

As a bioinformatics facility, we have been implementing data mining pipelines for next-generation sequencing projects including ChIP-seq, allele-specific gene expression profiling, SNP identification, and de novo EST and genome sequence assembly. The challenges for processing the next generation sequencing data include both developing software and building hardware infrastructure. We have applied tools for short reads sequencing analysis including ELAND, MAQ, BLAST, Mosaik and Newbler for sequence alignment and sequence assembly. We used Mosaik and various ace file viewers for manual inspection in regions of interest. The SNP detection pipeline was used to identify polymorphism in a region of the canine genome among multiple dog breeds. The results we obtained match the canine SNP data in the dbSNP. The SNP identification results on other species will also be presented. The Solexa sequencing was used in a ChIP-seq study on tumor samples, and a comparison with ChIP-CHIP data will be presented. On the hardware side, we are using a combination of a parallel HPC cluster and a large memory server to implement these pipelines. BioHPC, a HPC bioinformatics software package developed by our group is used extensively in the data analysis.