Introduction

Two groups have separately published draft assemblies of the naked mole-rat genome (Kim et al, 2011 & Keane et al, 2014). Both assemblies utilize a whole genome shotgun sequencing technique using high coverage Illumina data with various insert sizes. The purpose of this document is to explore the realm of genome assembly and address the utility of the NMR genomes.

Background

This section describes the relevant literature associated with the project, here I will discuss the history of sequencing technologies and the significance of algorithms associated with genome assembly.

*Sequencing technologies*

DNA sequencing was first introduced in 1975 by Fred Sanger & Alan Coulson, they had developed the “plus and minus” method for rapid DNA sequencing (Sanger & Coulson, 1975). Two years later, DNA sequencing with chain-terminating inhibitors (procedure) was developed, which were more accurate and rapid than both the plus and minus methods (Sanger et al, 1977). The automated Sanger method is considered ‘first-generation’ technology, and anything after is ‘next-generation’, then ‘third generation’ etc… Since first generation methods, computational instruments (sequencing-related) have been developing into the arrival of next generation sequencing (NGS) technologies (Illumina, 454, Ion Torrent etc.) which produce high-throughput data and has the ability to sequence DNA/RNA quicker and cheaper than traditional first gen Sanger methods (Metzker, 2009). The broadest application of NGS technology is the human re-sequencing project, built upon the original automated Sanger method which had produced the only finished grade human genome (International, 2004). A further example of NGS technology for genome assembly in a eukaryotic organism = the *de novo* assembly of the giant panda genome (Li et al, 2010). However, NGS methods have many hurdles and shortcomings, most notably short reads. Repeat sequences in a genome are often longer than NGS reads, which may lead to either gaps or misassemblies (Goodwin et al, 2016). To tackle this problem, long-read sequencing/third-generation technologies (Pacific Biosciences, Oxford Nanopore etc) have emerged, producing genome assemblies of unprecedented quality, providing a way to study genomes, transcriptomes, and metagenomes at high resolutions (van Dijk et al, 2018). After an assembly, the process of error correcting and removing lingering errors from contigs is called ‘polishing’. To further improve the quality of nanopore data, the polishing tool ‘Nanopolish’ can be implemented (Simpson et al, 2017). An example of an assembler, which uses long read technology for assembly is ‘Canu’, successor of Celera (Koren et al, 2017). The drawback with long-read technology is a high error rate (Jain et al, 2018).

*Assembly Algorithms*

Genome assembly from reads is an algorithm-driven automated process. Current assembly algorithms include greedy, overlap-layout consensus (OLC), de Bruijn graph (DBG) and hybrid algorithms (Miller et al, 2010). Greedy algorithm starts by joining reads that are most similar to each other in a pairwise, iterative fashion until all overlaps are combined. The process ends when no more overlaps are left or all the available overlaps conflict with already constructed segments (Pop et al, 2002). The problem with this approach are that it is fundamentally local in nature and ignores long-range relations between pairs of reads, useful for solving/detecting the problem of repeats (Pop et al, 2002). A recent example of a greedy approach assembler for *de novo* transcriptome assembly, ‘TraRECo’ (Yoon et al, 2018).

OLC and DBG are two commonly used classes of algorithms (Li et al, 2011). OLC works in three stages, first all the overlaps in all reads are located (‘O’) then the overlaps are laid out (‘L’) on a graph and lastly, consensus (‘C’) string is extrapolated (Li et al, 2012). OLC is said to be an intuitionistic assembly algorithm (Li et al, 2011) where all overlapping nodes are combined once producing a genome sequence (Supratim Choudhuri, 2014). An examples of a eukaryotic genome assembly, which utilise OLC algorithms and the Celera assembler is the whole-genome assembly of the *Drosophila* (Myers et al, 2000).

The origins of the DBG dates back to 1946, where a Dutch mathematician called Nicolaas de Bruijn became interested in the superstring problem (Bruijn, 1946). In fact, before this, the theoretical discovery used in 1735 to solve the *Bridges of Konigsberg Problem* by the mathematician Leonhard Euler (Euler, 1736) enabled the assembly of billions of short sequencing reads (Compeau et al, 2011). The de Bruijn graph algorithm works by chopping down each sequence into smaller reads called *k*-mers to solve the problem of scalability & help improve the crisis of different initial read lengths (Khan et al, 2018). The graph is structures in a way where each individual read is represented as a node (‘O’) and overlaps (between reads) are represented by arrows (‘->’) also known as ‘directed edges’, merging two reads together. There are two sub strategies for genome assembly using the DBG algorithms they are (1) Hamiltonian cycles, and (2) Eulerian cycles (Compeau et al, 2011). In a Hamiltonian cycle, ‘vertices’ represent *k*-mers/nodes and ‘edges’ are the pairwise alignments. Following the edges in a numerical order, enables one to create a genome by combining pairwise alignments between consecutive reads (Compeau et al, 2011). Each vertex is only visited once in a Hamiltonian cycle, where *k*-mers are produced from vertices. For the Eulerian cycle, vertices/nodes are (*k*-1)-mers and edges are *k-*mers. Finding the Eulerian cycle enables one to build a genome assembly where each consecutive *k*-mer in an alignment is shifted by one place (Compeau et al, 2011). Each edge is only visited once in a Eulerian cycle, where *k*-mers are produced from edges. An example of an assembler which manipulates de Bruijn graph algorithms for genomic assembly is ‘Velvet’ (Zerbino, 2008).

*NMR Genomes*

Two groups have separately published draft genome assemblies of an individual naked mole-rat/NMR (Kim et al, 2011 & Keane et al, 2014). Both assemblies utilise whole genome shotgun sequencing using high coverage Illumina data with various insert sizes (Lewis et al, 2016). One major fault with this approach are that assemblies generated are fragmented and disjointed, containing high levels of unfilled gaps (15%) preventing good analysis of gene expression & regulation (Lewis et al, 2016). Initial short-read Illumina assemblies were shown to contain significant levels of misassemblies and collapsed homologous genes and pseudogenes (Zimin et al, 2012). However, with the emergence of long-read sequencing/third-generation technologies (Pacific Biosciences, Oxford Nanopore etc) the quality of the sequencing data will continue to improve, effectively constructing genome assemblies of higher unprecedented qualities and paving the way for new genome, transcriptome and metagenomic studies at higher resolutions (van Dijk et al, 2018).