

Cancer Analysis Workflow to process normal/tumor WGS data

Maxime Garcia¹, Szilveszter Juhas², Malin Larsson³, Teresita Diaz de Ståhl⁴, Johanna Sandgren⁴, Jesper Eisfeldt⁵, Sebastian DiLorenzo⁶, Marcel Martin⁷, Pall Olason⁸, Björn Nystedt⁸, Monica Nistér⁴, Max Käller⁹

- 1 - BarnTumörBanken, Dept. of Oncology Pathology, Science for Life Laboratory, Karolinska Institutet
- 2 - Dept. of Biochemistry and Biophysics, Science for Life Laboratory, Stockholm University
- 3 - Dept. of Physics, Chemistry and Biology, National Bioinformatics Infrastructure Sweden, Science for Life Laboratory, Linköping University
- 4 - BarnTumörBanken, Dept. of Oncology Pathology, Karolinska Institutet
- 5 - Clinical Genetics, Dept. of Molecular Medicine and Surgery, Karolinska Institutet

- 6 - Dept. of Medical Sciences, National Bioinformatics Infrastructure Sweden, Science for Life Laboratory, Uppsala University
- 7 - Dept. of Biochemistry and Biophysics, National Bioinformatics Infrastructure Sweden, Science for Life Laboratory, Stockholm University
- 8 - Dept. of Cell and Molecular Biology, National Bioinformatics Infrastructure Sweden, Science for Life Laboratory, Uppsala University
- 9 - Science for Life Laboratory, School of Biotechnology, Division of Gene Technology, Royal Institute of Technology



WGS normal/tumor pairs analysis workflow written in



Follows GATK best practices. Provides SNVs, small indels, structural variants, heterogeneity, ploidy CNVs, annotation and QC reports

- Tools used:
- ASCAT
 - Freebayes
 - HaplotypeCaller
 - Manta
 - MultiQC
 - MuTect1
 - MuTect2
 - snpEff
 - Strelka
 - VEP

Can also analyse normal only samples

Easily deployable, supports containers



Can be used on



Open source, contribute on github



Join the chat on gitter



Aknowledgements



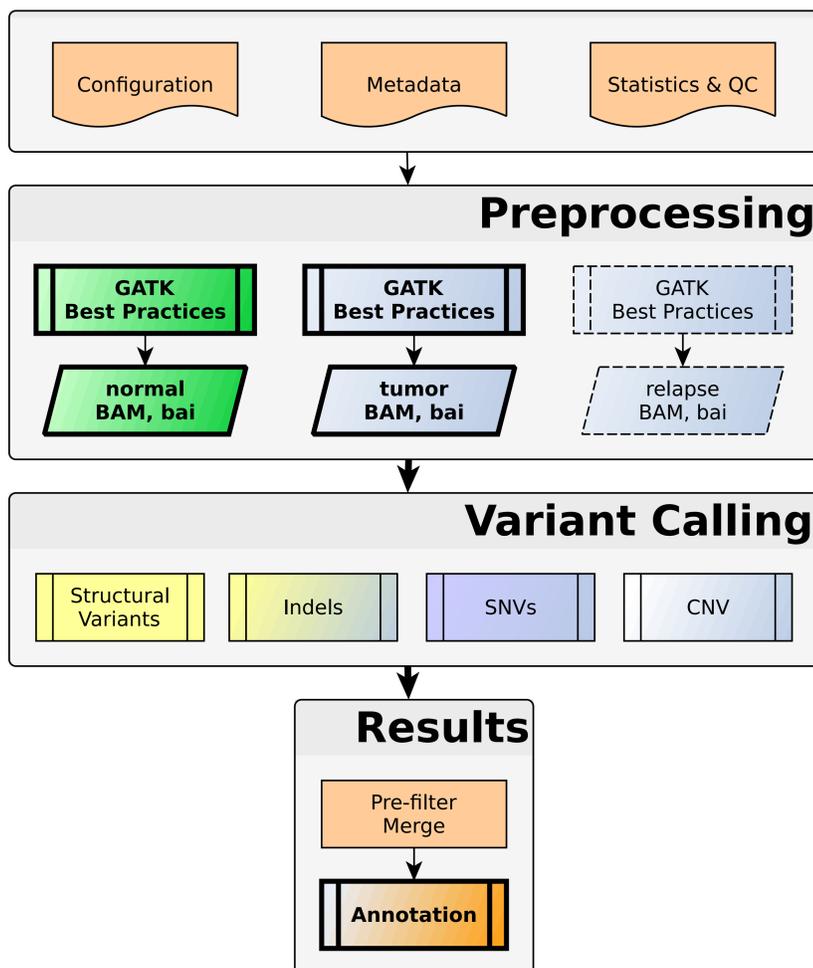
UPPMAX

The MIT License (MIT)
Copyright © 2016
SciLifeLab

Abstract

As WGS (whole genome sequencing), the broadly used research tool is getting cheaper and being introduced to clinics, it is now possible to compare data from normal and tumor samples of numerous patients. There are still many challenges, mostly regarding bioinformatics: datasets are huge, workflows are complex, and there are multiple tools to choose from for somatic and structural variant detection and quality control.

We are presenting CAW (Cancer Analysis Workflow), a complete open source pipeline to resolve somatic variants from WGS data: it is written in Nextflow*, a domain specific language for workflow building.



Links

- <http://opensource.scilifelab.se/projects/caw>
- <https://github.com/SciLifeLab/CAW>
- <https://gitter.im/SciLifeLab/CAW>
- <https://www.scilifelab.se/facilities/genomics-applications/>

References

- * : <http://dx.doi.org/10.1038/nbt.3820>

Core principles

We are utilizing GATK best practices to align, realign and recalibrate short-read data in parallel for both tumor and normal samples. After preprocessing, several somatic variant callers scan the resulting BAM files; MuTect1, MuTect2 and Strelka are used to find somatic SNVs and small indels. For structural variants we use Manta. Furthermore, we are applying ASCAT to estimate sample heterogeneity, ploidy and CNVs. The workflow also provides quality controls presented by MultiQC.

CAW can start the analysis from raw FASTQ files, from the realignment step, or directly with any subset of variant callers. Resulting VCF files can be annotated using snpEff or VEP. At the end of the analysis the final VCF files are merged to facilitate further downstream processing, though the individual results are also retained. The flow is capable of accommodating further variant calling software or CNV callers.

Workflow specificities

CAW is prepared to process normal/tumor pairs and can handle additional relapse samples. It can also be used to preprocess and analyse normal only samples using GATK best practices and HaplotypeCaller.

Preprocessing takes around 4 days on one Huawei XH620 V3 compute node with dual CPUs (Intel Xeon E5-2630 v3, each with 8 cores) with 45 coverage normal/tumor pairs. Further Variant Calling can take up to 10 hours for each variant caller.

The pipeline can use GRCh37 or GRCh38 as a reference genome. Docker containers are also available for easier deployment and testing purposes. CAW can be downloaded from our GitHub repository.

Aknowledgements

The authors thank the Swedish Childhood Cancer Foundation for the funding of BarnTumörbanken. We would like to acknowledge support from Science for Life Laboratory, the National Genomics Infrastructure, NGI, and UPPMAX for providing assistance in massive parallel sequencing and computational infrastructure.