BIOINFORMATICS ANALYSIS OF WHOLE EXOME SEQUENCING DATA

Peter J. Ulintz, Weisheng Wu, and Chris M. Gates

Overview

This research paper provides a detailed protocol for analyzing next-generation sequencing data from tumor and matching normal samples to identify somatic single nucleotide polymorphisms (SNPs) and small insertions/deletions (Indels). The protocol is based on the Broad Institute's "Best Practices" guidelines and utilizes their Genome Analysis Toolkit (GATK) platform. Variants are annotated with population allele frequencies and information from curated resources like GnomAD and ClinVar using VCFtools, SnpEff, and SnpSift. The chapter highlights the importance of whole exome sequencing (WES) for cost-effective analysis of protein-coding regions. The Mutect2 somatic variant caller is employed, with instructions provided for setting up the required files and software. Alternative workflows using VarScan Somatic are also discussed. The chapter concludes by acknowledging the contributions of various informatics tools and public resources in genomics research. Overall, this chapter offers a comprehensive guide for the bioinformatics analysis of WES data for somatic variant detection in cancer research.

Introduction

Whole Exome Sequencing is a genomic technique for sequencing all of the protein-coding regions of genes in a genome. It utilizes a set of oligonucleotide hybridization probes that target known exon sequences. Applications of WES include; 1) Somatic variant detection, 2) Characterization of new therapeutic targets, 3) Profiling of copy-number variations (CNVs) and the detection of structural variations, 4) Mutational analysis: the detection of single-nucleotide variants (SNVs) or small insertions and deletions (Indels). Somatic Variant Detection is performed using algorithms and software tools specialized for the task. It can classify a variant in a cancer sample as either germline or somatic with a second measure of likelihood. Mutect2 somatic variant caller workflow used, largely following the Broad GATK4 Somatic SNVs + Indels Best Practices workflow. There is also a supplementary workflow based on a second popular caller: VarScan Somatic. This workflow is demonstrative of a class of variant callers that do not perform local haplotype assembly, thus benefitting from a specific GATKbased step known as Indel Realignment which would be superfluous in the main workflow.

Methodology

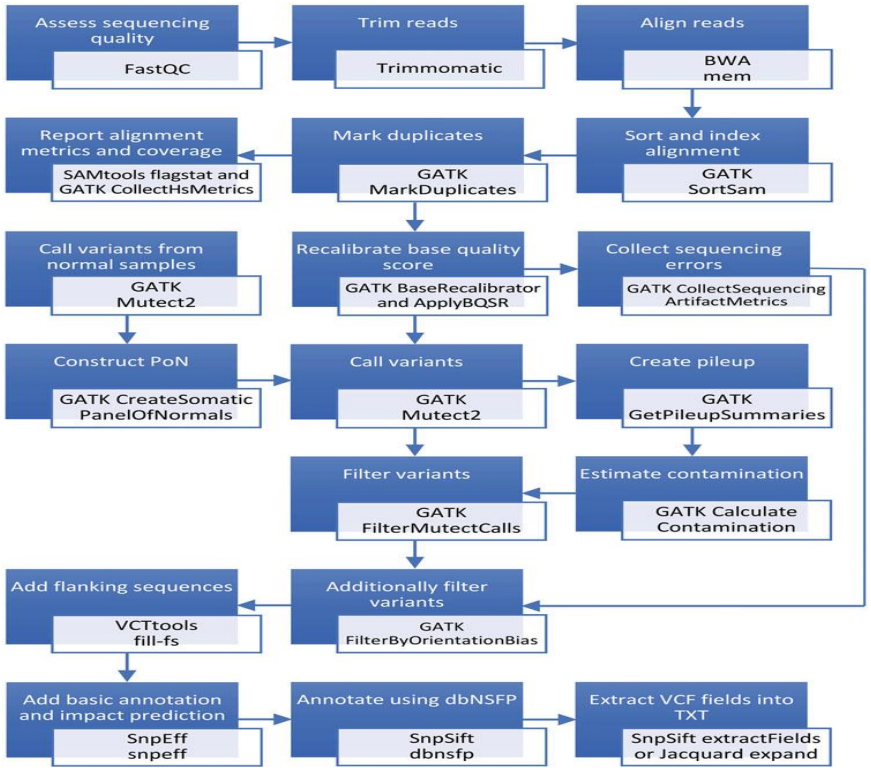

Fig. 1 Workflow of somatic variant identification and annotation using GATK4 and Mutect2

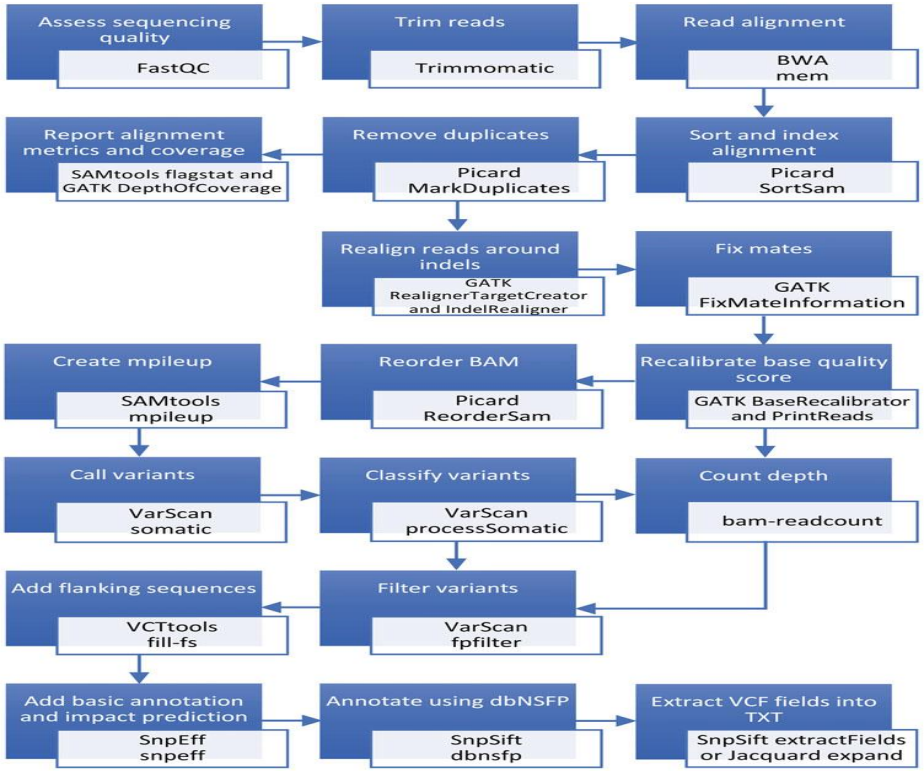

Fig. 2 Workflow of somatic variant identification and annotation using GATK3 and VarScan (Alternative)