"The evolution of assembly algorithms"

Since the advent of Sanger sequencing, the development of new approaches to sequence DNA has gone hand in hand with the development of algorithms to 'stitch together' the resulting fragments into an intact genome assembly. Initial forays into the construction of a reference mammalian genome in the 1990s relied on a clone-by-clone based sequencing approach - assembling 200 kb long fragments cloned in bacterial artificial chromosomes (BACs) after constructing a 'genome map'[1]. However, the extensive time and financial input required by this approach prompted the adaptation of 2nd generation sequencing technologies onwards of 2003.

The problem of putting together the BAC amplified fragments into an entire genome could be approximated to the shortest superstring problem - given a set of input strings, find the shortest string T such that every input string is a substring of T. Since the shortest superstring problem is NP-hard, scientists used an efficient approximation instead - 'given any read or contig, add the next high-scoring overlapping read or contig'[2]. This greedy merging heuristic worked with the human genome assembly due to the a-priori chromosome maps, which informed local *and* long-range relationships between reads. However, greedy algorithms are inherently local in nature[3]. The application of greedy algorithms for processing short-read (2nd generation) data could hence result in local maxima due to a single contig accruing reads that would otherwise have helped other contigs grow longer. These inherent drawbacks of this approach ultimately spurred the development of new algorithms for 2nd generation sequencing data[3].

The first alternative approach for assembling 2nd generation sequencing data, Overlap-Layout-Consensus (OLC), relies on finding a 'Hamiltonian path' where each sequence fragment (or node) is visited only once: overlaps between reads are identified in a pairwise manner, sufficiently overlapping fragments assembled into contigs, which in turn are assembled into scaffolds[1]. However, overlap identification stage is rather time consuming due to the need for pairwise comparisons between all reads. Additionally, while the OLC approach aims to reach a global minimum with consensus overlaps, it is still confounded by repeats - infact, the identification of Hamiltonian paths has been shown to be an NP complete problem.

Given the drawbacks of a Hamiltonian assembly approach, an alternative Eulerian approach (with linear time implementation) based on de Bruijn graphs was proposed. In this approach, instead of explicitly calculating pairwise overlaps, the reads are broken into overlapping k-mers. The resulting graph structure has single k-mers as nodes, the edges represent a single base shift in the connected k-mer nodes, and the Eulerian path traces each edge only once[1,3]. This approach is thus able to represent single base sequencing errors (bubbles in the graph), polymorphisms (equally high coverage for alternate paths between two non-contiguous nodes), and repeats ('cycles' formed in the graph) far more efficiently than OLC. ALLPATHS and ABySS are two popular assembly methods that use this assembly approach.

Over the years, linear approximations for OLC have also been developed, namely SGA and Fermi[4]. The String Graph Assembler approach (SGA) was proposed in 2010 to overcome the complete reliance of de Bruijn graphs on k-mers for full read construction. SGA built a de Bruijn graph equivalent based on full overlaps between reads (akin to an overlap consensus approach), instead of cutting up reads into k-mers. However, SGA was able to make this full-read adaptation less computationally intensive than the pure OLC methods by using a Burrows-Wheeler Transform and FM-indexing to accelerate string comparisons. The Fermi algorithm further extended SGA's FM-index based adaptation of the OLC paradigm. This algorithm managed to generate a 'phasing' context for the input reads in the final assembly by preserving read information in the final set of 'unitigs' instead of collapsing them into a single consensus call. This allows the preservation of heterozygous events from the raw reads in the final assembly, improving SNP and INDEL calling against a reference genome.

The short reads generated by 2nd generation sequencing technologies make it difficult to characterize repeats and other long-range genomic variants. The advent of 3rd generation sequencing technologies such as PacBio hopes to overcome this issue by generating longer reads (10 kilo base pairs on average) - sequencing a single DNA molecule in real time instead of chopping it up into fragments. This, however, comes at the cost of a higher error rate, which currently ranges between 15% (PacBio) to 33% (Oxford NanoPore). Concordantly, assembly algorithms have needed to be significantly optimized to utilize the informational enrichment in assembly from long reads while accommodating for their high error rates. Hybrid methods overcome this high error rate by combining low-pass long read data with short, high accuracy sequences. This approach improves read accuracy from as low as 80% to over 99.9%[5]. Alternatively, long read errors can be reduced by consensus base calling from the same template by looping the read-generating polymerase back over the same dsDNA template with 'circular' terminal adapters (PacBio CCS). PacBio CCS long reads have been found to be the best combination with Illumina short-range paired data in the Celera hybrid assembler.

Repeats have remained a problem in assembly ever since the heydays of sequencing. Long repetitive regions of the genome are difficult to cut and clone into BACs, and short reads are unable to resolve repeats that are longer than the reads themselves. In these cases, identification of contiguous repeats, low frequency repeats, or even regions of the genome that share perfect repeats, relies on either using read pairs that span the repeat region, or by correlating reads with patterns in their base compositions[2]. Several assemblers simply evade this problem by masking out repeats[3], others rely on deep coverage to identify statistically significant clusters of reads spanning interspersed repeats. Long read data with hybrid assembly can help resolve repeats, whether they be low frequency, single base repeats (homopolymers), tandem (>1 base) repeats, and interspersed repeats (separated by unique sequences), simply by capturing a longer contiguous genomic region that would contain the repeat region(s)[5]. Lastly, palindromic DNA sequences can often cause short reads-based assemblies to 'fold onto themselves', but this problem can be overcome in DBG assemblers by requiring odd-numbered k-mers[5].

The Genome Assembly Gold-standard Evaluations (GAGE)[6] have shown that data quality has a dramatic effect on the quality of an assembled genome. Viral inserts and mitochondrial genome reads are common contaminants that can be usually filtered out by using a short read aligner like BWA to delete all fragments mapping to the contamination sequences. However, over-representation of contaminant-derived sequences (ex. a disproportionately high amount of mtDNA in energetically active tissues) can result in extreme read depth, which can confound identification of duplicated regions of the nuclear genome, and ultimately be a waste of sequencing effort[5]. At the other end of the spectrum, adapter dimerization and protocol biases can result in fewer reads representing a genomic region, subsequently resulting in poorer assembly. Targeted sequencing and deeper coverage have been suggested to 'fill in' such areas, and single molecular sequencing holds promise to address this issue as well[5].

Works Cited

1. Chaisson, MJP et al. **Genetic variation and the de novo assembly of human genomes**. *Nat Rev Genetics* 16:627-640 (2015).
2. Miller, JR et al. **Assembly algorithms for Next-Generation Sequencing Data**. *Genomics* 95(6): 315-327 (2010).
3. Pop, M et al. **Genome Sequence Assembly: Algorithms and Issues**. *Computer* 35(7):47-54 (2002).
4. Ekblom, R and Wolf, JBW. **A field guide to whole-genome sequencing, assembly and annotation**. *Evolutionary Applications* 7(9):1026-1042 (2014).
5. Koren, S et al. **Hybrid error correction and de novo assembly of single-molecule sequencing reads**. *Nature Methods* 12(8): 693-700 (2012).
6. Salzberg, SL et al. **GAGE: A critical evaluation of genome assemblies and assembly algorithms**. *Genome Res* 22(3): 557-567 (2012).