Physlr: Construct a Physical Map from Linked Reads

Shaun D Jackman Justin Chu Vladimir Nikolić Amirhossein Afshinfard
Lauren Coombe Gokce Dilek Sauparna Palchowdhury Hamid Mohamadi
Rene L Warren Joerg Bohlmann Inanc Birol

Background

Sequencing large molecules of DNA has drastically improved the contiguity of genome sequence assemblies. Long read sequencing has reduced sequence fidelity compared to short read sequencing, and it is currently more expensive. With linked read sequencing from 10x Genomics Chromium, reads originating from the same large molecule all have the same 16 nucleotide barcode identifier. Linked read sequencing combines the benefits of large DNA molecules with the sequence fidelity and cost of short read sequencing. Two key challenges present themselves when assembling linked read sequencing. First, each of the roughly one million barcodes contain approximately ten large molecules of DNA. Second, each individual molecule is sequenced to less than one fold depth of coverage. For this reason, the assembly of linked reads typically follows a bottom-up approach, starting with an assembly of all the reads, at first ignoring the linked-read barcodes. These initial contigs are then scaffolded using the long-range information of the linked reads. Scaffolding may be difficult if, for any reason, the contiguity of this initial assembly is low. A top-down approach could be modeled after the overlap-layout-consensus (OLC) assembly paradigm, first determining the order of the large DNA molecules, before determining the fine-scale sequence. Our tool, Physlr, implements this top-down approach to construct a physical map of large molecules sequenced using linked reads.

Methods

Each barcode is a set of reads, and Physlr treats each barcode as a bag of k-mers. To reduce memory requirements and improve computational efficiency, only those k-mers that are minimizers of each read are retained. Minimizers that occur in only a single barcode are discarded, as likely resulting from a sequencing error. The most frequent minimizers are also removed, as likely derived from repetitive sequence. All pairs of barcodes that share a substantial number of minimizers are then identified. A weighted, undirected barcode-overlap graph is constructed, wherein each vertex represents a barcode, each edge represents the intersection of the minimizers of two barcodes, and the weight of the edge is the number of common minimizers. If each barcode contained a single molecule of DNA, we would have the desired molecule-overlap graph, and we could move on to the layout step of OLC. Because each barcode is composed of multiple molecules, each vertex must first be separated into its component molecules.

For each vertex we identify a vertex-induced subgraph of its adjacent vertices and the edges between those vertices. This subgraph is composed of communities of vertices, where each community is born from a DNA molecule. These communities are identified using k-clique community detection. This barcode vertex is then removed and replaced by one molecule vertex for each of its detected communities, and each of its incident edges is assigned to one of these newly-created molecule vertices. This method reconstructs the underlying molecule-overlap graph from the barcode-overlap graph.

With the molecule-overlap graph constructed, we proceed to the layout stage of OLC assembly. A maximum spanning tree (MST) is used to compute a fast and approximate solution to the traveling salesman problem. The MST is computed for each connected component of the graph, identifying the longest path through the MST. This path naturally does not visit every vertex, but it is not necessary at this point to include every molecule in the physical map. All vertices in a path and their neighbours are removed from the graph, and this process is repeated to complete the construction of the physical map.

This method determines the number of molecules present in each barcode, and it constructs a physical map of these molecules. The physical map is a set of contigs, where each contig is an ordered list of barcodes. With the physical map constructed, two applications readily present themselves. First, a subset of reads taken from a set of barcodes from one physical map contig (or a portion of one) may be assembled to yield a local assembly of a particular genomic region. This process may be repeated to assemble the entire genome. Second, an existing assembly may be mapped to the contigs of the physical map to order and orient the contigs (or scaffolds) of that existing assembly.

Results

We constructed a physical map of the 137 Mbp fruit fly (*Drosophila melanogaster*) genome from linked reads using Physlr. The physical map has a single contig for five of six chromosomes (excluding chrY), and three large contigs compose the sixth chromosome (fig. 1). An ABySS 2.0 (Jackman et al. 2017) assembly of the linked reads was scaffolded by mapping it to this physical map. The resulting assembly has an NG50 of 16.9 Mbp, approaching the 20.8 Mbp NG50 of Supernova (Weisenfeld et al. 2017). When the Supernova contigs (derived by splitting its scaffolds at Ns) are scaffolded using the physical map, the resulting assembly has an NG50 of 20.5 Mbp. Comparing this assembly to the reference genome indicates no translocation misassemblies.

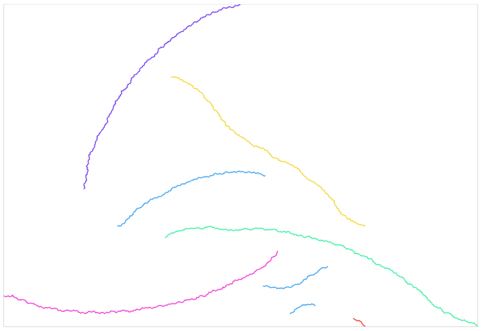

Figure 1: The physical map of fruit fly (*Drosophila melanogaster*) constructed using Physlr

We constructed a physical map of the 1.34 Gbp zebrafish (*Danio rerio*) genome from linked reads using Physlr. The scaffolds of the Supernova assembly of the linked reads was further scaffolded by mapping it to this physical map. The NG50 of the Supernova assembly improved from 4.8 Mbp to 9.1 Mbp by scaffolding with Physlr. Comparing this assembly to the reference genome indicates only three translocation misassemblies. Development is ongoing to identify and correct these misassemblies.

Conclusions

Physlr constructs a physical map of large DNA molecules sequenced using the linked reads from 10x Genomics Chromium without first assembling those reads. This physical map may be used to scaffold an existing assembly, assembled either from the same linked read sequencing data set, or from an entirely separate sequencing data set, enabling hybrid assemblies of linked reads and other types of sequencing. Physlr can employ multiple libraries of linked reads, necessary for genomes larger than mammals. We are currently developing Physlr to scale to the assembly of conifer genomes, which can exceed 20 Gbp.

References

Jackman SD, Vandervalk BP, Mohamadi H, Chu J, Yeo S, Hammond SA, Jahesh G, Khan H, Coombe L, Warren RL, et al. 2017. ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter. *Genome Research* **27**: 768–777. https://doi.org/10.1101/gr.214346.116.

Weisenfeld NI, Kumar V, Shah P, Church DM, Jaffe DB. 2017. Direct determination of diploid genome sequences. *Genome Research* **27**: 757–767. https://doi.org/10.1101/gr.214874.116.