DNA-seq

Whole genome sequencing data was analyzed for variants using GenPipes [PMID 31185495]. This pipeline follows the stepwise procedures of the BROAD Institute GATK best practices. Raw reads derived from the sequencing instrument are quality trimmed and adapter clipped using Trimmomatic [1] to obtain a high-quality set of reads for sequence alignment (sam/bam) file generation. The trimmed reads are aligned to a reference genome (typically build 37) using a fast, memory-efficient Burrows-Wheeler transform (BWT) aligner BWA-mem [2]. Mapped reads are further refined using GATK and Picard program suites [3,4] to improve mapping near insertions and deletions (indels; GATK indel realigner), remove duplicate reads with same paired start site (Picard mark duplicates) and improve quality scores (GATK base recalibration). Variants are called using GATK haplotype caller in gvcf mode to allow efficient downstream merging of multiple samples into one variant file to streamline downstream variant processing procedures which include normalization and decomposition of multi nucleotide polymorphisms (MNPs)[5], functional annotation with SNPeff [6] and variant annotations using the Gemini [7] framework which provides quality metric and extensive metadata to help further prioritize variants.

1. Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics*, btu170.
2. Li, H., Durbin, R. Fast and accurate long-read alignment with Burrows—Wheeler transform. Bioinformatics. 2010. 26, 589–595.
3. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–303 (2010).
4. Picard program suite: http://broadinstitute.github.io/picard
5. Tan A, Abecasis GR, Kang HM. 2015. Unified representation of genetic variants. Bioinformatics. 2015 Jul 1;31(13):2202-4
6. Cingolani, P.; Platts, A.; Wang le, L.; Coon, M.; Nguyen, T.; Land, S.J.; Lu, X.; Ruden, D.M. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 2012, 6, 80–92.
7. Paila, U., Chapman, B. A., Kirchner, R. & Quinlan, A. R. GEMINI: integrative exploration of genetic variation and genome annotations. PLoS Comput. Biol. 9, e1003153 (2013).