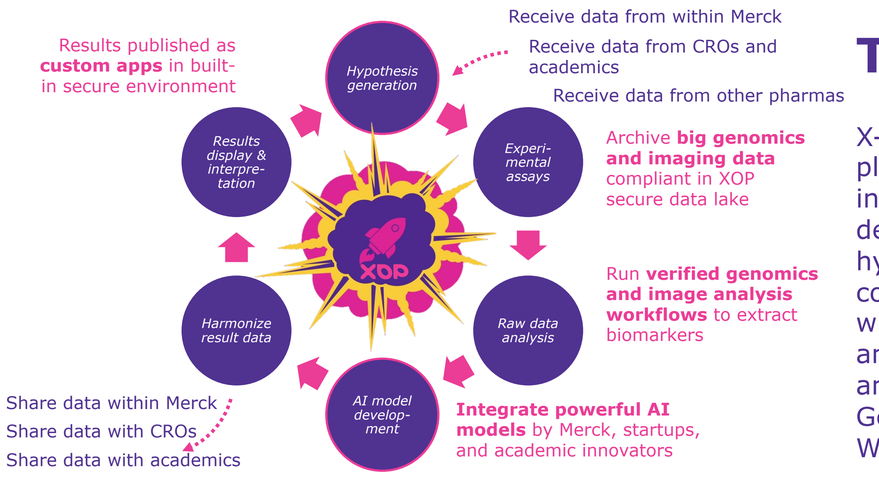
FOSS-based best-practice genomics workflows, GxP-compliant data analysis, and open science at Merck

Sven-Eric Schelhorn[1], Anna Coenen-Stass[1], Dmitriy Drichel[2], Thomas Grombacher[1], Stefan Pinkert[1], Alex Rolfe[3], Jing Yang[1], Sheng Zhao[1]

[1] Merck Healthcare KGaA, Germany • [2] Drichel Analytics, Germany • [3] EMD Serono Inc., USA

Introduction

Large biopharmaceutical organizations widely use free and open-source software (FOSS)-based genomics workflows. However, sharing practical approaches for analyzing genomics data from pre-clinical disease models and human patients is not common in the industry community, which counteracts open science principles and makes improvement of specialized workflows difficult. Here, we present a FOSS-based, best-practice genomics workflow developed at Merck for clinical use that promotes FAIR principles and GxP/GDPR regulatory compliance as well as helped identify a high-impact regression in `Mutect2`.



The Merck X-Omics platform

X-Omics is an AI-driven data management and analytics platform which supports the whole R&D innovation cycle including raw data analysis, data harmonization, AI model development, result visualization and interpretation, and hypothesis generation. This setup facilitates data durability, compute elasticity, cost efficiency, and continuous validation while being easy to use by external collaborators. It is based an Amazon AWS components such as Redshift Spectrum, S3, and EC2 Autoscaling Groups in conjunction with the Genedata Profiler FAIR Data Lake, Podman, Gitlab, and Posit Workbench, Package Manager, and Connect (for Shiny apps).

Verified clinical genomics workflows at Merck

Context

Merck provides several high-quality reference workflows in our cloud-based X-Omics system. These reference workflows enable both pre-clinical and clinical data analyses under full GxP and GDPR compliance and consist of a range of open-source activities that are extensible via `podman` containers. Besides managing data governance, elastic cluster computing, and long-term reproducibility via globally immutable file identifiers, the system also includes automatic data provenance as each results data item also remembers the full (re-)executable workflow that produced it.

Read mapping and quality control

1 After import of `FASTQ`, `BAM` or `CRAM` files from the cloud-based data lake, read quality control and adapter trimming are conducted prior to read mapping. For WGS, reads are mapped in chunks so that mapping of single samples can be parallelized across the cluster.

2 For the analysis of patient-derived xenograft models, reads are mapped in parallel against both human and mouse references. Only read pairs that map better to the human reference than to the mouse reference are retained to remove host contaminants.

Variant calling & annotation

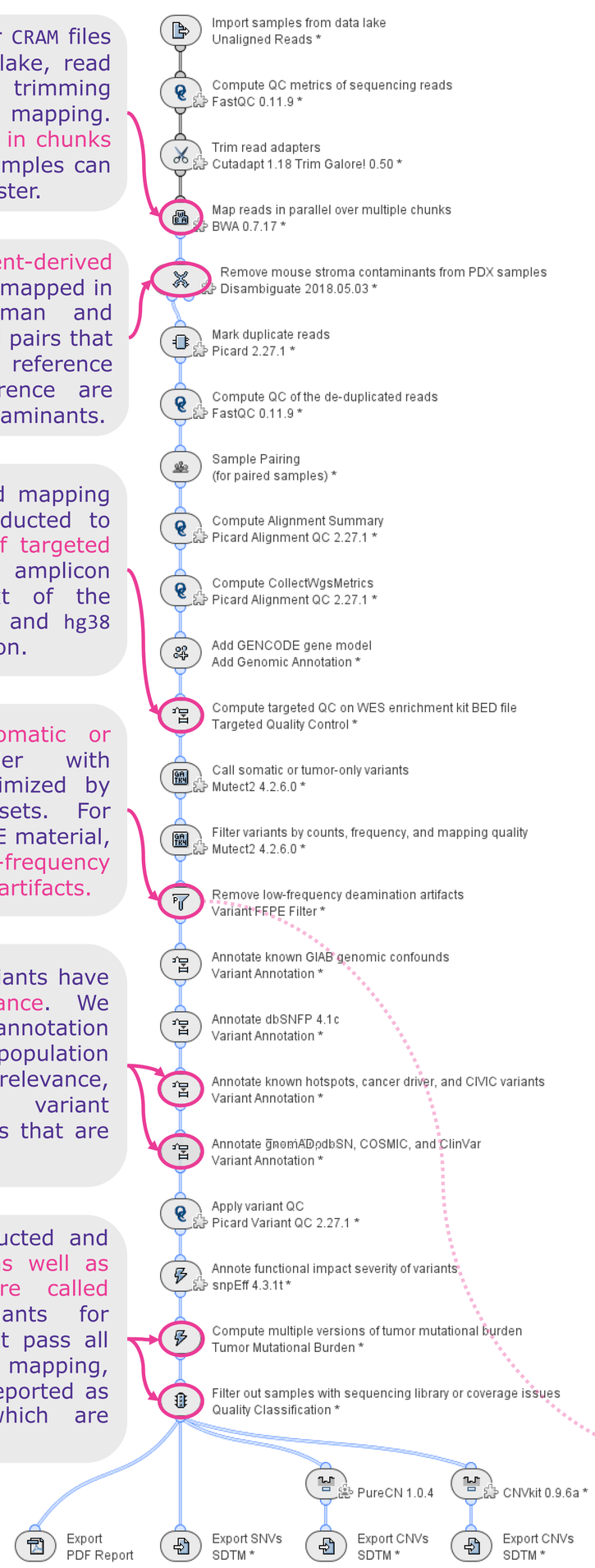
3 Prior to variant calling, read mapping enrichment analysis is conducted to evaluate the performance of targeted assays such as WES and amplicon sequencing in the context of the sequencing library BED file and `hg38 GENCODE 33` genome annotation.

4 We utilize Mutect2 as somatic or tumor-only variant caller with parameters and filters optimized by public benchmark data sets. For samples processed from FFPE material, we additionally filter out low-frequency deamination & DNA damage artifacts.

5 In oncology, most small variants have unknown functional relevance. We therefore employ several annotation sources to annotate population frequencies, clinical relevance, computationally predicted variant effects, and genomic regions that are functionally blacklisted.

Results verification

6 Finally, variant QC is conducted and tumor mutational burden as well as copy number variants are called (which used small variants for phasing). Only samples that pass all QC steps (read, library, mapping, variant) pass. Results are reported as SDTM-like `tsv.gz` files which are converted to `parquet` files.



Usecase: identifying a regression in Mutect2

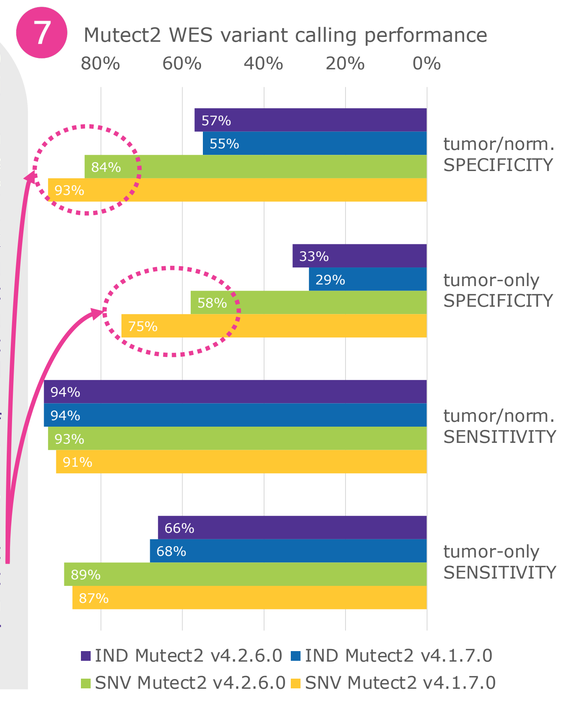
Impact

The somatic variant caller Mutect2 developed by the Broad Institute's GATK team is one of the most popular tools for DNA-Seq analysis currently in use. A Scopus search in Q2/2023 revealed 260 academic papers referencing Mutect2 while Google scholar reported 4.390 documents mentioning the tool. Due to its focus on cancer patient samples and its good performance in multiple benchmarks, Mutect2 is broadly used for clinical analyses: the Broad Institute alone reported in Q4/2022 that it had sequenced two million human samples; conservatively estimated, it is likely that at least tens of thousands of these were human WES tumor samples analyzed by Mutect2. In addition, many other institutions and consortia such as the German DKFZ-ODCF, ICGC and TCGA utilize Mutect2 as part of their clinical cancer sequencing pipelines.

Benchmark and bug detection

An issue in Mutect2 became apparent after we updated our benchmark-verified Mutect2 v4.1.7.0 to v4.2.6.0, motivated by the discovery of log4j vulnerability in v4.1.7.0.
The SEQC2 benchmark we employ is based on HCC1395 tumor cell line, sequenced at 7 centers [see *Somatic Mutation Working Group of the SEQC2 Consortium, Nature Biotechnology (2021)*, `PMID:34504347`]. Comprising ~40k SNVs and ~2k INDELs in ~2.4Gb high-confidence regions, and using a combined 1,500x WGS coverage for variant discovery and multiple orthogonal sequencing technologies for validation, it represents the most comprehensive and reliable somatic benchmark call set available to date.
In this scenario, we utilized the WES validation set of the SEQ2 benchmark generated at the Fudan University (~170x coverage for both tumor and normal tissues). After benchmarking the workflow using the more recent version of Mutect2 and comparing to our previous results, we found that SNV calling specificity dropped by 9-17% (indels not affected). The drop in specificity corresponds to a significant increase of false positive SNPs by a factor of 2.3 in tumor-only and 2.7 in tumor/normal calls.

7



Technical root cause analysis

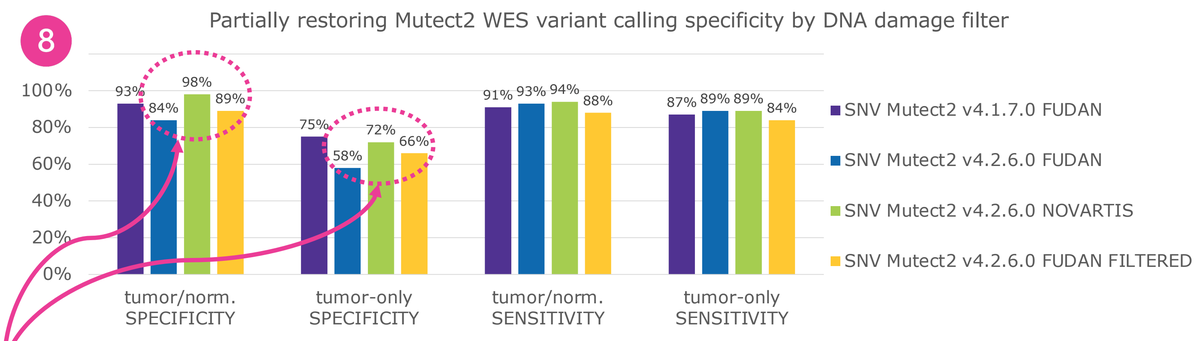
Upon identification of the issue in June 2022, our team created a bug report in the GATK GitHub repository (issue `#7921`, `tinyurl.com/mutect2`) with a detailed description of the problem affording full reproducibility. Several other users confirmed the regression and it was acknowledged and flagged as high-priority by the GATK team.

As the GATK team was not able to provide support in the foreseeable future due to time constraints, we conducted a Git bisection analysis and identified the line `private static final int ONE_THIRD_QUAL_CORRECTION = 5;` in commit `a304725` as the culprit. The commit was introduced in October 2020 as part of a correction to the Mutect2 quality score model that slightly increased sensitivity on the DREAM3 synthetic benchmark. We confirmed that *disabling* the quality score model correction *restored* specificity in the more recent and, as we argue, superior, SEQC2 HCC1395 benchmark.

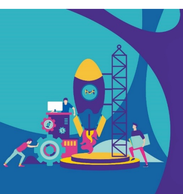
Next, we aimed to understand better why certain WES variants seem to be falsely included by Mutect2 given the corrected quality score model. Luckily, we could identify additional experimental replicates of the SEQC2 HCC1395 somatic variant benchmark samples that underwent sample preparation at Novartis instead of at Fudan.

Effects of sample preparation

8



According to two 2021 Science and Nature Biotechnology publications (`PMID:28209900`, `PMID:34504346`), the Fudan sample exhibits higher DNA damage (as indicated by strongly increased `G>T/C>A` ratios) compared to the Novartis samples, independent of library type. Indeed, filtering low-frequency variants associated with deamination of cytosine bases, a common form of DNA damage occurring during sample storage and preparation, partially restored specificity selectively in the Mutect2 version affected by the regression (see 8). While not fully restoring specificity to values prior to the regression, we believe this indicates that changes to the Mutect2 quality model may inadvertently have made it overly sensitive towards artefactual variant calls resulting from DNA-damaged samples.



Here, we have demonstrated how open science principles of reproducibility, provenance, and benchmarking based on public data can increase the quality of FOSS-based data analyses performed on human clinical samples. In particular, we could demonstrate how running benchmarks between version changes can identify regressions, and how pharmaceutical companies can contribute back to the FOSS genomics community by contributing best practice workflows and issue reports. Going forward, we recommend establishment of automatic regression tests in FOSS tools used to analyze clinical data. Ideally, such tests should include multiple real-world benchmarks that take clinical realities such as DNA damage due to sample degradation into account.

INNOVATIVE PRACTICES AWARDS WINNER BioIT World 2022