Mouse Variant Pipeline

Main Pipeline

Variant calling was performed using the nf-core/sarek pipeline [Garcia2020, Ewels2020], supplemented with additional custom scripts available on GitHub [MusVar2024]. The sarek pipeline was executed in somatic mode, utilizing the following callers: Mutect2, Strelka, and FreeBayes. Version 3.4.0 was used; however, genomic resources were updated and modified as detailed below.

Upon completion of the standard Sarek workflow, an additional set of calls were computed using the VarDict caller, version 1.8.3 [VarDict2021]. The outputs from Sarek and VarDict were then post-processed to merge the calls from the various algorithms into a single table. Since the FreeBayes method does not mark the standard VCF FILTER field, we chose to mark FreeBayes events as passed if their VCF QUAL score was > 15,000. The raw, unfiltered set of calls was output in the standard MAF format, and the pipeline then generated filtered versions in a separate file. All tables were filtered using the following rules:

- `Variant Allele Frequency (VAF) >= 0.05`
- `Mutation Allele Depth >= 8`
- `Total Depth >= 20`
- `Tumor VAF > 5* Normal VAF`
- `Non-silent events only`

Additionally, two filter sets were used. For a list of high-confidence events, mutations had to have been called with `FILTER == PASS` in at least two of the four callers. A second table with a high sensitivity level was also generated, which only required one caller with a PASS filter.

Genomic Resources

Most of the reference files were sourced from the included igenomes, with the following modifications: The main genome FASTA file was not sourced from igenomes but rather the standard GRCm38 reference file from Ensembl (Release 68, patch level 0). The contigs were reordered to standard ordering, as detailed in the genome dictionary file available in our GitHub repository, and the index was rebuilt using BWA version 7. A germline reference file, used by Mutect2, was created by reformatting and merging variant files from the Mouse Genome Project available through the igenomes Ensembl resource. Finally, a panel of normals reference file was also created. These custom VCF files are available at https://zenodo.org/records/10914483.

References

Garcia M, Juhos S, Larsson M, et al. Sarek: A portable workflow for whole-genome sequencing analysis of germline and somatic variants. F1000Research 2020, 9:63 (https://doi.org/10.12688/f1000research.16665.2)

Ewels, PA, Peltzer, A, Fillinger, S, et al. The nf-core framework for community-curated bioinformatics pipelines. Nat Biotechnol 38, 276–278 (2020). (https://doi.org/10.1038/s41587–020–0439-x)

MusVar 2024, version 0.7.1 (MusVar-v0.7.2)

VarDict 2021, version 1.8.3 (vardict-v1.8.3)