The KAPA Target Enrichment Bioinformatics Container Offers a Streamlined and Portable Solution for Germline Variant Analysis

Shobana Sekar, Jieqiong Dai, Nikita Raymond Dsouza, and Alejandro Quiroz Zarate
Roche Diagnostics Corporation, Wilmington, MA

INTRODUCTION

Germline variants, inherited at the time of conception, are not only associated with several genetic diseases, but can also contribute to cancer risk and tumor progression.[1,2] Targeted next generation sequencing (NGS) enables high-throughput and cost-effective analysis of such germline variants.[3] Roche's KAPA target enrichment (TE) workflows (KAPA HyperCap and KAPA HyperPETE,* **Figure 1**) offer an end-to-end, comprehensive solution for targeted germline variant calling. These TE workflows enable sensitive and precise detection of germline single nucleotide variants (SNVs) and small insertions/deletions (indels), achieving a true positive rate of over 97% when tested on reference cell line samples.[4]

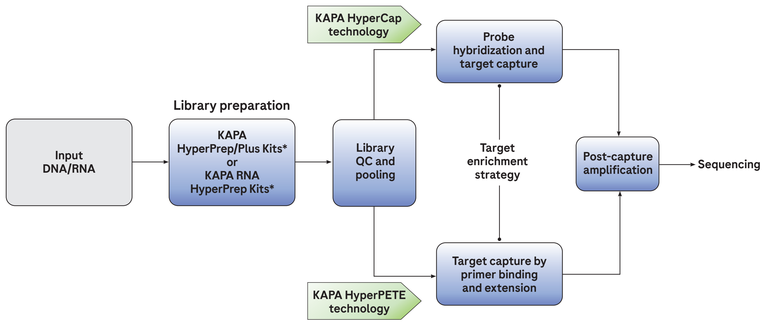

Figure 1: Two different workflows for KAPA Target Enrichment. In this demonstration of the software container, the KAPA HyperPETE workflow, using KAPA HyperPlus Kit, was used.

BIOINFORMATICS CONTAINER FOR KAPA TARGET ENRICHMENT ANALYSIS

Several open-source bioinformatics tools are available for the analysis of NGS TE data. This creates a major challenge in the creation and implementation of reproducible analysis pipelines, since software installation and setup of the runtime environment can be time-and-resource consuming, and can also be variable between laboratories.

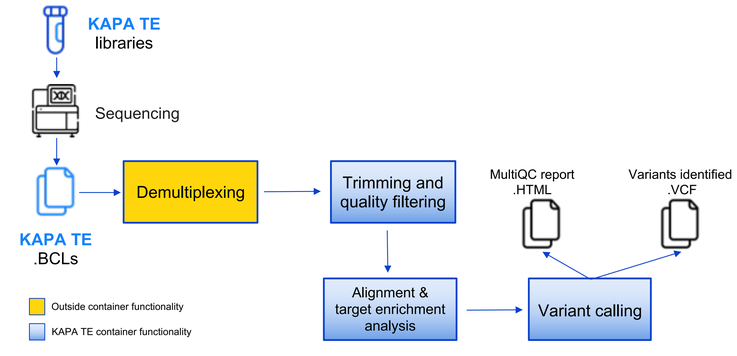

Figure 2: Overview of KAPA TE bioinformatics container implementation.

A container-based solution has been created using Singularity,[5] providing a consistent computational environment and a streamlined path for the analysis of KAPA TE germline data (**Figure 2**). This bioinformatics container encapsulates all the major analysis steps of the pipeline shown in (**Figure 3**).

ANALYSIS WORKFLOW

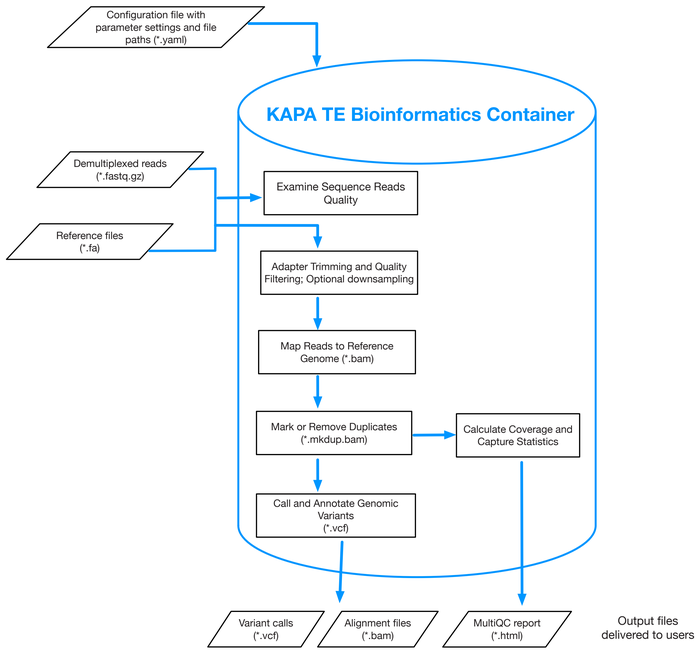

Figure 3: Bioinformatics analysis workflow to detect germline variants from KAPA TE data. A configuration file with parameter settings and file paths has to be provided as input to the container.

QC AND TARGET ENRICHMENT METRICS

	Hereditary_ Oncology_1plex	Hereditary_ Oncology_8plex	Newborn_ screening_1plex	Newborn_ screening_8plex
Total reads	900,000	900,000	900,000	800,000
Duplication rate	2.90%	10.50%	4.90%	10.20%
Percentage aligned	100%	100%	100%	100%
Error rate	0.24%	0.24%	0.19%	0.21%
Artifact rate	0.00815	0.00847	0.00867	0.00833
Fraction of targets with at least 20X coverage depth	99.80%	99.97%	96.64%	96.21%
Fraction of targets with at least 30X coverage depth	99.61%	99.70%	95.64%	95.12%
Fold-80 base penalty	1.5	1.5	1.6	1.6
Percentage of reads on target	75.80%	77.90%	71.10%	73.10%

Sequencing data from NA12878 libraries (n=4) prepared using two different KAPA HyperPETE panels (Hereditary Oncology Panel - 203 kb, covering 47 genes, and Newborn Screening Panel - 294 kb, covering 89 genes, singleplex and multiplexed) were sequenced (2 X 151 bp) on the Illumina® NextSeq platform. Raw sequencing data was then demultiplexed and run through the KAPA TE bioinformatics container to demonstrate its functionality.

VARIANT CALLING RESULTS

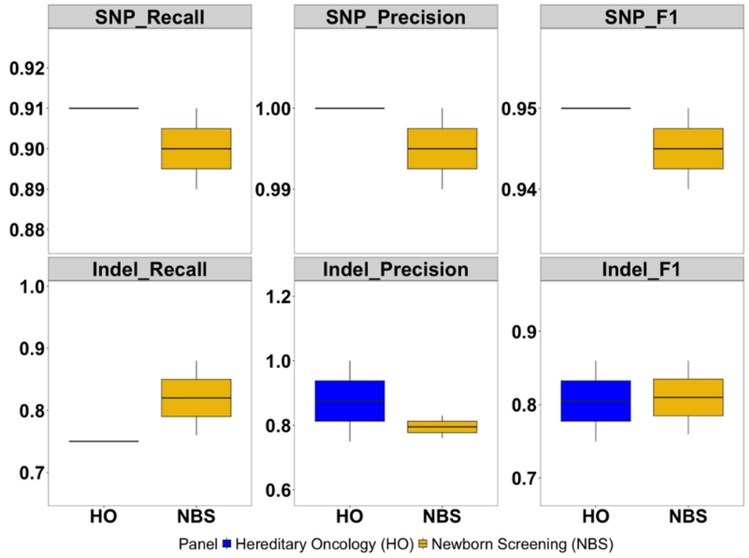

Figure 4: KAPA TE bioinformatics container shows reliable performance in germline variant calling. This evaluation used benchmark variants within the target regions of the respective panels. Recall (True positive rate) = FP/(FP+TN), Precision (Positive predictive value) = TP/(TP+FP), F1=2× (Precision+Recall) / (Precision×Recall).

CONCLUSIONS

Here, a container-based solution has been implemented to streamline the bioinformatics analysis of KAPA target enrichment data for germline variant detection. The container includes all the major analysis steps of the pipeline, starting with sequence quality assessment, trimming and subsampling as necessary, and then followed by alignment, TE metrics generation, and germline variant calling.

By leveraging the KAPA TE bioinformatics container, users with limited bioinformatics resources can swiftly evaluate the efficacy of KAPA TE offerings, thereby delivering a user-friendly bioinformatics solution.

REFERENCES

1. Jackson, M., Marks, L., May, G. H., & Wilson, J. B. (2018). The genetic basis of disease. *Essays in biochemistry, 62*(5), 643-723.
2. Milanese, J. S., & Wang, E. (2019). Germline mutations and their clinical applications in cancer. *Breast Cancer Management, 8*(1), BMT23.
3. Di Resta, C., Galbiati, S., Carrera, P., & Ferrari, M. (2018). Next-generation sequencing approach for the diagnosis of human diseases: open challenges and new opportunities. Ejifcc, 29(1), 4.
4. The novel, end-to-end KAPA HyperPETE Target Enrichment Workflow enables high-performance germline variant analysis. Roche Sequencing Solutions, Inc. (2022)
5. Kurtzer, G. M., Sochat, V., & Bauer, M. W. (2017). Singularity: Scientific containers for mobility of compute. *PloS one, 12*(5), e0177459.

Data on file.
Project: KAPA TE Germline Container
*For Research Use Only. Not for use in diagnostic procedures.

MC-US-15328 07/24