BITS :: Call for Abstracts 2022 - Oral communication

Type	Oral communication
Session	Algorithms for Bioinformatics
Title	A novel affordable and reliable framework for accurate detection and comprehensive analysis of somatic mutations in cancer.
All Authors	Atzeni R(1), Massidda M(1) , Pieroni E(1), Pisu M(1), Fotia(1)

Affiliation

(1) Center for Advanced Studies, Research and Development in Sardinia (CRS4), Pula, Italy

Motivation

Cancer is a disease of the genome which is determined by mutations in somatic cells, with a number of mutations ranging ten to millions during disease evolution.
Accurate detection, key features identification and comprehensive analysis of somatic mutations are thus the major tasks in processing cancer sample data, now widely available thanks to the growing abilities of Next Generation Sequencing (NGS) to parallel interrogate multiple cancer genomes.
Somatic variant calling represents the baseline for many heterogeneous downstream analyses which are routinely carried out by combining many third-party software with complex dependency trees and configuration requirements, as well as laborious, error-prone and time-consuming data format conversions. This classic approach definitely fails in terms of accuracy, reproducibility and portability of the analysis workflow. The lack of standardization of data analysis procedures limits the full exploitation of results in clinical practice.
In order to overcome these limitations, we developed MUSTA, a scalable and flexible framework.

Methods

Musta is based on a Python command-line tool that easily handles matched tumor-normal or tumor-only samples for accurate detection and comprehensive analysis of somatic mutations. This user-friendly approach allows researchers, without specific computer programming experience, to process cancer data by using a single command line.
The core is a Snakemake based workflow that contains all analysis steps commonly performed in cancer genomics, following GATK Best Practices Somatic Pipeline and exploiting mafTools R package. In details, Musta is thus able to perform in an integrated way the following tasks:
- Variant Calling. Calls somatic SNVs and indels. Users can choose between two modes: (i) tumor-normal mode, where a tumor sample is matched with a normal sample in analysis and (ii) tumor-only mode, where a single sample's alignment data is analyzed.
- Variant Annotation. Functional annotation of called somatic variants based on a set of data sources, each with its own matching criteria.
- Driver Gene Detection. Identification of cancer driver genes based on positional clustering. The output contains the list of genes ordered according to their p-values and a weighted scatter plot.
- Pathway Analysis. Check for enrichment of known oncogenic pathways. The output contains a fraction of the affected pathway and samples and an oncoplot of the oncogenic pathway.
- Estimation of Tumor Heterogeneity. Inferring tumor clonality by clustering variant allele frequencies. The output contains clustering results and the related density plot.
- Deconvolution of Mutational Signatures. De-novo extraction of mutational signatures followed by a comparison with the reference COSMIC signatures by means of similarity score. The output contains deconvoluted signatures, cosine similarities, aetiologies, best match and a barplot of decomposed mutational signatures.

The first step accepts as input a BAM file, the second one a VCF file and all other steps a MAF file. The workflow contains separate rules for each step. Each rule with additional dependencies has a separate Conda environment that will be automatically created by running the workflow for the first time. Thus, Snakemake ensures reproducibility and scalability seamlessly to different computing environments, while Conda offers version control of the utilized programs, leading to a simple installation without the risk of dependency conflicts.
Musta is conceived for an easy installation on any computational platform and any operating system through the Docker platform. A simple Makefile bootstraps Musta, taking care of the installation, configuration and running steps and allowing the execution of the entire pipeline or any individual step depending on the starting data.

Results

Musta is currently used for cancer sample data analysis at the CRS4-NGS Core.
Its reliability has been extensively tested on published data from “The Cancer Genome Atlas” repository, particularly for head and neck, breast cancers, lung adenocarcinoma. Each cohort counts hundreds of samples in distinct patients. Furthermore, Musta was also tested on published data from the genome archive of Beijing Institute of Genomics, with 23 samples of liver cancer from the same patient.
Musta is easy to install and bootstrap by users without specific skills in computational programming and results can be fully reproduced by any other user.
Musta, both in tests and routine analysis, has proven to be a robust and flexible framework for accurate detection and comprehensive analysis of somatic variants in cancer and is freely available by contacting the authors and on GitHub.

Info

Atzeni R. performed his activity in the framework of the International PhD in Innovation Sciences and Technologies at the University of Cagliari, Italy.	
filename	-
Figure	
-	
Availability	https://github.com/next-crs4/musta
Corresponding Author	
Name, Surname	Atzeni, Rossano
Email	ratzeni@crs4.it
Submitted on	29.04.2022