# Assignment 2: Related Work

- there are many challenges in genome assembly such as read length(short) and repeated sequences .
- so we will use de Bruijn Graph dependes on k_mers( A k-mer is a substring over a read with specific L length) .
- we find the k-mers of every read without repetion of every edge
- number of k_mers= L(read length)-K(k_mers length)+1
- example read(ATTTGC) and K=3 so first k_mers is (ATT) second is (TTT) third (TTG) and no.of k_mers =L-K+1=6-3+1=3
- A de Bruijn graph Gk(V, E) represents overlaps between k-mers, in which:
- – The set of vertices is defined by V = S = {s1, s2, ..., sp}, where S is a set of unique k-mers over a given set of reads
- – The set of edges is defined by E = {e1, e2, ..., eq}, where e = (si, sj) if and only if the k - 1-th suffix of si matches exactly the k - 1-th prefix of sj. si and sj must be adjacent k-mers in at least one read
- The life cycle of DBG for genome assembly can be summarized in two steps.

  First, construction involves the generation of all k-mers to generate a node per distinct k-mer and an edge between two nodes if these k-mers have a k - 1 overlapped in at least one read. In the second step, the processing is carried on by simplifying the graph and traverse it to generate contiguous genome regions called contigs