Data preprocessing

Low-quality and adapters sequences are trimmed from the raw sequencing reads using Trimmomatic (v. 0.39)[1]. Trimmed reads are then aligned to the human hg38 reference genome using BWA mapping software (v. 0.7.17)[2]. Duplicate reads are marked using Samblaster (v. 0.1.25)[3] and sorted using samtools (v. 1.8). Finally, base quality score recalibration is performed as indicated in the GATK4 (v. 4.2.2.0) best practices [4].

Germline variant calling

HaplotypeCaller from GATK4 (v. 4.2.2.0) is used to call germline variants, parallelized across chromosomes, and all samples in the cohort are joint genotyped together [4, 5].

Somatic variant calling

Somatic variant calling (SNPs and Indels) is performed using Mutect (v. 1.1.7)[6], Mutect2 (GATK v. 4.2.0)[7], Strelka2 (v. 2.9.0)[8], and VarDict (v. 1.4)[9] in tumor-normal mode. Variants from all callers are merged using the CombineVariants tool from GATK version 3.8-1. Genomic, functional and consequence annotations are added using Variant Effect Predictor (VEP v. 99) [10] and converted to Mutation Annotation Format (MAF) using the vcf2maf tool (v. 1.6.16)[11].

For Copy Number Variants (CNVs), Control-Freec (v. 11.6)[12] is used to generate pileups, which are used as input for the R package ‘*sequenza’* (v. 3.0.0)[13]. The complete Control-Freec workflow is then re-run using ploidy and cellularity estimates from *sequenza*.

Quality and identity metrics

Ancestry and relatedness scores are generated using Somalier (v. 0.2.13)[14]. Contamination analyses are performed against viral and bacterial genomes from NCBI using Kraken2 (v. 2.1.2)[15], as well as against mouse, human, and UniVec databases using FastQ Screen (v. 0.14.1)[16]. Sequence, mapping and variant statistics are computed using FastQC (v. 0.11.9), Qualimap (v. 2.2.1)[17] and SNPeff (v. 4.3t)[18]. All of these metrics are combined into an interactive HLTML report using MultiQC (v. 1.11)[19].

Pipeline Orchestration

Job execution and management is done using Snakemake (v. 6.8.2)[20] using custom-built Singularity (v. 3.8.5) containers for reproducibility.

1. Bolger, A.M., M. Lohse, and B. Usadel, *Trimmomatic: a flexible trimmer for Illumina sequence data.* Bioinformatics, 2014. **30**(15): p. 2114-20.

2. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-60.

3. Faust, G.G. and I.M. Hall, *SAMBLASTER: fast duplicate marking and structural variant read extraction.* Bioinformatics, 2014. **30**(17): p. 2503-5.

4. Van der Auwera, G.A. and B.D. O'Connor, *Genomics in the cloud : using Docker, GATK, and WDL in Terra*. First edition. ed. 2020, Sebastopol, CA: O'Reilly Media.

5. Poplin, R., et al., *Scaling accurate genetic variant discovery to tens of thousands of samples.* bioRxiv, 2018: p. 201178.

6. Cibulskis, K., et al., *Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples.* Nat Biotechnol, 2013. **31**(3): p. 213-9.

7. Benjamin, D., et al., *Calling Somatic SNVs and Indels with Mutect2.* bioRxiv, 2019: p. 861054.

8. Kim, S., et al., *Strelka2: fast and accurate calling of germline and somatic variants.* Nat Methods, 2018. **15**(8): p. 591-594.

9. Lai, Z., et al., *VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research.* Nucleic Acids Res, 2016. **44**(11): p. e108.

10. McLaren, W., et al., *The Ensembl Variant Effect Predictor.* Genome Biol, 2016. **17**(1): p. 122.

11. Memorial Sloan Kettering Cancer Center. *vcf2maf*. 2013 [cited 2019; Available from: <https://github.com/mskcc/vcf2maf>.

12. Boeva, V., et al., *Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data.* Bioinformatics, 2012. **28**(3): p. 423-5.

13. Favero, F., et al., *Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data.* Ann Oncol, 2015. **26**(1): p. 64-70.

14. Pedersen, B. *somalier: extract informative sites, evaluate relatedness, and perform quality-control on BAM/CRAM/BCF/VCF/GVCF*. 2018 [cited 2020; Available from: <https://github.com/brentp/somalier>.

15. Wood, D.E., J. Lu, and B. Langmead, *Improved metagenomic analysis with Kraken 2.* Genome Biol, 2019. **20**(1): p. 257.

16. Wingett, S.W. and S. Andrews, *FastQ Screen: A tool for multi-genome mapping and quality control.* F1000Res, 2018. **7**: p. 1338.

17. Okonechnikov, K., A. Conesa, and F. Garcia-Alcalde, *Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data.* Bioinformatics, 2016. **32**(2): p. 292-4.

18. Cingolani, P., et al., *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**(2): p. 80-92.

19. Ewels, P., et al., *MultiQC: summarize analysis results for multiple tools and samples in a single report.* Bioinformatics, 2016. **32**(19): p. 3047-8.

20. Koster, J. and S. Rahmann, *Snakemake-a scalable bioinformatics workflow engine.* Bioinformatics, 2018. **34**(20): p. 3600.