MOSCAL: Detection of mosaic variants using linked-read sequencing

Genomic mosaicism describes the presence of multiple cell lineages derived from distinct fertilized eggs. Detection of mosaic variants has unraveled the genomic pathogenicity of many diseases including early developmental disorders and cancers. However, accurate identification of mosaic variants has been frequently confounded owing to the low variant allele frequency and the absence of a clear matched control for germline variant filtration. While several strategies to achieve higher precision, such as read-backed phasing with nearby heterozygous germline SNPs have been employed, the lack of such SNPs within a short read hindered wider application. To secure a sufficient number of phased heterozygous SNPs, we applied the Linked-read sequencing technology (10x Genomics), which leverages the barcode DNA to generate data type provides contextual information about the genome from short-reads. Using this technology, we developed a novel variant pipeline MOSCAL that utilizes distant heterozygous germline SNPs that are phased into the regions of interest. In benchmarks on datasets used for Single-sample Mosaic SNV calling with linked read (Samovar) training, our pipeline achieved improved accuracy (F1-score of 0.729 and 0.604 in 60x and 30x sequencing) than previous methods (MuTect2; 0.542 and 0.318, and Samovar; 0.619 and 0.477, in 60x and 30x sequencing, respectively). We expect that the use of linked-read sequencing would provide new options for identifying mosaic variants.