Abstract ID:- 174

Abstract Topic:- Evolutionary and population genetics

Abstract Title:- Beyond short variants: Building a structural variation resource of Indian populations

Presenting author name :- Sofia Banu

Presenting author institute:- CSIR - Centre for Cellular and Molecular Biology, Hyderabad; Academy of Scientific and Innovative Research (AcSIR), Ghaziabad

Co-authors name:- Sreelekshmi MS, A Sreenivas, Lamuk Zaveri, Payel Mukherjee, Genome India Consortium, Karthik Bharadwaj Tallapaka, Divya Tej Sowpati

Co-authors institute:-CSIR - Centre for Cellular and Molecular Biology, Hyderabad; Academy of Scientific and Innovative Research (AcSIR), Ghaziabad

Aims:-Structural variants (SVs) are large genomic rearrangements which span at least 50 bp and their large size represents a formidable mutational force in terms of diversity. Although heavily implicated in diseases, SVs have also been observed to contribute to variation in healthy individuals. Despite the uptick in population scale genome sequencing projects, representation of Indian diversity has been limited. The Genome India project aims to fill this gap by sequencing 10,000 genomes across representative communities, with a goal of capturing SNP and small InDel variation amongst the groups. To complement this, we embarked on a pilot study to identify structural variants from different Indian populations.

Methods:- Since comprehensively identifying SVs is limited by solely using short reads, we aim to illustrate the merit of inclusion of newer technologies such as long read sequencing. We generated long-read data for a subset of samples from Genome India on the Oxford Nanopore platform. Raw nanopore data was basecalled using dorado in high accuracy mode. Reads were aligned to the reference genome with minimap2. SV identification and merging was performed using sniffles2. In addition, a comparative analysis was performed using hg38 and T2T reference genomes.

Results:- We validated our SV identification pipeline using trio datasets. >99% of the SVs identified in the offspring were inherited from either of the parents, in line with expected Mendelian inheritance. Using this benchmarked pipeline, SVs were identified in a population-specific manner. We see improved SV detection using long read data: we identified ~12000 SVs per individual with short reads and ~27000 SVs per individual using long reads. We observe an increase in the number of SVs identified per individual when using CHM13 T2T genome as the reference as compared to GRCh38.

Conclusions:- Our study provides a starting point in maximizing the variant identification efforts in Indian populations and additionally proposing a strategy for improved structural variation identification. We also showcase improved SV detection in genomic blindspots such as repetitive regions using the Telomere-to-Telomere (T2T) reference genome.

Keywords:- Structural variation, population genomics, variation, genome