Pangenome-based genome inference

Jana Ebler, Heinrich-Heine University, Düsseldorf

Typical analysis workflows map reads to a reference genome in order to genotype genetic variants. Generating such alignments introduces reference biases and comes with substantial computational burden. In contrast, recent k-mer based genotypers are fast, but struggle in repetitive or duplicated genomic regions. We introduced a new algorithm, PanGenie, that leverages a haplotype-resolved pangenome reference in conjunction with k-mer counts from short-read sequencing data to genotype a wide spectrum of genetic variation – a process we refer to as genome inference. We could demonstrate that our method produces better results compared to mapping-based approaches. Improvements are especially pronounced for structural variants (SVs) and variants in repetitive regions. We studied SVs across large cohorts sequenced with short-reads, using pangenome graphs generated by the HGSVC and HPRC consortia, which enables the inclusion of these classes of variants in genome-wide association studies.