**De bruijn graphs**

**Greedy algorithms**

**SGA & Fermi algorithms**

Hybrid assembly

Homopolymers, repeats, palindromes

Contamination and mtDNA

Construction of a genome sequence currently relies on acquisition of fragmented sequences (reads), which in turn are ‘stitched together’ by computer programs into intact assemblies. Ultimately, we want an assembly with the largest contigs and scaffolds, so as to get the maximum degree of contiguity and accuracy while spanning a long genomic region (or alternatively, an entire genome). The initial assembly approach relied on finding overlaps between reads, assembling sufficiently overlapping fragments into contigs, which in turn are assembled into scaffolds. This Overlap Layout Consensus (OLC) approach worked well for assembly tools in the heydays of DNA sequencing, when reads were short and high quality, and BAC clones were the primary DNA source being sequenced. With the advent of short-read sequencing using the shotgun approach, the sequence assembly problem got more complicated. The conservative assemblers that traditionally relied on extensive overlap between reads and used shorter contigs cannot be directly applied to reconstruct the organization of an entire genome. Common examples are non-uniform coverage of the target in WGS, and processing of fragments representing repeats – a genomic feature not handled that well with OLC since they can only be resolved by a spanning read or a read pair that is uniquely anchored on both sides. Repeat separation is assisted by high coverage but confounded by high sequencing error. As a method for avoiding misassemblies using OLC, the de Bruijn graph was adapted for sequence assembly. The nodes of a de BRuijn graph, in the assembly paradigm, represent fixed-length subsequences (fixed length reads, or k-mers), and edges represent a single base shift in the k-mer from one node to the other. This approach is quickly able to discount single base sequencing errors by excluding short, dead-end paths, and can represent short tandem repeats as ‘cycles’ in the graph. It is also able to identify mis-matches and SNPs due to the formation of ‘bubbles’, although the discrimination between a sequencing error and a SNP is difficult and relies on read count estiamtes for the alternative reads.

Due to the iterative extension model of the OLC approach, initial assemblers were implemented using greedy algorithms. *Greedy algorithms can get stuck at local maxima if the contig at hand takes on reads that would have helped other contigs grow even larger*.

*In order to put together partly overlapping fragments, one could follow a Hamiltonian path (whereby one overlaps certain k-mers – Overlap Layout Consensus), or a Eulerian path (whereby one looks for a 1-mer shift in each subsequent node of a de Bruijn Graph)*.

SGA & Fermi algorithms remain the linear approximations of the OLC assembly paradigm. Alternative algorithms that seek to get the maximum degree of contiguity in the contig assembly stage generally fall under the category of greedy algorithms.

The advent of 3rd generation sequencing technologies (PacBio, Oxford NanoPore), has taken away the need for redundancy in sequences during assembly. This, however, comes at the cost of a higher error rate in a single long read. While PacBio can generate long reads of up to 30 kilo base pairs (kbp), and Oxford NanoPore claims to generate ones that are **???**, the error rate for these sequencing methodologies goes up to 15%. Assembly algorithms have hence been significantly optimized for long reads data. One such way to generate contiguous assemblies with a lower error rate, while using 3rd generation sequencing data, is a hybrid method that can combine long read data with short, high accuracy sequences (Koren et al, 2012), or simply a long read sequencing run with a high coverage (30x – 70x or above). Application of the hybrid assembly protocol on simulated and real data has shown an improvement in read accuracy from as low as 80% to over 99.9% (Koren et al, 2012).

The recently finished Genome Assembly Gold-standard Evaluations(**GAGE**), have shown that data quality, rather than the assembler, often has a dramatic effect on the quality of an assembled genome. A lot of short-read assemblers have inbuilt QC metrics (for example, Illumina’s Chastity filter) that identify reads with base calling errors (lower k-mer coverage), ensure that poor quality reads with a low signal to noise ratio get filtered out. Numerous genomic sources of contamination can nevertheless contribute to assembly errors – viral insertions (transposons) and mitochondrial genome reads being the common ones. Targeted sequencing and deeper coverage have been suggested to ‘fill in’ poorly assembled areas of the genome under consideration. It should also be noted that sometimes too deep a coverage can lead to increased sequencing errors, since in that case a greater possibility exists for a sequencing error to be disproportionally represented in reads. This can eventually cause misalignment issues. Another interesting issue that highlights the importance of the tissue retrieval and sequencing methodology is that energetically active tissues often yield a disproportionately high amount of mtDNA – the resulting extreme read depth for reads derived from the mtDNA can eventually cause downstream issues with identification of duplicated regions of the genome, and ultimately be a waste of sequencing effort.

Works Cited

1. Chaisson, MJP et al. **Genetic variation and the de novo assembly of human genomes**. *Nat Rev Genetics* 16:627-640 (2015).