COMPUTATIONAL BIOLOGY WEBINAR @ IMSc

REPEAT-AWARE METHODS FOR MAPPING AND ANALYSES OF UNDER-EXPLORED REGIONS IN THE HUMAN GENOME

DR. CHIRAG JAIN

INDIAN INSTITUTE OF SCIENCE, BENGALURU

THURSDAY, 10 DECEMBER 2020, 3 PM IST

The recent completion of human chromosomes X and 8 by the Telomere-to-Telomere Consortium has revealed highly repetitive satellite and segmentally duplicated sequences that were previously inaccessible to both de novo genome assembly and re-sequencing approaches. Over 10% of the current human reference cannot be reliably mapped with short sequencing reads, and this number will only grow as additional reference gaps are completed. Thus, the challenge will be to map reads and call variants within such repetitive, yet functionally important, regions of the genome. Long-read sequencing technologies hold promise, but the problem of accurately mapping long reads to complex genomic repeats must still be addressed. In this talk, I'll highlight the fact that existing long read mappers often yield incorrect alignments and variant calls within long, near-identical repeats, as they remain vulnerable to allelic bias. In the presence of a non-reference allele within a repeat, a read sampled from that region could be mapped to an incorrect repeat copy because the standard pairwise sequence alignment scoring system penalizes true variants.To address the above problem, we propose a novel, long read mapping method that addresses allelic bias by making use of minimal confidently alignable substrings (MCASs). MCASs are formulated as minimal length substrings of a read that have unique alignments to a reference locus with sufficient mapping confidence (i.e., a mapping quality score above a user-specified threshold). This approach treats each read mapping as a collection of confident sub-alignments, which is more tolerant of structural variation and more sensitive to paralog-specific variants (PSVs) within repeats. We mathematically define MCASs and discuss an exact algorithm as well as a practical heuristic to compute them. The proposed method, referred to as Winnowmap2, is evaluated using simulated as well as real long read benchmarks using the recently completed gapless assemblies of human chromosomes X and 8 as a reference. We show that Winnowmap2 successfully addresses the issue of allelic bias, enabling more accurate downstream variant calls in repetitive sequences.

GOOGLE MEET LINK:

meet.google.com/yah-wfwv-dns