Recovery and annotation of satellite DNA sequences from assembled and unassembled reads

Aleksey Komissarov

Institute of Cytology RAS, St. Petersburg, Russia, aleksey.komissarov@gmail.com

Tandemly repeated DNA represents a significant portion of eukaryotic genomes. The large tandem repeats including satellite DNA are the main component of centromeric and pericentromeric regions that are mostly unassembled. More than 18 percent of mouse unassembled reads are similar to the mouse major satellite DNA assembled into just several arrays in the mouse reference genome. The incomplete assembly and characterization of large tandem repeats and satellite DNA limits experimental studies. Here, we present a workflow for satellite DNA recovery and annotation from genome assemblies or from unassembled reads.

In the case of assembled genome, a non-redundant set of tandem repeats found with TRF in assembled sequences is divided into following types: microsatellites, perfect minisatellites, minisatellites, tandem repeats related to mobile elements, and large tandem repeats including satellite DNA. We suggest two following approaches to the distance computation between arrays: a distance based on a pairwise sequence alignment and a distance based on a number of common k-mers between two arrays; using these distances we constructed tandem repeats similarity graph that allows to define tandem repeats families and subfamilies for large tandem repeats, each family named accordingly to proposed uniform nomenclature. Annotation step includes prediction of the position in the reference genome, estimated copy number, presence of known DNA motifs (e.g. CENP-B box, G-quadruplex , or pJalpha), presence of high order repeats, and predicted chromosome specificity.

In the case of unassembled genome, we pick out all reads containing highly repeated k-mers. Each selected read is divided into following groups based in its k-mer profile: reads similar to known microsatellites, reads similar to dispersed repeats from Repbase collection, and reads that could be satellite DNA. Then we divide satellite DNA reads into group with similar

k-mer profiles and try to recover monomer consensus for each group using the most frequent k-mer for given group as an anchor for monomer assembly. Finally assembled monomers are used for annotation as described for tandem repeats found with TRF in assembled sequences.