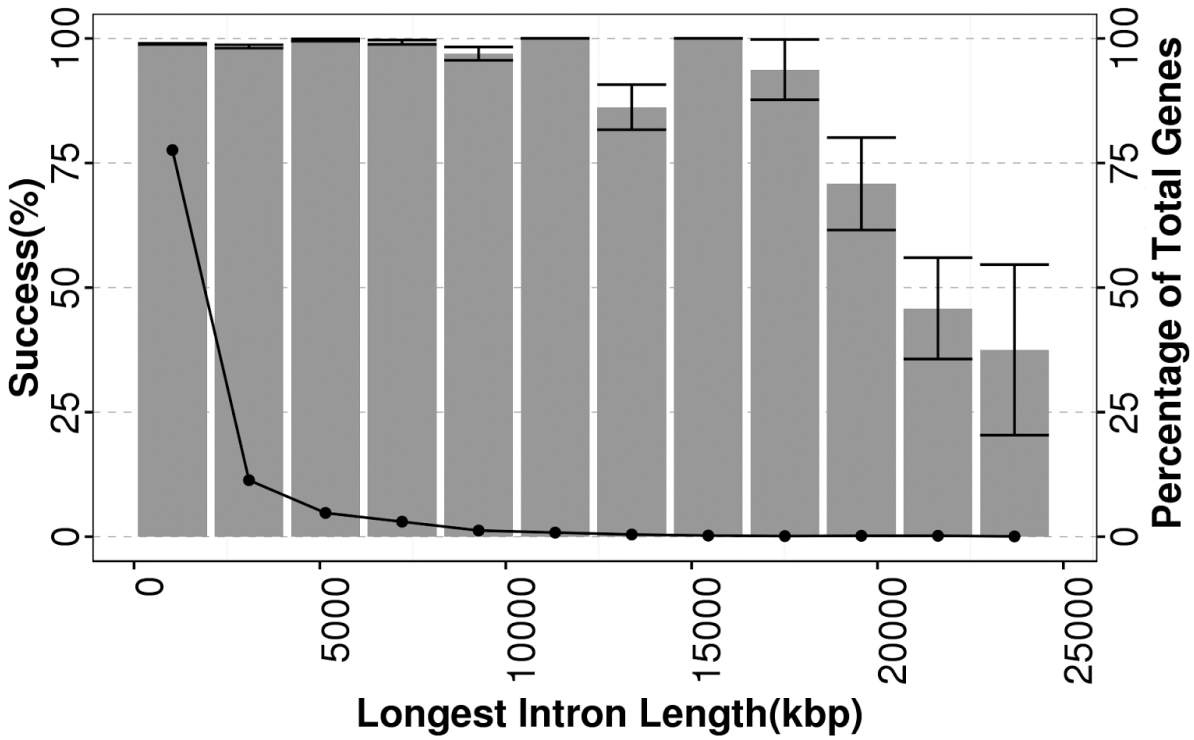

Figure S1. Proportion of successful gene assemblies vs longest introns (bars), with percentage of total genes in each bin(lines). 86% of the genes are concentrated in the first two bins

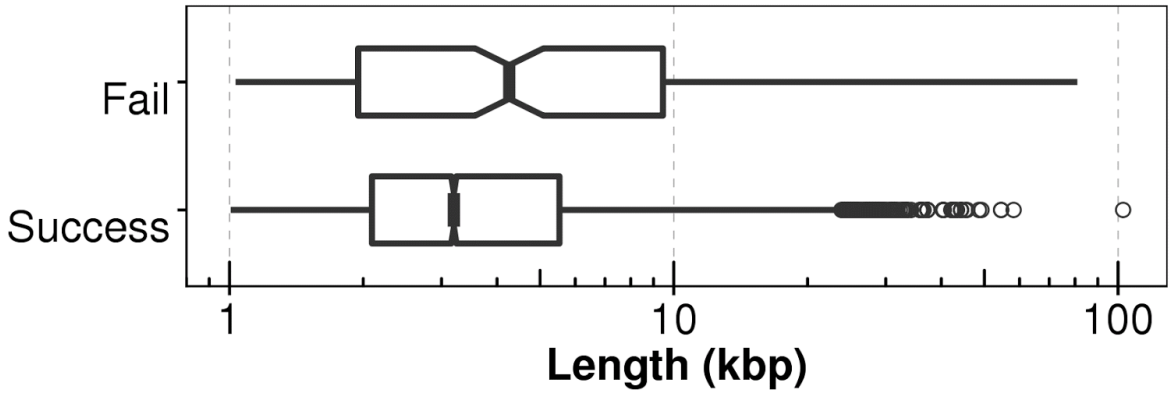

Figure S2. Length comparison between *C.elegans* target genes that are successfully assembled by Kollector and those failed to assemble. Notches in the boxes represent a 95% confidence interval around the median. Length difference between two groups is found to be statistically significant by Student's t-test ($p=1.5x10^{-5}$)