**Optical mapping**

Citation: Shelton. J.M., Coleman M.C., Herndon. N et al (2015) Tools and pipelines for BioNano data: molecule assembly pipeline and FASTA super scaffolding tool. *BMC Genomics.* **16**(734) [Accessed on 21/02/2019] Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4587741/>

Source: <https://bionanogenomics.com/wp-content/uploads/2018/04/30073-Bionano-Solve-Theory-of-Operation-Hybrid-Scaffold.pdf>

Tools:

Bionano softwares primarily used for hybrid scaffolding and analysing structural variation in genome.

Bionano Irys hardware is used for optical mapping with long-DNA sequencing reads as input and a long genome map as output.

hybrid-scaffold = long genome map + sequencing assembly.

Uses of a long genome map:

* validate order of sequencing fragments.
* identify chimeric joins in sequence assembly
* evaluate the size of gap between adjacent sequences.

How to construct hybrid scaffold:

* long-read sequencer produces long DNA reads from sample.
* Irys used for optical mapping. After which a long-genome map is produced from long DNA reads.
* hybrid-scaffold = long genome map/ bionano assembly + sequencing assembly.
* Assume the assembly has good contiguity and quality, >100kbp

Purpose of a hybrid scaffold:

* A representative study of chromosomal structures.
* Improve N50 of assembly.

How Irys works (optical map + hybrid scaffolding automated):

1) generate in silico maps for sequence data;

2) align sequences against Bionano genome maps to identify and resolve potential conflicts (chimeric joints) in either data set;

3) merge the non-conflicting maps into hybrid scaffolds;

4) generate an alignment between the sequence maps and the hybrid scaffolds;

5) generate AGP and FASTA files for the scaffolds.

Chimeric join leads to junction conflicts:

short-reads unable to span across long DNA repeat.

After alignment the junctions for alignment with bionano assembly will be different to alignment with sequencing assembly.

**BUSCO-tests.**

The BUSCOs are evolutionarily well conserved and inherited as orthologous single-copy genes diverging only with speciation. Most of the BUSCOs found in more than 90% of all species have been catalogued within the OrthoDB platform. The tool downloaded and used in this investigation is BUSCOv3, a package that contains a multitude of tools such as tblastn to align AA sequence from OrthoDB to translated query sequence, HMMER which produces multiple sequence alignments and build HMM profiles, as well as Augustus gene predictor that obtains the locations and annotates the functions of recovered genes and determine if they are orthologous.

A BUSCO test assigns 4 different metrics on recovered genes based on expected gene length recovered from sequencing. If the ortholog gene length is within 2 standard deviation of the mean BUSCO length from the set the expectation would fall within 95% and the content is considered “complete.” If the expectation is below 95% then the content is fragmented. With little or no length of gene recovered the genetic content would be labelled as missing. Should the data contain more than one “complete” gene, the content would be labelled as duplicate but this could indicate errors as BUSCOs evolve in single copy. Regardless of the metric, the no of genes would be assigned to each dataset as a measure of the quality to result’s interpretation.

A BUSCO set is constructed by creating multiple sequence alignments, MSAs from the known sequences of a clade’s species with Clustal Omega. Subsequently the MSA need to be processed with HMMER3, a software packages that constructs a Hidden Markov Model profile of the MSAs at the level of amino acids. This is to ensure that the genes compared against BUSCO sets are orthologs. (Simao, et al 2015)

**Supplementary**

Terms of questions:

contiguity, N50, scaffold, chimeric join/junction conflict between assemblies,

These terms are relevant in lectures:

[ w2:l5, w5:l12-15 ]