Related works:

1505.06550.pdf (arxiv.org)

Abstract:

Motivation:

The assembly of massive overlapping short reads randomly sampled from DNA fragments is a significant impediment to next-generation genome sequencing. To finish assembling, a simple task in many leading assembly algorithms must be completed: counting the number of occurrences of k-mers. The k-mer counting task will easily consume a vast amount of memory for large genomes.

Results: MSPKmerCounter is a disk-based method for effectively performing k-mer counting for large genomes with limited memory. It is based on a novel methodology known as Minimum Substring Partitioning (MSP) MSP divides fast reads into several disjoint partitions, allowing each partition to be loaded into memory and interpreted separately. It can achieve astonishing compression ratios by exploiting the overlaps among the k-mers extracted from the same short read.
