Assignment 1:Abstract and Introduction summary

. De-novo this mean that we will assemble reads together without reference genome this difficullte but we will find a way.

. genome assembly hase volvedto respond to the changes ininput datatype.

. we will talk about De-novo Genome Assembly; Short Read Genome Assembly; Long Read Genome Assembly; Hybrid Genome Assembly

. In this review, we provide

(1) An algorithmic description of the important processes in the workflow that introduces fundamental concepts and improvements.

(2) A review of existing software that explains possible options for genome assembly

(3) (benchmarking)A comparison of the accuracy and the performance of existing methods executed on the same computer using the same processing capabilities and using the same set of real and synthetic datasets.

.we allows a fair and precise comparison of accuracy in all aspects

. we identifies both the strengths and weaknesses of each method we will use.

. we will provide a detailed comparison of a broad spectrum of cutting-edge algorithms and methods.

introduction

. Sequencing is the process of reading DNA molecular chain into strings of A, C, T and G where each letter (called a niytogne base) represents one of the small molecules. Due to the available technology today, the DNA strand is broken into small pieces calld fragments and each fragment is then sequenced separately.

. The process of de-novo assembly has to do with assembling a species' genome for the first time.

.The resulting genome can then be used to facilitate genome assembly of other individuals of the same species, in a process called the reference-guided assembly. (To understand this, one might think of solving a jigsaw puzzle by using its cover photo for reference.)

.In contrast, de-novo assembly refers to assembling the genome without having a draft genome; this is a complex and difficult problem to solve.

. the purpose of the assembly, there are several expectations about the accuracy of the assembled genome and the time it takes to be assembled. Furthermore, the type and volume of sequenced data play important roles in the assembly process.

• The types of data that are available for assembly:

-It is important to compare short-read and long read assembly approaches to each other and to hybrid assembly techniques.

–Ultra-short read sequencing is suitable for de Bruijn graph assembly but not an OverlapLayout-Consensus (OLC) assembly approach.

–Next-Generation Sequencing (NGS) machines such as Illumina produce reads up to hundreds of bases (short-reads) with high accuracy. The error rate is less than 2%; most errors are of the substitution type and are less frequently short Indels .

–paired-end reads are sequenced from two ends of a long DNA fragment where the distance between them is approximately known. Such information can be used to improve the quality of the assembled genome.–PacBio and Oxford Nanopore are Single-Molecule Real-Time (SMRT) technologies that can sequence reads up to many thousands of bases long. However, they suffer from a high base calling error rate and can include long Indels.

• Sequencing remains an expensive process: - It is important to understand how different software programs respond to low and high read coverage depth when assembling a genome.

–If only short-read or long-read are provided, the coverage depth should be considered in configuring the assembly pipeline. For high coverage data, more filtering can be applied to collect high-confidence data for more accurate assembly.

Subject Areas