An algorithm for assembling human centromeres from long reads

Shubhakar R. Tipireddy[1], Yuta Suzuki[1], Shinichi Morishita[1]

1. Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, University of Tokyo

Despite the advancements in the sequencing technologies, assembling the human centromere regions still remains a challenge due to its near-identical tandem repeat structure. Assembling these regions relies on a small number of SNVs (single nucleotide variants) present in the alpha-satellite monomers acting as unique markers for the sequence assembly. Identifying these sparsely distributed SNVs from only long-reads is almost infeasible due to their high error rate. Our method utilizes both the highly accurate short reads from the Illumina sequencing and erroneous long reads from the PacBio sequencing of a personal genome and attempts to assemble human centromere regions.

Starting with an initial set of 3736 monomers belonging to the human centromere region (K. Miga et al. 2014), we generated 590 clusters of monomers and corresponding representative monomer sequences such that the inter-cluster sequence similarity is at most 90%. Sets of monomer variants were identified for each representative monomer through alignment with the short-reads. Each monomer variant is characterized by a set of SNVs present in it and has a unique identifier (ID) assigned to it. Occurrences of representative monomers in PacBio reads are detected through sequence alignment and a set of mismatches for each occurrence is obtained. Replacing each occurrence of a representative monomer in PacBio reads with the best-suited monomer variant ID will effectively reduce the sequencing error in PacBio reads while carefully preserving the SNVs. Comparing the SNVs of each monomer variant with the mismatches of PacBio-monomer alignment leads to the identification of the best suiting monomer variant for each monomer occurrence. Our method encodes each PacBio read into a series of monomer variant IDs (used as unique markers) and computes the alignments between them. Through this method, we have successfully obtained a read overlap-layout of length up to ~500,000 base pairs.