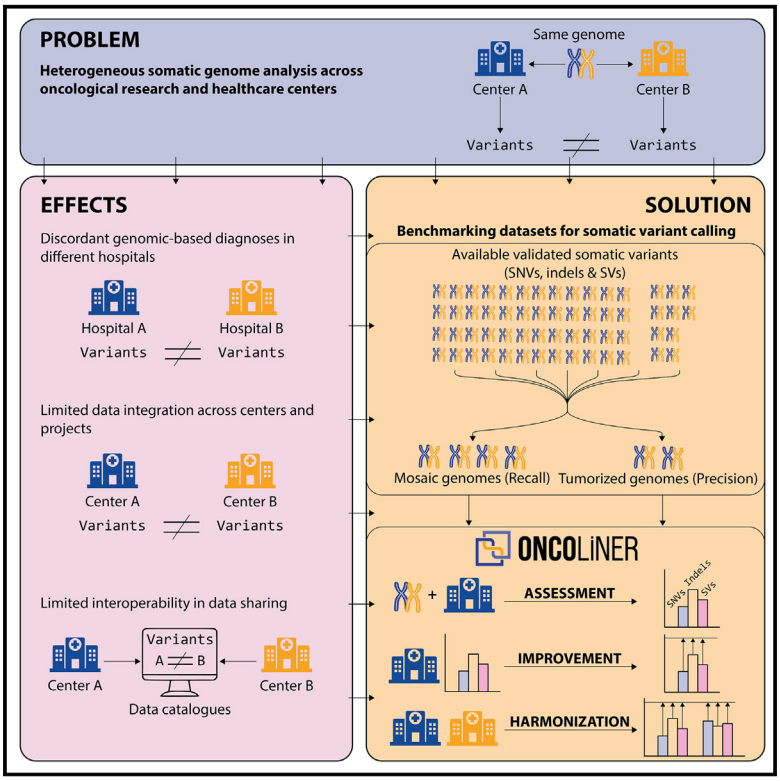
Cell Genomics

Article

ONCOLINER: A new solution for monitoring, improving, and harmonizing somatic variant calling across genomic oncology centers

Graphical abstract



Highlights

- Heterogeneity in cancer genome analysis affects diagnosis and sharing of data
- ONCOLINER is a solution to align somatic variant analysis across centers
- ONCOLINER provides recommendations to improve somatic variant calling
- Here, we also provide adapted benchmark datasets to assess somatic variant calling

Authors

Rodrigo Martín, Nicolás Gaitán, Frédéric Jarlier, ..., Romina Royo, Philippe Hupé, David Torrents

Correspondence

david.torrents@bsc.es

In brief

Current methods to identify somatic variants in oncological research and healthcare are highly heterogeneous across centers. Martín et al. present a new paradigm for assessing, improving, and harmonizing somatic variant calling across genomic oncology centers. This will allow the sharing of oncology data and provide consistency in diagnosis across hospitals.

Martín et al., 2024, Cell Genomics *4*, 100639
September 11, 2024 © 2024 The Authors. Published by Elsevier Inc.
https://doi.org/10.1016/j.xgen.2024.100639

CellPress

Cell Genomics

CellPress
OPEN ACCESS

Article

ONCOLINER: A new solution for monitoring, improving, and harmonizing somatic variant calling across genomic oncology centers

Rodrigo Martín,[1,13] Nicolás Gaitán,[1,13] Frédéric Jarlier,[2,3,4,5] Lars Feuerbach,[6] Henri de Soyres,[2,3,4,5] Marc Arbonés,[1] Tom Gutman,[2,3,4,5] Montserrat Puiggròs,[1] Alvaro Ferriz,[1] Asier Gonzalez,[1] Lucía Estelles,[7] Ivo Gut,[7] Salvador Capella-Gutierrez,[1] Lincoln D. Stein,[8,9] Benedikt Brors,[6,10] Romina Royo,[1] Philippe Hupé,[2,3,4,5,11] and David Torrents[1,12,14,*]

[1]Life Sciences Department, Barcelona Supercomputing Center (BSC), Barcelona, Spain
[2]Institut Curie, Paris, France
[3]U900, Paris, France
[4]PSL Research University, Paris, France
[5]Mines Paris Tech, Fontainebleau, France
[6]Division of Applied Bioinformatics, German Cancer Research Center (DKFZ), Heidelberg, Germany
[7]Centro Nacional de Análisis Genómico, Barcelona, Spain
[8]Department of Molecular Genetics, University of Toronto, Toronto, ON, Canada
[9]Ontario Institute for Cancer Research, Toronto, ON, Canada
[10]German Cancer Consortium (DKTK), Heidelberg, Germany
[11]UMR144, CNRS, Paris, France
[12]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain
[13]These authors contributed equally
[14]Lead contact
*Correspondence: david.torrents@bsc.es
https://doi.org/10.1016/j.xgen.2024.100639

SUMMARY

The characterization of somatic genomic variation associated with the biology of tumors is fundamental for cancer research and personalized medicine, as it guides the reliability and impact of cancer studies and genomic-based decisions in clinical oncology. However, the quality and scope of tumor genome analysis across cancer research centers and hospitals are currently highly heterogeneous, limiting the consistency of tumor diagnoses across hospitals and the possibilities of data sharing and data integration across studies. With the aim of providing users with actionable and personalized recommendations for the overall enhancement and harmonization of somatic variant identification across research and clinical environments, we have developed ONCOLINER. Using specifically designed mosaic and tumorized genomes for the analysis of recall and precision across somatic SNVs, insertions or deletions (indels), and structural variants (SVs), we demonstrate that ONCOLINER is capable of improving and harmonizing genome analysis across three state-of-the-art variant discovery pipelines in genomic oncology.

INTRODUCTION

Understanding how somatic genomic variation drives the biology of tumors is the foundation of modern personalized oncology. The characterization of somatic changes in cancer genomes has already uncovered hundreds of tumor-associated genes that can potentially be used as diagnosis, prognosis, and treatment markers.[1–4] For this reason, the analysis of tumor genomes has become a critical step within cancer genomic research and for its downstream clinical translation into personalized medicine protocols.

This has motivated the development of multiple somatic variant calling solutions over the past years, providing a wide catalog of different available tools and methods, each of them typically focused on specific types and sizes of variants.[5–22] Furthermore, the combination of different variant callers into complex pipelines has proven to be the best solution for different types of analyses in both research and clinical settings.[23–26] Selecting the best-performing tools and deciding how to best combine them to maximize recall and precision of all types of variant discovery, as well as their precise characteristics (e.g., tumor allele frequency, ploidy, exact break-junction sequence, etc.), are critical and challenging steps when developing genome analysis pipelines, as they require know-how as well as high-quality benchmarking information.[27–30] In addition, pipelines can be designed to prioritize recall or precision depending on the scenario of application, such as research or healthcare.

This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

Despite the publication of many benchmarking efforts, their results are usually descriptive and difficult to translate into practical decisions for the design and development of variant calling pipelines. This is due, for example, to the tendency to generate and use theoretical scenarios of application, combined with full *in silico* benchmarking datasets, which do not properly capture the nature of the biological behavior, noise, and variation of the data. That, together with the limited availability of real benchmarking data for somatic variant detection, defines the current scenario where variant calling across most of the different research and clinical centers is often limited and highly heterogeneous. This is particularly severe when considering the analysis of whole-genome sequencing (WGS) and the identification of structural variants (SVs), which are critical for cancer genomic studies. In this context, an increasing number of initiatives worldwide are aiming at introducing the sequencing and analysis of whole genomes as a routine for healthcare in genomic oncology, facing the challenge of genome analysis pipeline design and implementation.[31–37] It is therefore necessary to have actionable solutions in place, adjusted to the specific needs and scenarios that require variant calling pipeline design, development, or improvement.

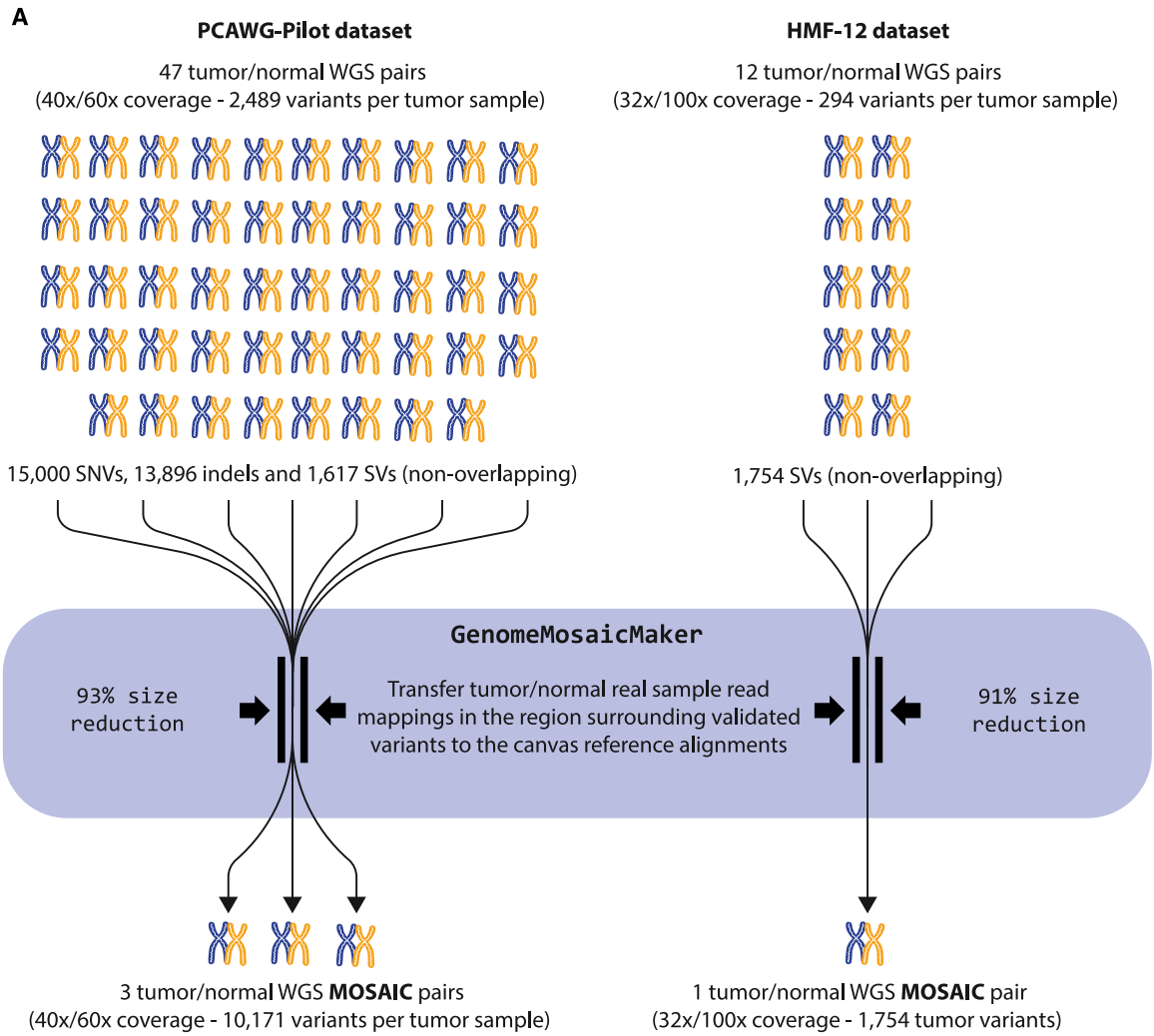
Current limitations in somatic variant calling have a direct impact on the quality and scope of downstream variant interpretation in research studies and clinical applications that impede the possibility of sharing, combining, and integrating data and results across different research groups and centers. In fact, growing national and worldwide efforts toward designing and building multicentric research ecosystems,[38,39] operating under federated or centralized schemes, require output harmonization of their different analysis platforms at both the level of quality and the scope of the variant calling, as well as at the level of standards that allow interoperability. Despite the high number of global initiatives to standardize and harmonize the management of biomedical data such as GA4GH[40] and ELIXIR (https://elixir-europe.org/), limited efforts have been devoted to the harmonization and standardization of analysis pipelines. The existing heterogeneity across different cancer research and clinical centers currently frustrates any attempt to integrate data and results, restricting the possibilities of new scientific discoveries, as well as generating potential discordant tumor reports from different healthcare centers. This heterogeneity also limits the chances for interoperability at other levels, for example through variant-based data discovery engines (e.g., Beacon from GA4GH), which ideally also require homogeneous and consistent variant data across centers. Because the adoption and maintenance of identical reference pipelines across many research centers is not a realistic solution for harmonization within growing federated (decentralized) data-sharing scenarios, we need applicable and practical solutions for the improvement and harmonization of somatic genome analysis across these data environments. This will allow us to answer more ambitious biomedical research questions and will enhance and globalize genomic oncology.

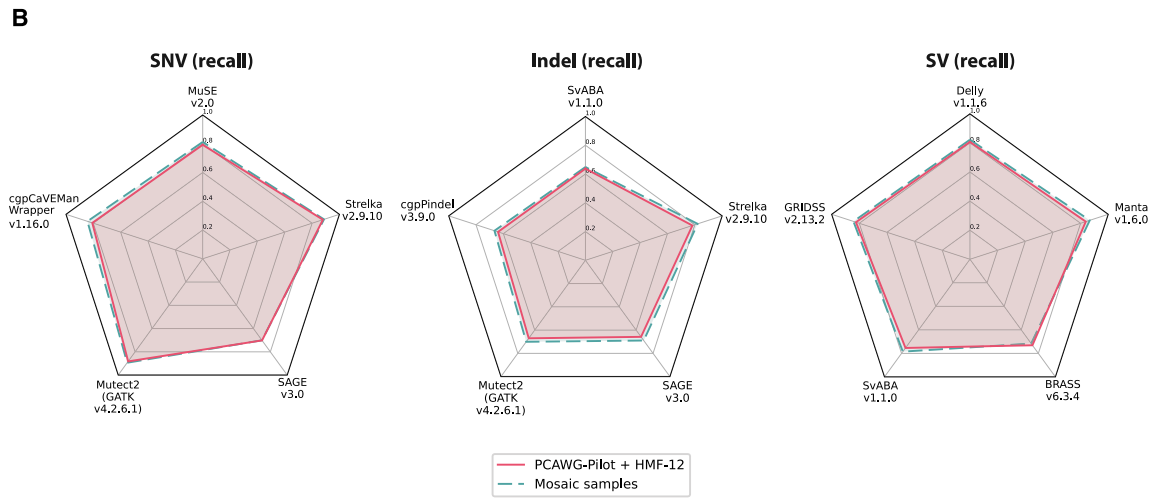
With this objective, we have designed a new actionable benchmarking paradigm, implemented as the ONCOLINER platform, a modular, configurable, and easy-to-use software solution that provides users with direct and personalized recommendations for building, improving, and harmonizing somatic variant calling pipelines from whole-genome short-read sequencing data within and across research and clinical oncology centers. The recommendations for improvement and harmonization are based on quality standards measured through common accuracy metrics using a comprehensive set of validated somatic variants. We have also processed and integrated these experimentally validated somatic variant data into more accessible benchmarking datasets that, in contrast to traditional *in silico* simulations, capture real data noise and variability impacting the calling of SNVs, insertions or deletions (indels), and SVs. We here demonstrate that the combination of these developments and resources significantly improved and harmonized real somatic variant calling pipelines that represent current research and clinical scenarios in genomic oncology.

RESULTS

Large data integration efforts in cancer research[1,4,21,41,42] have already highlighted significant heterogeneity in performance, quality, and scope of short-read-based genome analysis across different centers. With the ultimate aim of providing actionable and personalized solutions for the improvement and harmonization of genome analysis across cancer research centers and hospitals, we have first assessed the extent to which existing somatic variant calling pipelines need improvement and harmonization. For this, we have measured and compared the analysis performance (recall and precision) of three selected state-of-the-art variant calling pipelines from three active genome analysis research centers (A, B, and C). As recall and precision values reflect different properties of the variant calling, their measurement also requires specific and dedicated benchmarking datasets and scenarios.

To measure the recall (sensitivity) performance of these pipelines, we used a comprehensive collection of reported and validated somatic variants encapsulated within mosaic tumor-normal samples. These samples have been designed to retain the recall assessment properties of the original samples while reducing the analysis burden considerably (Figure 1A). Tumor and normal mosaic samples preserve the intrinsic noise and properties from the sample preparation (e.g., purity) and sequencing (e.g., insert sizes). To produce these mosaics, all the original sample reads mapped around a 2 kb window surrounding the coordinates of each validated variant are inserted into a WGS genome simulated from the GRCh37 reference while removing the artificial reads that overlap the span of the window. The same process is equally applied to the tumor and normal samples. For this study, we have built four mosaic genomes selecting a total of 32,267 validated variants (with non-overlapping genomic positions), distributed across 47 selected tumor-normal samples from the Pilot-50[4] and 12 from the HMF-12[9] datasets. These variants include 15,000 SNVs, 13,896 indels (composed of insertions, duplications, and deletions with a length less than 100 bp), and 3,371 SVs (i.e., variants with length greater than 100 bp). SVs are further subclassified in sizes and classes defined through their breakpoints, resulting in 736 translocations, 667 inversions, 1,318 deletions, and 650 duplications. While all the variants of the 12 HMF-12 samples could be integrated into a single mosaic tumor-normal pair with 1,754 SVs,





(legend on next page)

the heterogeneity of insert sizes across the 47 PCAWG Pilot-50 tumor-normal samples forced us to group them in three mosaic samples with consistent insert sizes and an average of 10,171 variants each, including SNVs, indels, and SVs (see Figures S1–S4). The variant allele frequency (VAF) of the mosaic genome variants ranges from 5% to 98%, following the original distribution of the PCAWG samples (see Figures S5 and S6). Within the scope of these variants and using a wide and representative collection of available somatic variant callers, we demonstrate that the mosaic genomes retain practically identical recall assessment properties to the original benchmarking datasets (Figure 1B).

On the other hand, to measure the reliability of the calling, we have calculated the precision performance of pipelines A, B, and C using tumorized genomes (Figure 2). In contrast to real or fully simulated benchmarking genomes, tumorized samples are hybrids that can accurately capture precision values and false positive rates. To balance the control of the variants and, at the same time, provide a real testing scenario to evaluate false positive calls, we combined samples from the GIAB project (NA12878 and HG002)[43] with a collection of synthetically reproduced true somatic variants (11,987 SNVs, 1,397 indels, and 239 SVs) from the PCAWG consensus callsets.[4] Tumorized samples contain only 0.2% of reads that are modified to accurately represent the sequence of the variant, while the other 99.8% of reads remain unaltered. Importantly, in contrast to mosaics, which may retain germline variants within somatic variant windows and require controlled access agreements, tumorized samples can be shared openly within the community as described by the GIAB project.[43]

Analysis of heterogeneity

To precisely assess the extent to which pipelines A, B, and C need improvement and harmonization, we tested them against the four mosaics and the two tumorized benchmarking samples described above and extracted recall and precision values for each variant type and size. An initial inspection of the results already shows some degree of heterogeneity and possibilities for improvement for some pipelines across different variant types and sizes (see Figure S7). As a signal of how different two or more pipelines perform against the same benchmark samples, we have defined a performance heterogeneity score (PHS), reflecting the differences between their performances, as of recall and precision (Figure 3A). Moreover, we also studied the functional and clinical impact that these pipelines are able to capture and how heterogeneous this is across centers by evaluating the fraction of coding genes, including cancer drivers,[2] that are affected by non-synonymous variants (i.e., gene discordance ratio [GDR]).

Overall, the three pipelines together could correctly identify 28,644 variants (13,834 SNVs, 11,704 indels, and 3,106 SVs) of the 32,267 original truth variants from all four mosaic samples. From the correctly identified variants, only 19,805 (10,052 SNVs, 7,273 indels, and 2,480 SVs) were concordantly identified by the three centers, whereas 8,839 (3,782 SNVs, 4,431 indels, and 626 SVs) were missed by at least one of the pipelines. Pairwise comparisons between pipelines A-B, A-C, and B-C showed 7,612, 3,616, and 6,450 discordant variants, respectively. Translating these differences into potential functional and clinical impact, from a total of 3,217 genes with somatic coding mutations, 2,624 were found by all three pipelines. Conversely, 593 (18%) of them, including 24 coding cancer drivers, were missed or excluded by at least one of the pipelines. Among these, we found genes that are key in decision-making processes within tumor boards for the diagnosis, treatment, and prognosis of different cancer types (ONCOKB[45,46]). For example, the gene *KIT*, which codes for a receptor tyrosine kinase, is a proto-oncogene and a US Food and Drug Administration-approved therapy target of kinase inhibitor drug groups.[47–50] Similarly, among the discordant variants, we found a translocation disrupting the *RARA* gene, a retinoic acid receptor that is found translocated as a gene fusion in certain types of leukemia. In particular, some *RARA* mutations are standard diagnostic biomarkers for acute promyelocytic leukemia.[50,51] Another important discordance was a translocation affecting *CCND3*. Alterations in *CCND3* are used as prognostic biomarkers of various hematologic malignancies.[50,52] Lastly, other genes, used as support for the diagnosis of different cancer types, such as *FOXP1* and *NOTCH2*, also presented mutations that were discordantly detected between the three pipelines.[50]

Overall, PHSs among the three pipelines range from 1.92% to 36.53%, showing heterogeneity across all variant types and sizes (Figure 3B). We also observe differences in PHSs affected by precision and recall heterogeneity values asymmetrically. For SNVs, in agreement with previous studies,[53] we find a significant performance heterogeneity (PHS = 11.71%), mostly affecting recall values (Figure 3B), which translates into 3,782 discordant variants from all 13,834 SNVs identified by at least one center, as well as 58 discordant coding and 4 cancer driver genes. We also detected a degree of heterogeneity (PHS = 4.14%) for

Figure 1. Generation of the mosaic tumor/normal genome pairs

(A) To assess the levels of recall for somatic analysis pipelines, we have designed mosaic tumor-normal genome pairs, which condense the benchmarking information of a total of 59 (47 + 12) validated tumor-normal pairs into only four resulting mosaic pairs. This is achieved by literally transferring (copying) a 2 kb region of mapped reads around each validated variant of the original genomes into a canvas genome generated by simulated reads from the human reference genome (GRCh37). Whereas HMF-12 samples have homogeneous insert sizes and all the variants could be condensed into a single tumor-normal mosaic genome pair, the heterogeneity across samples of the PCAWG-Pilot dataset forced us to generate three sample pairs with similar ranges of insert sizes (see Figure S1). This avoids conflicts from some variant callers that depend on and need to generate internal decision criteria based on read pair distances. Mosaic genomes reproduce the exact same genomic context for variant detection as in the original samples. In fact, in (B), we show that these mosaic genomes retain the recall benchmarking properties of the original source samples. The results that we obtained using 12 different variant callers on the recall for SNVs, indels, and SVs are practically identical between the 59 original samples (yellow line) and the derived mosaic genomes (blue line). During the quality check of the data, we discarded two outlier samples from the 49 available tumor-normal pairs from the PCAWG-Pilot dataset. Defined by the original data, this benchmarking set considers effective sequencing coverages ranging from 27× to 62× for the normal samples and from 27× to 149× for the tumor samples (see Figure S2), as well as variants with a defined variant allele frequency (VAF) larger than 1% (see Figure S6).

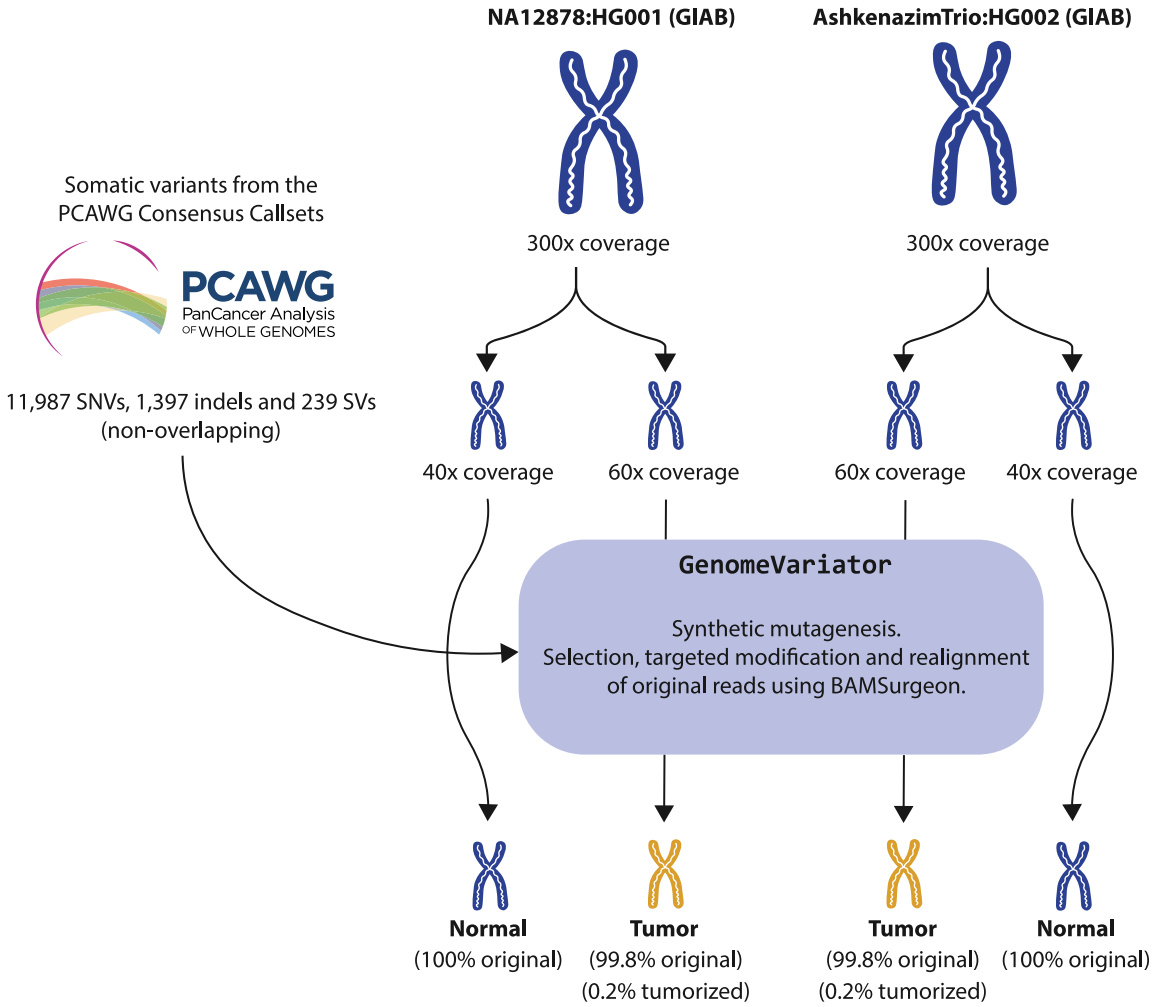


Figure 2. Design and construction of the tumorized tumor-normal genome pairs

For an accurate assessment of the precision and to calculate the rate of false positives during somatic variant discovery, we have generated tumorized genomes. These consist of real WGS samples from the GIAB project[43] with synthetic cancer somatic variants extracted from the PCAWG consensus callsets.[4] For each introduced variant, a subset of reads in the tumor sample are modified to represent the variant. The number of modified reads depends on the depth of the region where the variant is located and the selected VAF (see Figure S6). This method is implemented in GenomeVariator, a wrapper tool that enhances the functionalities of BAMSurgeon.[25,26,44] The high coverage of these samples (300×) allows the generation of tumor and normal genomes with a different composition of reads, recreating real tumor-normal analysis scenarios. Furthermore, the fact that only 0.2% of the reads have been modified to reconstruct the variants in the tumor samples of the tumorized genome pairs makes these samples ideal for an accurate evaluation of precision, as they retain 99.8% of the original sequencing and mapping properties. In order to avoid potential sample bias, we have generated two tumor-normal samples with the same validated variants: one derived from the NA12878 GIAB sample and the other from the HG002 GIAB sample.

indels, resulting from the discordance of 4,431 variants that have been missed by at least one center, affecting 39 coding and 4 cancer driver genes. Finally, the calling of SVs also shows low concordance across centers, with 16 discordant cancer driver genes disrupted by SVs in total (PHS = 6.63%). Particularly, the 100–500 bp range shows the two highest PHS values, with 36.53% for duplications and 29.22% for deletions, translating into 95 duplications and 68 deletions of the total 626 SVs differently identified by these pipelines, respectively (Figure 3C). Conversely, deletions and duplications with lengths above 500 bp present the two lowest PHSs (1.92% and 3.7%, respectively). This demonstrates that the detection of SVs within the 100–500 bp range remains challenging with short-read technologies. Altogether, duplications and deletions in the 100–500 bp range cause 47 coding and 1 cancer driver and 34 coding and 4 cancer discordant genes, respectively (Figure 3D). Interestingly, discordant false positive deletions produce significant differences in precision across pipelines, whereas duplications of the same size differ mostly in recall (see Figure 3B).

ONCOLINER solution

To overcome these quantitative and qualitative differences across centers, and driven by the specific needs of each of the pipelines, we designed and implemented a solution called

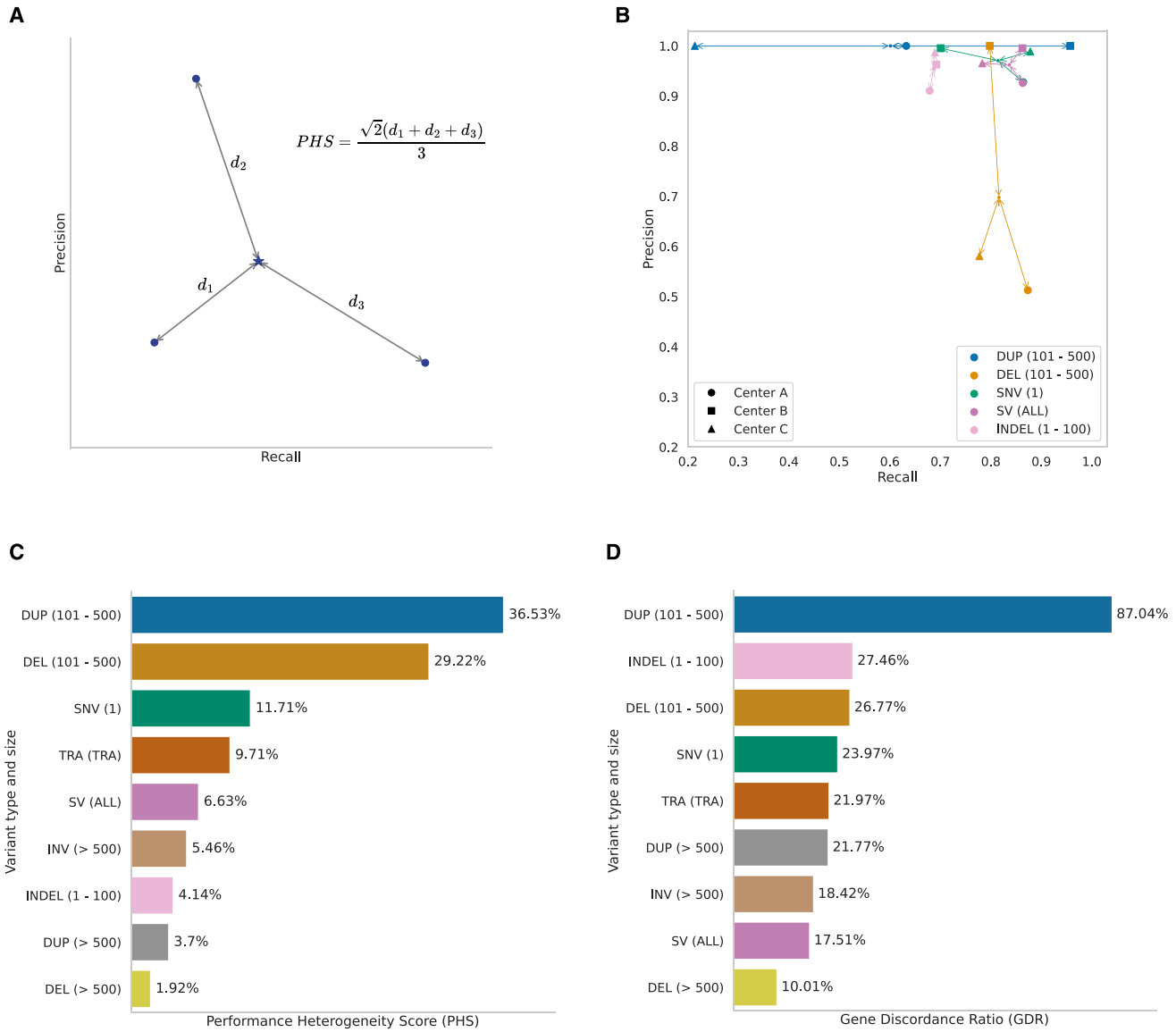


Figure 3. Comparative study of somatic variant analysis across different research and clinical scenarios

(A) A graphic and mathematical representation of the underlying rationale for the definition of PHS. This score measures the degree of heterogeneity in somatic variant calling performance across two or more centers, independently of the overall quality of each pipeline. It is calculated as the normalized average distance from each pipeline to the centroid in a Euclidean space delimited by recall and precision values. PHS values of 0% indicate no heterogeneity, whereas those of 100% indicate maximum heterogeneity.

(B) The results for the top four variant types and sizes by PHS for centers A, B, and C, representative of real and active genome analysis pipelines. SVs are grouped by type, such as deletions (DEL), duplications (DUP), translocations (TRA) and inversions (INV) and by size (from 101 to 500 bp and more than 500 bp). Performance parameters, including recall and precision, were obtained using a validated set of variants with a total of 38,947 SNVs, 16,688 indels, and 3,851 SVs of different subtypes and sizes. The samples only contained inversions above 500 bp, so smaller ones could not be assessed (see Figures 1 and 2).

(C) The PHSs for all variant types.

(D) The effect of this calling heterogeneity on the functional reach and impact of the analysis in the form of the fraction of discordant genes identified as mutated by at least one center and missed by another (GDR).

ONCOLINER to assist in the development and improvement of somatic genome analysis pipelines for cancer genomics research and clinical oncology. ONCOLINER provides users with multiple functionalities encapsulated in different interoperable modules, covering the assessment, improvement, and harmonization of already operational variant calling pipelines, to the *de novo* generation of optimized pipelines (Figure 4). In brief, this tool first analyzes targeted pipelines, and then makes a diagnosis based on calling performances and harmonization levels, to finally provide improved and harmonized solutions in

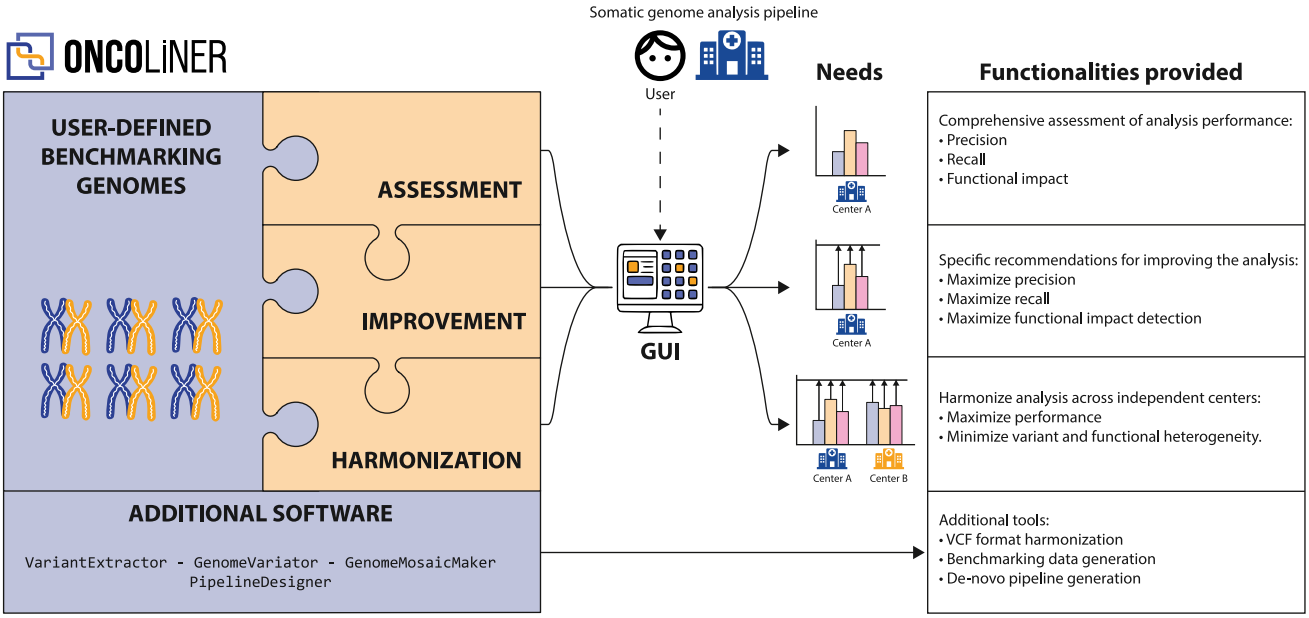


Figure 4. Conceptual and functional map of ONCOLINER

Left of the image, a box represents the different functional modules and basic elements of ONCOLINER. In our use case, the defined benchmarking genomes are composed of three mosaic and two tumorized tumor-normal genome pairs that have been designed to capture recall and precision performance, respectively (see Figures 1 and 2). The three modules in orange correspond to the three main functionalities, assessment, improvement, and harmonization, which are described on the right part of the image. In addition, shown in the blue box on the bottom, ONCOLINER also provides different standalone solutions for the improvement and harmonization of genome analysis in additional contexts. In the middle of the image, a user interface (GUI) allows interaction with the platform and the selection of the best solution for each specific clinical or research need.

the form of specific and actionable recommendations, such as adding or intersecting with additional variant callers.

The first step of ONCOLINER requires users to execute their analysis pipelines against the benchmarking data, corresponding in this case to the four mosaics and the two tumorized samples. After entering the pipeline's results (in VCF format), the assessment module first calculates basic performance metrics, such as recall and precision, across a wide range of variant types and sizes (see Figure S7), highlighting parts that can be improved. Next, targeting these limitations, the improvement module provides specific recommendations that increase detection power or precision for each of the variant types and sizes. These recommendations are based on the current knowledge in state-of-the-art somatic variant detection, obtained through thorough testing and benchmarking of 12 selected variant callers and their combinations over the mosaic and tumorized benchmarking genomes. Recommendations targeting the improvement of recall values for certain pipelines and variant types involve adding one or more of these variant callers, whereas recommendations to lower the rate of false positives and improve precision involve recommending intersections. As different research and clinical scenarios prioritize either recall or precision differently, ONCOLINER generates different possibilities for improvement, from which users can select the most suitable option for their needs. Finally, if more than one pipeline has been provided, then the harmonization module enables selecting the recommendations for improvement that also maximize the harmonization across centers, quantitatively and qualitatively.

Along with ONCOLINER, we also provide additional standalone software solutions to meet multiple needs associated with improving, standardizing, and harmonizing genome analysis across centers. For example, to ensure interoperable interpretation and representation of variants from VCF files, we developed VariantExtractor as a library used by ONCOLINER and also as a standalone package. VariantExtractor reads and interprets SNV, indel, and SV records by applying a set of consistent rules across all VCFs. This also facilitates downstream analysis with no information losses, especially for SVs, as there are different ways to encode the same variant that are biologically identical but very different in the VCF format (see Figure S8). In addition, we also provide GenomeVariator and GenomeMosaicMaker for the generation of custom tumorized and mosaic genomes, respectively. These tools can be used to generate genomes adapted to specific needs for benchmarking data.

Beyond the functionalities aiming at improving and harmonizing existing pipelines, we also include PipelineDesigner, a standalone tool that helps users to find the best strategy to combine and merge specific variant callers to maximize recall and precision over all variant types. Using PipelineDesigner, we designed and implemented a *de novo* variant calling pipeline with the combination of the best-performing variant callers of this study (see Table S1) that can be readily adopted. To facilitate the application of the recommendations from both ONCOLINER and PipelineDesigner, we provide each variant caller we used in a container and the necessary tools for merging and combining their results to provide a consistent framework for applying

Table 1. Pipeline heterogeneity and performance after ONCOLINER

	SNV	Indel	SV
Discordant variants	1,548 (2,234 ↓)	1,212 (3,219 ↓)	92 (534 ↓)
Discordant genes	22 (36 ↓)	12 (27 ↓)	60 (436 ↓)
Discordant driver genes	2 (2 ↓)	1 (3 ↓)	0 (16 ↓)
GDR, %	9 (15 ↓)	8 (19 ↓)	2 (15 ↓)
PHS, %	4 (7 ↓)	4 (0 =)	3 (4 ↓)
Average F1-score, %	92 (3 ↑)	87 (7 ↑)	90 (2 ↑)

improvement and harmonization recommendations. In addition, to increase flexibility and ensure future applicability over a wider range of variant calling benchmarking scenarios, we allowed ONCOLINER to function with other provided benchmarking datasets and other reference variant callers, adapting to the specific needs of the user.

Application to a real research scenario

In order to prove ONCOLINER's functional applicability to a common research scenario, we have used pipelines A, B, and C. First, the assessment module calculated and provided all variant calling metrics for each pipeline and each variant type, including affected coding and cancer driver genes, from the analysis of the tumorized and mosaic genomes (see Figures S9–S11). Next, the improvement module provided recommendations that enhance recall (at least 5%), precision (at least 5%), or F1-score. For example, ONCOLINER's recommendation of adding GRIDSS2 to pipelines A and C improved their recall for detection of SVs from 86% to 91% and from 78% to 89%, respectively. This recommendation also increased the number of discovered mutated protein-coding genes for both pipelines by 277 and 274 and cancer driver genes by 12 and 8 for A and C, respectively. The recommended intersection of pipeline A with mutect2 (from GATK) and SAGE removed 1,741 false positives for SNVs and increased the precision from 93% to 99.99% (see Figures S12 and S13).

Finally, the harmonization module evaluated and selected those recommendations for improvement that also minimized heterogeneity scores across centers. For example, among multiple choices with similar outcomes (see Figures S14 and S15), the addition of GRIDSS2 to pipelines A and C not only improved their recall as described above but also reduced PHSs from 6.63% to 2.57% for SVs, with notable effects in duplications between 101 and 500 bp decreasing PHSs from 36.53% to 4.03%. Moreover, this harmonization option decreased the GDR for SVs from 17.50% to 2.09% (notably from 87.04% to 3.64% for 100–500 bp duplications), which translated into 436 less discordant affected genes across pipelines. Overall, prioritizing PHSs, the recommended strategies for harmonization and improvement made consistent a total of 5,987 true variants (2,234 SNVs, 3,219 indels, and 534 SVs) out of the initial 8,839 discordant variants across the three centers and also recovered variants that affected 499 protein-coding and 21 cancer driver genes, including the five actionable genes previously missed by at least one of the three pipelines (Table 1). Despite considerably improving performance and homogeneity across centers, a total of 2,852 true variants remained discordant after a first iteration with ONCOLINER, including 94 affected genes and three drivers. The improvement and harmonization reached here apply to these specific pipelines. We expect even higher improvement and harmonization levels across research centers and hospitals that, for example, require the *de novo* inclusion of WGS and SVs into their protocols.

DISCUSSION

Naturally, the ultimate value of benchmarking efforts during the development and improvement of genome analysis pipelines critically depends on the quality and scope of the reference (truth) set of validated variants. They will determine the reach of the assessment and its final level of trust. Unfortunately, there are only a handful of available and suitable patient-derived datasets with enough numbers and varieties of validated somatic variants in WGS for building and calibrating somatic variant identification and classification pipelines. Of these, we have used the PCAWG-Pilot[4] (SNVs, indels, and SVs) and the HMF-12[9] (SVs) datasets, both generated in benchmarking contexts of somatic variant calling. While the HMF dataset is more homogeneous and internally consistent, the PCAWG set derives from samples collected, processed, and sequenced in multiple centers with different quality standards, such as tumor purity, insert size, and sequencing coverage. This affects not only the mosaic strategy, which requires homogeneous insert sizes, but also the scope of this study. For this reason, we cannot assure that pipelines calibrated with these specific datasets will actually translate in calling improvement when applied outside their sample purity, sequencing error rate, coverage, and insert size ranges (see Figure 1). Considering these limitations, we plan to improve further benchmarking datasets by generating publicly accessible tumor-normal benchmarking genomes for the evaluation of recall and precision of somatic variant calling in a single run.

At the same time, the identification and validation methods used to generate truth sets of variants also determine the value and reach of benchmarking studies. For instance, at the level of variant type, our benchmarking datasets cover SNVs, indels, and breakpoint-definable SVs (deletions, inversions, translocations, and duplications) but do not include, for example, large insertions and coverage-derived copy-number variants. In addition, other data quality issues can also affect the reliability of pipeline assessment. Among these, even validated somatic variation datasets still contain a certain amount of germline contamination and sequencing errors recalled as low-VAF SNVs. This

could slightly generate underestimations of recall values, even in this study. Upcoming efforts aiming to generate more accurate, customized, and targeted benchmarking datasets for other potential applications will surely improve the usability and quality of the functionalities of ONCOLINER.

The reach of ONCOLINER is determined by the collection of preselected variant callers used to generate recommendations for improvement and harmonization. Although we have selected twelve variant callers based on their acceptance and use within the community, we cannot discard that other variant callers could, in fact, outperform this set and generate better recommendations for improvement and harmonization. For this reason, together with customized benchmarking datasets, we have also allowed the inclusion by the user of additional variant callers into the platform to be able to improve the calling in general or to target specific types of variants not included here. Among other benchmarking datasets that can be used with ONCOLINER, we can find high-quality tumor-normal pairs in previous studies.[23,30] Nevertheless, even considering the limitations of the datasets used as the ground truth, ONCOLINER managed to substantially improve and harmonize the performance of the pipelines of the three centers, leading to the recovery of mutations on five discordant clinically actionable cancer genes. Thus, further benchmarking efforts applying this paradigm shift with curated gold-standard variants will be able to generate actionable recommendations for health centers.

Taking these results together, we present and validate a new concept for the benchmarking of somatic variant discovery with actionable recommendations to users for improving and harmonizing across centers the identification of somatic variants associated with cancer. The application of ONCOLINER to align genome analysis across research centers and hospitals can provide consistency in the diagnosis and selection of treatment within primary care, as well as for the possibility of improving scientific discovery by allowing an interoperable integration and sharing of cancer genomics datasets within emerging federated data spaces around the world.

Limitations of the study

There are two major aspects of ONCOLINER that can have generic and specific limitations, with potential consequences for users. One relies on the quality of the benchmarking dataset, which determines the scope and the reliability of all functionalities of ONCOLINER. Benchmarking datasets with low quality or low diversity of variants will result in inaccurate and poor performance assessments, which will also affect all improvement and harmonizing recommendations. For our study, we have taken two specific datasets[4,9] that cover SNVs, indels, and SVs that have passed different rounds of quality check, but we cannot discard, for example, that a fraction of variants labeled as somatic within original datasets are, in fact, germline. Other sources of limitations rely on the sequencing and preprocessing methods that have been used on those original benchmark datasets, which also determine the scope of application of ONCOLINER. To solve these limitations, users can provide their own benchmarking dataset with specific parts of their methodology considered. Finally, other limitations rely on the algorithm and implementation of ONCOLINER, which has been designed to offer full functionality within low computational requirements. This compromise forced us to allow the prioritization of the harmonization to rely on the improvement of either the recall or the precision but not both at the same time. Although we do not expect much impact during implementations, this can sometimes result in suboptimal harmonization recommendations for precision and require follow-up executions of ONCOLINER.

RESOURCE AVAILABILITY

Lead contact

Further information and requests should be directed to and will be fulfilled by the lead contact, David Torrents (david.torrents@bsc.es).

Material availability

No materials were generated in this study.

Data and code availability

- The interactive HTML report generated by ONCOLINER for centers A, B and C has been deposited at a custom HTTP server and is publicly available as of the date of publication. URL is listed in the key resources table.
- Tumorized genomes CRAM and VCF files have been deposited at ENA and are publicly available as of the date of publication. Accession numbers are listed in the key resources table.
- Mosaic tumor-normal genome pairs CRAM files and Gold Standard VCF files from PCAWG-Pilot have been deposited at ICGC Data Portal, and accession numbers are listed in the key resources table. They are available upon request if access is granted.
- Mosaic tumor-normal genome pairs CRAM files and Gold Standard VCF files from HMF-12 have been deposited at EGA, and accession numbers are listed in the key resources table. They are available upon request if access is granted.
- All original code for ONCOLINER, PipelineDesigner, GenomeVariator and GenomeMosaicMaker has been deposited at GitHub and is publicly available as of the date of publication. DOIs are listed in the key resources table.
- All original code for VariantExtractor has been deposited at GitHub and is publicly available as of the date of publication. DOI is listed in the key resources table. VariantExtractor is also available through PyPi.
- All the variant callers included in this study are listed and can be downloaded as Singularity containers from GitHub and are publicly available as of the date of publication. URL is listed in the key resources table.

Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

BSC-CNS discloses support for the research of this work from the European Union's Horizon 2020 research and innovation programme under EUCANCan (grant agreement no. 825835), the Instituto de Salud Carlos III (ISCIII) and "Unión Europea NextGenerationEU/Mecanismo para la Recuperación y la Resilencia (MRR)/PRTR" under project PMP21/00015, the Departament de Recerca i Universitats de la Generalitat de Catalunya (code: 2021 SGR 01626), and the Science and Innovation Spanish Ministry under project BenchSV (PID2020-119797RB-I00/AEI/10.13039/501100011033). Institut Curie discloses support for the research of this work from the European Union's Horizon 2020 research and innovation programme under EUCANCan (grant agreement no. 825835) and Cancéropôle Île-de-France (grant GENOPROFILE - RIC2021). The German Cancer Research Center (DKFZ) received funding from the European Union's Horizon 2020 research and innovation programme and the Canadian Institutes of Health Research under the grant agreement no. 825325. CNAG institutional support was from the Spanish Instituto de Salud Carlos III, Fondo de Investigaciones Sanitarias, and cofunded with ERDF funds (PI19/01772); the Spanish Ministry of Science and Innovation through the Instituto de Salud Carlos III; the 2014–2020 Smart Growth Operating Program and

institutional co-financing with the European Regional Development Fund (MINECO/FEDER, BIO2015-71792-P); and the Generalitat de Catalunya through the Departament de Salut, Departament d'Empresa i Coneixement. We acknowledge Hartwig Medical Foundation for their help and contributions to different parts of the study. We specifically thank and highly appreciate the valuable contribution of Matias Mendeville. In addition, we would like to express our sincere gratitude to the Genome in a Bottle Consortium (GIAB), the International Cancer Genome Consortium (ICGC), and The Cancer Genome Atlas (TCGA) for making high-quality data accessible through secure protocols, which have been instrumental to the success of this study.

AUTHOR CONTRIBUTIONS

D.T. and P.H. contributed to study conception and design and jointly directed the work. D.T., P.H., B.B., L.D.S., S.C.-G., and I.G. were principal investigators and contributed to study initiation. R.M., N.G., F.J., L.F., H.d.S., M.A., T.G., M.P., A.F., A.G., L.E., and R.R. contributed to the study methodology. All authors had unrestricted access to final study data and were responsible for data interpretation, preparation of the manuscript, and the decision to submit for publication. The manuscript was written and compiled by D.T., P.H., R.M., N.G., F.J., and L.F. R.M. and N.G. have accessed and verified the data. All authors attest to study completeness and the accuracy of the data and data analysis and approved the final version of the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- METHOD DETAILS
  - Benchmarking datasets: Mosaic and tumorized genomes
  - Variant caller selection
  - Assessment module
  - Improvement module
  - Harmonizer module

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xgen.2024.100639.

Received: March 19, 2024
Revised: June 13, 2024
Accepted: August 7, 2024
Published: August 30, 2024

REFERENCES

1. International Cancer Genome Consortium; Hudson, T.J., Anderson, W., Artez, A., Barker, A.D., Bell, C., Bernabé, R.R., Bhan, M.K., Calvo, F., Eerola, I., et al. (2010). International network of cancer genome projects. Nature *464*, 993–998. https://doi.org/10.1038/nature08987.
2. Martínez-Jiménez, F., Muiños, F., Sentís, I., Deu-Pons, J., Reyes-Salazar, I., Arnedo-Pac, C., Mularoni, L., Pich, O., Bonet, J., Kranas, H., et al. (2020). A compendium of mutational cancer driver genes. Nat. Rev. Cancer *20*, 555–572. https://doi.org/10.1038/s41568-020-0290-x.
3. Tate, J.G., Bamford, S., Jubb, H.C., Sondka, Z., Beare, D.M., Bindal, N., Boutselakis, H., Cole, C.G., Creatore, C., Dawson, E., et al. (2019). COSMIC: the catalogue of somatic mutations in cancer. Nucleic Acids Res. *47*, D941–D947. https://doi.org/10.1093/nar/gky1015.
4. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium (2020). Pan-cancer analysis of whole genomes. Nature *578*, 82–93. https://doi.org/10.1038/s41586-020-1969-6.
5. Wala, J.A., Bandopadhayay, P., Greenwald, N.F., O'Rourke, R., Sharpe, T., Stewart, C., Schumacher, S., Li, Y., Weischenfeldt, J., Yao, X., et al. (2018). SvABA: genome-wide detection of structural variants and indels by local assembly. Genome Res. *28*, 581–591. https://doi.org/10.1101/gr.221028.117.
6. Rausch, T., Zichner, T., Schlattl, A., Stütz, A.M., Benes, V., and Korbel, J.O. (2012). DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics *28*, i333–i339. https://doi.org/10.1093/bioinformatics/bts378.
7. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M.A. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303. https://doi.org/10.1101/gr.107524.110.
8. Kim, S., Scheffler, K., Halpern, A.L., Bekritsky, M.A., Noh, E., Källberg, M., Chen, X., Kim, Y., Beyter, D., Krusche, P., and Saunders, C.T. (2018). Strelka2: fast and accurate calling of germline and somatic variants. Nat. Methods *15*, 591–594. https://doi.org/10.1038/s41592-018-0051-x.
9. Cameron, D.L., Baber, J., Shale, C., Valle-Inclan, J.E., Besselink, N., van Hoeck, A., Janssen, R., Cuppen, E., Priestley, P., and Papenfuss, A.T. (2021). GRIDSS2: comprehensive characterisation of somatic structural variation using single breakend variants and structural variant phasing. Genome Biol. *22*, 202. https://doi.org/10.1186/s13059-021-02423-x.
10. Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Källberg, M., Cox, A.J., Kruglyak, S., and Saunders, C.T. (2016). Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics *32*, 1220–1222. https://doi.org/10.1093/bioinformatics/btv710.
11. Fan, Y., Xi, L., Hughes, D.S.T., Zhang, J., Zhang, J., Futreal, P.A., Wheeler, D.A., and Wang, W. (2016). MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. Genome Biol. *17*, 178. https://doi.org/10.1186/s13059-016-1029-6.
12. Raine, K.M., Hinton, J., Butler, A.P., Teague, J.W., Davies, H., Tarpey, P., Nik-Zainal, S., and Campbell, P.J. (2015). cgpPindel: Identifying somatically acquired insertion and deletion events from paired end sequencing. Curr. Protoc. Bioinformatics *52*, 15.7.1–15.7.12. https://doi.org/10.1002/0471250953.bi1507s52.
13. Jones, D., Raine, K.M., Davies, H., Tarpey, P.S., Butler, A.P., Teague, J.W., Nik-Zainal, S., and Campbell, P.J. (2016). cgpCaVEManWrapper: Simple execution of CaVEMan in order to detect somatic single nucleotide variants in NGS data. Curr. Protoc. Bioinformatics *56*, 15.10.1–15.10.18. https://doi.org/10.1002/cpbi.20.
14. Hansen, N.F., Gartner, J.J., Mei, L., Samuels, Y., and Mullikin, J.C. (2013). Shimmer: detection of genetic alterations in tumors using next-generation sequence data. Bioinformatics *29*, 1498–1503. https://doi.org/10.1093/bioinformatics/btt183.
15. Rimmer, A., Phan, H., Mathieson, I., Iqbal, Z., Twigg, S.R.F., WGS500 Consortium; Wilkie, A.O.M., McVean, G., and Lunter, G. (2014). Integrating mapping-assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. Nat. Genet. *46*, 912–918. https://doi.org/10.1038/ng.3036.
16. Layer, R.M., Chiang, C., Quinlan, A.R., and Hall, I.M. (2014). LUMPY: a probabilistic framework for structural variant discovery. Genome Biol. *15*, R84. https://doi.org/10.1186/gb-2014-15-6-r84.
17. Fan, X., Abbott, T.E., Larson, D., and Chen, K. (2014). BreakDancer: Identification of genomic structural variation from paired-end read mapping. Curr. Protoc. Bioinformatics *45*, 15.6.1–15.6.11. https://doi.org/10.1002/0471250953.bi1506s45.
18. Larson, D.E., Harris, C.C., Chen, K., Koboldt, D.C., Abbott, T.E., Dooling, D.J., Ley, T.J., Mardis, E.R., Wilson, R.K., and Ding, L. (2012).

SomaticSniper: identification of somatic point mutations in whole genome sequencing data. Bioinformatics *28*, 311–317. https://doi.org/10.1093/bioinformatics/btr665.

19. Christoforides, A., Carpten, J.D., Weiss, G.J., Demeure, M.J., Von Hoff, D.D., and Craig, D.W. (2013). Identification of somatic mutations in cancer through bayesian-based analysis of sequenced genome pairs. BMC Genom. *14*, 302. https://doi.org/10.1186/1471-2164-14-302.
20. Moncunill, V., Gonzalez, S., Beà, S., Andrieux, L.O., Salaverria, I., Royo, C., Martinez, L., Puiggròs, M., Segura-Wang, M., Stütz, A.M., et al. (2014). Comprehensive characterization of complex structural variations in cancer by directly comparing genome sequence reads. Nat. Biotechnol. *32*, 1106–1112. https://doi.org/10.1038/nbt.3027.
21. Martínez-Jiménez, F., Movasati, A., Brunner, S.R., Nguyen, L., Priestley, P., Cuppen, E., and Van Hoeck, A. (2023). Pan-cancer whole-genome comparison of primary and metastatic solid tumours. Nature *618*, 333–341. https://doi.org/10.1038/s41586-023-06054-z.
22. Nik-Zainal, S., Davies, H., Staaf, J., Ramakrishna, M., Glodzik, D., Zou, X., Martincorena, I., Alexandrov, L.B., Martin, S., Wedge, D.C., et al. (2016). Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature *534*, 47–54. https://doi.org/10.1038/nature17676.
23. Alioto, T.S., Buchhalter, I., Derdak, S., Hutter, B., Eldridge, M.D., Hovig, E., Heisler, L.E., Beck, T.A., Simpson, J.T., Tonon, L., et al. (2015). A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. Nat. Commun. *6*, 10001. https://doi.org/10.1038/ncomms10001.
24. Nadeu, F., Royo, R., Massoni-Badosa, R., Playa-Albinyana, H., Garcia-Torre, B., Duran-Ferrer, M., Dawson, K.J., Kulis, M., Diaz-Navarro, A., Villamor, N., et al. (2022). Detection of early seeding of richter transformation in chronic lymphocytic leukemia. Nat. Med. *28*, 1662–1671. https://doi.org/10.1038/s41591-022-01927-8.
25. Ewing, A.D., Houlahan, K.E., Hu, Y., Ellrott, K., Caloian, C., Yamaguchi, T.N., Bare, J.C., P'ng, C., Waggott, D., Sabelnykova, V.Y., et al. (2015). Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. Nat. Methods *12*, 623–630. https://doi.org/10.1038/nmeth.3407.
26. Lee, A.Y., Ewing, A.D., Ellrott, K., Hu, Y., Houlahan, K.E., Bare, J.C., Espiritu, S.M.G., Huang, V., Dang, K., Chong, Z., et al. (2018). Combining accurate tumor genome simulation with crowdsourcing to benchmark somatic structural variant detection. Genome Biol. *19*, 188. https://doi.org/10.1186/s13059-018-1539-5.
27. Olson, N.D., Wagner, J., Dwarshuis, N., Miga, K.H., Sedlazeck, F.J., Salit, M., and Zook, J.M. (2023). Variant calling and benchmarking in an era of complete human genome sequences. Nat. Rev. Genet. *24*, 464–483. https://doi.org/10.1038/s41576-023-00590-0.
28. Cameron, D.L., Di Stefano, L., and Papenfuss, A.T. (2019). Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software. Nat. Commun. *10*, 3240. https://doi.org/10.1038/s41467-019-11146-4.
29. Cortés-Ciriano, I., Gulhan, D.C., Lee, J.J.-K., Melloni, G.E.M., and Park, P.J. (2021). Computational analysis of cancer genome sequencing data. Nat. Rev. Genet. *23*, 298–314. https://doi.org/10.1038/s41576-021-00431-y.
30. Fang, L.T., Zhu, B., Zhao, Y., Chen, W., Yang, Z., Kerrigan, L., Langenbach, K., de Mars, M., Lu, C., Idler, K., et al. (2021). Establishing community reference samples, data and call sets for benchmarking cancer mutation detection using whole-genome sequencing. Nat. Biotechnol. *39*, 1151–1160. https://doi.org/10.1038/s41587-021-00993-6.
31. Schipper, L.J., Samsom, K.G., Snaebjornsson, P., Battaglia, T., Bosch, L.J.W., Lalezari, F., Priestley, P., Shale, C., van den Broek, A.J., Jacobs, N., et al. (2022). Complete genomic characterization in patients with cancer of unknown primary origin in routine diagnostics. ESMO Open *7*, 100611. https://doi.org/10.1016/j.esmoop.2022.100611.
32. Samsom, K.G., Schipper, L.J., Roepman, P., Bosch, L.J., Lalezari, F., Klompenhouwer, E.G., de Langen, A.J., Buffart, T.E., Riethorst, I., Schoenmaker, L., et al. (2022). Feasibility of whole-genome sequencing-based tumor diagnostics in routine pathology practice. J. Pathol. *258*, 179–188. https://doi.org/10.1002/path.5988.
33. Samsom, K.G., Bosch, L.J.W., Schipper, L.J., Roepman, P., de Bruijn, E., Hoes, L.R., Riethorst, I., Schoenmaker, L., van der Kolk, L.E., Retèl, V.P., et al. (2020). Study protocol: Whole genome sequencing implementation in standard diagnostics for every cancer patient (wide). BMC Med. Genom. *13*, 169. https://doi.org/10.1186/s12920-020-00814-w.
34. Horak, P., Klink, B., Heining, C., Gröschel, S., Hutter, B., Fröhlich, M., Uhrig, S., Hübschmann, D., Schlesner, M., Eils, R., et al. (2017). Precision oncology based on omics data: The nct heidelberg experience. Int. J. Cancer *141*, 877–886. https://doi.org/10.1002/ijc.30828.
35. Worst, B.C., van Tilburg, C.M., Balasubramanian, G.P., Fiesel, P., Witt, R., Freitag, A., Boudalil, M., Previti, C., Wolf, S., Schmidt, S., et al. (2016). Next-generation personalised medicine for high-risk paediatric cancer patients – the inform pilot study. Eur. J. Cancer *65*, 91–101. https://doi.org/10.1016/j.ejca.2016.06.009.
36. Lejeune, C., Amado, I.F., DEFIDIAG study group FHU Translad and Aviesan; Binquet, C., Deleuze, J.-F., Delmas, C., Dollfus, H., Esperou, H., Faivre, L., Frebourg, T., et al. (2022). Valuing genetic and genomic testing in france: current challenges and latest evidence. J. Community Genet. *13*, 477–485. https://doi.org/10.1007/s12687-020-00503-2.
37. Lévy, Y. (2016). Genomic medicine 2025: France in the race for precision medicine. Lancet *388*, 2872. https://doi.org/10.1016/s0140-6736(16)32467-9.
38. Solary, E., Blanc, P., Boutros, M., Girvalaki, C., Locatelli, F., Medema, R.H., Nagy, P., and Tabernero, J. (2022). UNCAN.eu, a european initiative to UNderstand CANcer. Cancer Discov. *12*, 2504–2508. https://doi.org/10.1158/2159-8290.cd-22-0970.
39. Bates, M. (2022). The cancer moonshot enters a new phase. IEEE Pulse *13*, 2–5. https://doi.org/10.1109/mpuls.2022.3227807.
40. Rehm, H.L., Page, A.J.H., Smith, L., Adams, J.B., Alterovitz, G., Babb, L.J., Barkley, M.P., Baudis, M., Beauvais, M.J.S., Beck, T., et al. (2021). Ga4gh: International policies and standards for data sharing across genomic research and healthcare. Cell Genom. *1*, 100029. https://doi.org/10.1016/j.xgen.2021.100029.
41. Priestley, P., Baber, J., Lolkema, M.P., Steeghs, N., de Bruijn, E., Shale, C., Duyvesteyn, K., Haidari, S., van Hoeck, A., Onstenk, W., et al. (2019). Pan-cancer whole-genome analyses of metastatic solid tumours. Nature *575*, 210–216. https://doi.org/10.1038/s41586-019-1689-y.
42. Tikellis, G., Dwyer, T., Paltiel, O., Phillips, G.S., Lemeshow, S., Golding, J., Northstone, K., Boyd, A., Olsen, S., Ghantous, A., et al. (2018). The international childhood cancer cohort consortium (i4c): A research platform of prospective cohorts for studying the aetiology of childhood cancers. Paediatr. Perinat. Epidemiol. *32*, 568–583. https://doi.org/10.1111/ppe.12519.
43. Zook, J.M., Catoe, D., McDaniel, J., Vang, L., Spies, N., Sidow, A., Weng, Z., Liu, Y., Mason, C.E., Alexander, N., et al. (2016). Extensive sequencing of seven human genomes to characterize benchmark reference materials. Sci. Data *3*, 160025. https://doi.org/10.1038/sdata.2016.25.
44. adamewing, L.,M., Rapsssito, Mjko1210, SebastianHollizeck, Xia, C., Cook, D.E., GILLET-Markowska, A., Richter, D., Hammerbacher, J., St. John, J., et al. (2021). Dami Rebergen, and Zhmz90. BAMSurgeon, URL:. https://doi.org/10.5281/ZENODO.5116421
45. Chakravarty, D., Gao, J., Phillips, S.M., Kundra, R., Zhang, H., Wang, J., Rudolph, J.E., Yaeger, R., Soumerai, T., Nissan, M.H., et al. (2017). Oncokb: A precision oncology knowledge base. JCO Precis. Oncol. *2017*, 1–16. https://doi.org/10.1200/po.17.00011.
46. Suehnholz, S.P., Nissan, M.H., Zhang, H., Kundra, R., Nandakumar, S., Lu, C., Carrero, S., Dhaneshwar, A., Fernandez, N., Xu, B.W., et al. (2024). Quantifying the expanding landscape of clinical actionability for patients with cancer. Cancer Discov. *14*, 49–65. https://doi.org/10.1158/2159-8290.cd-23-0467.

47. Gajiwala, K.S., Wu, J.C., Christensen, J., Deshmukh, G.D., Diehl, W., DiNitto, J.P., English, J.M., Greig, M.J., He, Y.-A., Jacques, S.L., et al. (2009). Kit kinase mutants show unique mechanisms of drug resistance to imatinib and sunitinib in gastrointestinal stromal tumor patients. Proc. Natl. Acad. Sci. USA *106*, 1542–1547. https://doi.org/10.1073/pnas.0812413106.

48. Heinrich, M.C., Blanke, C.D., Druker, B.J., and Corless, C.L. (2002). Inhibition of kit tyrosine kinase activity: A novel molecular approach to the treatment of kit-positive malignancies. J. Clin. Oncol. *20*, 1692–1703. https://doi.org/10.1200/jco.2002.20.6.1692.

49. Bauer, S., Duensing, A., Demetri, G.D., and Fletcher, J.A. (2007). Kit oncogenic signaling mechanisms in imatinib-resistant gastrointestinal stromal tumor: Pi3-kinase/akt is a crucial survival pathway. Oncogene *26*, 7560–7568. https://doi.org/10.1038/sj.onc.1210558.

50. Zehir, A., Benayed, R., Shah, R.H., Syed, A., Middha, S., Kim, H.R., Srinivasan, P., Gao, J., Chakravarty, D., Devlin, S.M., et al. (2017). Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. Nat. Med. *23*, 703–713. https://doi.org/10.1038/nm.4333.

51. Lo-Coco, F., Avvisati, G., Vignetti, M., Thiede, C., Orlando, S.M., Iacobelli, S., Ferrara, F., Fazi, P., Cicconi, L., Di Bona, E., et al. (2013). Retinoic acid and arsenic trioxide for acute promyelocytic leukemia. N. Engl. J. Med. *369*, 111–121. https://doi.org/10.1056/nejmoa1300874.

52. Wang, H., Nicolay, B.N., Chick, J.M., Gao, X., Geng, Y., Ren, H., Gao, H., Yang, G., Williams, J.A., Suski, J.M., et al. (2017). The metabolic function of cyclin d3–cdk6 kinase in cancer cell survival. Nature *546*, 426–430. https://doi.org/10.1038/nature22797.

53. Garcia-Prieto, C.A., Martínez-Jiménez, F., Valencia, A., and Porta-Pardo, E. (2022). Detection of oncogenic and clinically actionable mutations in cancer genomes critically depends on variant calling tools. Bioinformatics *38*, 3181–3191. https://doi.org/10.1093/bioinformatics/btac306.

54. Cunningham, F., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., et al. (2015). Ensembl 2015. Nucleic Acids Res. *43*, D662–D669. https://doi.org/10.1093/nar/gku1010.

55. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics *25*, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324.

56. Huang, W., Li, L., Myers, J.R., and Marth, G.T. (2012). ART: a next-generation sequencing read simulator. Bioinformatics *28*, 593–594. https://doi.org/10.1093/bioinformatics/btr708.

57. McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The ensembl variant effect predictor. Genome Biol. *17*, 122. https://doi.org/10.1186/s13059-016-0974-4.

58. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The sequence alignment/map format and SAMtools. Bioinformatics *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

59. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with bowtie 2. Nat. Methods *9*, 357–359. https://doi.org/10.1038/nmeth.1923.

60. Kim, D., Paggi, J.M., Park, C., Bennett, C., and Salzberg, S.L. (2019). Graph-based genome alignment and genotyping with hisat2 and hisat-genotype. Nat. Biotechnol. *37*, 907–915. https://doi.org/10.1038/s41587-019-0201-4.

61. Narzisi, G., Corvelo, A., Arora, K., Bergmann, E.A., Shah, M., Musunuri, R., Emde, A.-K., Robine, N., Vacic, V., and Zody, M.C. (2018). Genome-wide somatic variant calling using localized colored de bruijn graphs. Commun. Biol. *1*, 20. https://doi.org/10.1038/s42003-018-0023-9.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Tumorized genomes	This paper	ENA: PRJEB68324
Mosaic genomes (PCAWG-Pilot)	This paper	https://docs.icgc-argo.org/docs/data-access/icgc-25k-data; pilot50-mosaic
Mosaic genome (HMF-12)	This paper	EGA: EGAD50000000668
Compendium of cancer driver genes (release 2023.05.31)	IntOGen[2]	https://www.intogen.org/download
Human gene annotation transcriptome (GRCh37.87)	Ensembl-EBI[54]	https://ftp.ensembl.org/pub/grch37/current/gff3/homo_sapiens/Homo_sapiens.GRCh37.87.gff3.gz
PCAWG-Pilot	PCAWG[4]	https://docs.icgc-argo.org/docs/data-access/icgc-25k-data
HMF-12	Cameron et al.[9]	N/A
NA12878	GIAB[43]	https://github.com/genome-in-a-bottle/giab_data_indexes/
HG002	GIAB[43]	https://github.com/genome-in-a-bottle/giab_data_indexes/
PCAWG Consensus Callsets	PCAWG[4]	https://docs.icgc-argo.org/docs/data-access/icgc-25k-data
Software and algorithms		
ONCOLINER (and PipelineDesigner)	This paper	https://github.com/EUCANCan/oncoliner; https://zenodo.org/doi/10.5281/zenodo.12755026
GenomeVariator	This paper	https://github.com/Computational-Genomics-BSC/GenomeVariator; https://zenodo.org/doi/10.5281/zenodo.12755117
GenomeMosaicMaker	This paper	https://github.com/Computational-Genomics-BSC/GenomeMosaicMaker; https://zenodo.org/doi/10.5281/zenodo.12755101
VariantExtractor	This paper	https://github.com/EUCANCan/variant-extractor; https://zenodo.org/doi/10.5281/zenodo.12755170
BWA-MEM	Li et al.[55]	https://github.com/lh3/bwa
ART Illumina	Huang et al.[56]	https://www.niehs.nih.gov/research/resources/software/biostatistics/art
VEP	McLaren et al.[57]	https://github.com/Ensembl/ensembl-vep; https://doi.org/10.1186/s13059-016-0974-4
BAMSurgeon	Edwin et al.[25,44] Lee et al.[26]	https://github.com/adamewing/bamsurgeon; https://doi.org/10.5281/zenodo.7019232
Other		
Variant callers list and containers	This paper	https://github.com/EUCANCan/variant-callers
ONCOLINER HTML report	This paper	http://cg.bsc.es/cg/data/oncoliner_report.html

METHOD DETAILS

Benchmarking datasets: Mosaic and tumorized genomes

Experimentally validated variants are the best solution to properly benchmark variant calling algorithms and pipelines. Validated variant sets are highly valuable for providing realistic data in various fields, especially for cancer research. However, due to the difficulty of the validation process samples with validated variants are scarce. Additionally, existing validated datasets tend to be large and require computationally intensive processing due to the inclusion of numerous samples, even when they contain few validated variants. This computational demand poses challenges for researchers working with limited resources. To address this issue, we produced two alternative approaches in Mosaic and Tumorized genomes. We also provide the necessary open-source tools to generate them, GenomeVariator and GenomeMosaicMaker, intended for researchers interested in using these approaches for their studies.

First, Mosaic genomes provide a condensed representation of the complete benchmarking datasets, conveying the same information but significantly decreasing computational burden. Nevertheless, Mosaic genomes pose limitations. For once, the availability of restricted access benchmarking datasets due to privacy concerns or other confidentiality clauses impedes open access to Mosaic genomes. Additionally, source benchmarking datasets often lack control over false positives, affecting reliability for precision assessment on Mosaic genomes. To complement Mosaics and overcome these limitations, the Tumorized genomes were developed. Tumorized genomes address the need for patient data protection by ensuring that the genomes are de-identified, eliminating the need for bureaucratic processes that can impede progress in cancer research. By utilizing Tumorized genomes, researchers gain absolute control over the features and variants included in these datasets, allowing precise and controlled analysis. The method to bring these approaches into real datasets for our study is detailed below.

The short-read sequencing data needed to discover somatic variants in a sample consists of a pair of tumor-normal samples mapped to a reference genome. The ground for constructing the tumor-normal datasets comprising a Mosaic genome are the original reads sequenced from validated datasets such as the PCAWG-Pilot and HMF-12, and the validated variants within them. The first step is to estimate the average depth, read length, and insert sizes of each of the validated samples. Then, a canvas WGS dataset is simulated with these values from the same reference genome to which the samples are mapped. For our read simulations, we used the ART Illumina software.[56] Finally, read-alignments from the original tumor-normal samples overlapping with a 2kb window centered around each of the validated variants are inserted into the canvas genome. The simulated reads in these regions from the canvas genome are discarded in order to remove discrepancies in read depths or lengths. Thus, Mosaic genomes provide a realistic representation of the original somatic variants. We implemented this method as the GenomeMosaicMaker tool.

The HMF-12 samples contain a collection of experimentally validated short indels and SVs. These experiments were carried out after variant discovery using GRIDSS2, Manta, and Strelka.[9] Only SVs were used to construct the HMF-12 Mosaic genome because short indel calls presented inconsistent coordinates between the variant callers mentioned earlier. Then, we analyzed the features of the reads from each sample finding that all of them are homogeneous regarding sequencing depth, and lengths of reads and inserts. Thus, we generated one single Mosaic compiling all HMF-12 samples into a canvas, simulated from the GRCh37 reference with median (Mdn) read length values of 150 bp, and Mdn insert length of 500 bp with a standard deviation (SD) of 125 bp. The read depth of the normal Mosaic sample was 32x, and the tumor sample was 100x.

The PCAWG-Pilot data originates from different laboratories using different sequencing protocols and machines. Although the variant discovery and validation pipelines were homogenized, the original experimental conditions made it impossible to compile all samples into a single Mosaic. To find a consistent number of Mosaic genomes for this data, the following procedure was performed. First, we estimated the insert size median (Mdn) and standard deviation (SD) values for each sample. Using this information, we performed K-means clustering iteratively trying different cluster numbers to find the most consistent grouping for these samples. Then, for each configuration, we executed the variant callers and computed their performance metrics. These results were compared to their performances on the original samples, proving that the best results were obtained with the three Mosaic representations. Finally, these three Mosaics were generated with read depths of 40x for the normal samples and 60x for the tumor samples (see Figures S3 and S4). To select the better suited mapping method for the benchmarking and further evaluate the potential effect of different read aligners on ONCOLINER recommendations, the assessment process was performed over Mosaic genomes aligned with three state-of-the-art algorithms in BWA-MEM (v.0.7.17),[58] Bowtie2 (v.2.5.3),[59] and Hisat2 (v.2.2.1).[60] Based on these results BWA-MEM was chosen for the generation of the final Mosaics (see Figure S16). Overall, the construction process of the Mosaics from both datasets highlights the importance of careful control of sequencing features such as insert sizes, depths, and even mapping strategies to ensure the reliability of the Mosaic genome to represent real variation from the original samples.

In contrast to the Mosaic genome approach, a Tumorized genome provides a benchmark to test the precision of variant discovery. To overcome data sharing limitations, publicly available read mappings are used as the base. First, they are carefully split into the tumor-normal sets, ensuring that there are no duplicated reads between them, and that they come from different libraries, mimicking real scenarios where normal and tumoral samples are processed and sequenced independently. To reach the desired coverage for the benchmarking genome, balanced random downsampling is applied to maintain the original proportions of the read libraries. Then, for each variant in the validated VCF, a subset of reads in the tumoral sample is modified to represent the variant. The number of modified reads depends on the depth of the region where the variant is located and the provided VAF (see Figure S6). These modified reads are remapped to the original reference genome. Therefore, the Tumorized genome will only contain somatic variants that

were inserted into the reads (see Figure S17), eliminating external false positives while avoiding artificial variations in read depth. This method is implemented in GenomeVariator, a wrapper tool that enhances the functionalities of BAMSurgeon[25,26,44] to facilitate the construction of Tumorized genomes.

The Tumorized genomes (60x Tumoral - 40x Normal) used in this study were constructed from 300x depth Illumina reads from two real WGS samples from the GIAB project: NA12878 and AshkenazimTrio son HG002,[43] mapped to the GRCh37 reference using BWA-MEM.[55] The SNV, indel, and SV-validated variants from the PCAWG Consensus Callsets were simulated into the tumoral sample. To ensure the Tumorized genome contains a realistic amount of variants, the number of simulated variants of each type was set to the median number of total variants of the same type in the original samples, discovered by all callers. The number of SVs was set to double the median and overlapping SVs were excluded to avoid undesired results, prioritizing by SV type (SV type priority order: translocation, inversion, duplication, and deletion).

Variant caller selection

We originally selected 18 candidate programs for somatic variant identification, covering different calling algorithms, strategies, and scopes (SNV, indels, and SVs). From these 12 were selected for our study based on usability and efficiency: SvABA[5] (version 1.1.0), Delly[6] (version 1.1.6), mutect2 (from GATK[7] 4.2.6.1), Strelka2[8] (version 2.9.10), GRIDSS2[9] (version 2.13.2), MuSE[11] (version 2.0), Manta[10] (version 1.6.0), SAGE[21] (version 3.0), cgpPindel[12] (version 3.9.0), cgpCaVEManWrapper[13] (version 1.16.0), Shimmer[14] and BRASS[22] (version 6.3.4). The other 6 variant callers were excluded due to: excessive execution time, non-standard input requirements, not maintained, or with merged germline-somatic variants outputs. The excluded tools were Lancet[61] (version 1.1.0), Platypus[15] (version 0.8.1.1), Lumpy[16] (version 0.3.1), breakdancer[17] (version 1.4.5), SomaticSniper[18] (version 1.0.5.0) and Seurat[19] (version 2.5).

Assessment module

The ONCOLINER assessment module compares the discovered variants from the results of the input pipelines against the validated variants. Comparisons depend on the variant types and sizes. In particular, SNVs and small indels are considered true positives if their chromosome, coordinate, and alternate allele exactly match with a ground truth variant. To consider SVs as true positives their breakends must be located within a 100 bp window of the breakends of a gold standard SV, and their orientation must be equal. Conversely, even though large insertions are not present in the Mosaic or Tumorized datasets, their length would be compared instead of their second breakend. The assessment module counts true positive, false positive, and false negative calls from the evaluated pipeline to estimate recall, and precision using the Mosaics and Tumorized genomes, respectively. Then, the F-score is computed to provide a combined accuracy measure (see Figures S9–S11).

To assess the functional performance of the pipelines, every true positive variant inherits precomputed gene annotations of their matched gold standard variant. This avoids the need to annotate every incoming test VCF. The same method is used to identify affected cancer-driver genes. VEP[57] was used to annotate SNVs and indels, avoiding annotations without functional phenotypes. For SV annotation, VEP could not process multiple variants due to their VCF representation. To properly annotate SVs we developed scripts based on VariantExtractor, where each gene that intersected the span of an SV was added to the corresponding VCF record. In particular, for inversions and translocations, only the genes that intersected their breakpoints were considered as affected thereby disrupting the open reading frame of the gene. Finally, the source for the annotation of protein-coding genes to SVs was the Ensembl GHRCh37 transcriptome (v87),[54] and the cancer driver genes were collected from the IntOGen Catalog (release 2023.05.31).[2]

Improvement module

The ONCOLINER improvement module follows the assessment step and provides recommendations based on the performance evaluation of the input pipelines and the 12 selected variant callers. Specifically, the recommendations are the best combinations of variant callers to integrate into the pipeline to maximize performance metrics. The first step is to perform both the union and the intersection of the pipeline calls with the callers. Then, performance metrics are calculated for these merged results, and combinations are sorted based on them. In addition to a list of all possible combinations provided as a CSV file, the interactive GUI shows the most relevant recommendations. This allows the user to sort them by any of the metrics between recall, precision, F1-score, or even by the number of affected protein-coding or cancer-driver genes. To improve visualization of the most relevant recommendations, they are filtered by selecting the one in the top 5% for each performance metric, prioritizing those with the least number of callers. This follows the rationale that a better recommendation minimizes the cost of adding too many tools to a pipeline and the effort of going through redundant combinations (see Figure S18).

Harmonizer module

The ONCOLINER harmonizer module follows the improvement step generating recommendations to bring the performance of all input pipelines closer while maintaining the best possible performance. To achieve this, pipeline heterogeneity is quantified into two metrics, the Performance Heterogeneity Score (PHS) and Gene Discordance Ratio (GDR). The PHS is estimated from plotting precision-recall ranges in a Euclidean space, where pipelines are represented by their respective performance coordinates. Considering p_i (recall, precision) as a pipeline point from the set of n pipelines, a centroid (i.e., the gravity center) c is the point that minimizes the sum of its square distance to all p_i. Then, the PHS is computed as the normalized mean of all d_i. In other words, the PHS is the

normalized average Euclidean distance from the position of each pipeline to the centroid or the theoretical maximum homogeneity point. Equation 1 shows the calculation of the PHS.

$$PHS = \frac{\sqrt{2}}{n}\sum_{i=1}^{n} d_i, \text{where } d_i = \sqrt{\left(p_i^{recall} - c^{recall}\right)^2 + \left(p_i^{precision} - c^{precision}\right)^2} \quad \text{(Equation 1)}$$

To measure functional impact heterogeneity in variant discovery we created the Gene Discordance Ratio (GDR). The GDR is calculated as the complement of the proportion between the variant-affected genes found by every pipeline (intersection) over the total number of variant-affected genes even if only detected by one of the pipelines (union). Hence, a GDR value closer to 1 would imply a high level of heterogeneity in functional impact between the pipelines. This metric follows Equation 2 where G_i represents the number of genes affected by the discovered variants from the i-th pipeline.

$$GDR = 1 - \frac{|G_1 \cap G_2 \cap G_3 \ldots \cap G_n|}{|G_1 \cup G_2 \cup G_3 \ldots \cup G_n|} \quad \text{(Equation 2)}$$

To achieve comprehensive harmonization, this module first prioritizes the best recommendations to maximize performance for each of the pipelines. Then, it minimizes PHS and GDR to decrease heterogeneity in accuracy and functional impact. This priority order avoids optimized homogenization where the distance from a pipeline to the centroid could be 0, but would sometimes worsen accuracy. As the harmonization module works with the performance enhancements generated by the improvement module, it will highlight recommendations where heterogeneity decreases, but most importantly where performance improves in all possible homogenization scenarios. Finally, following the filtering logic of the improvement step, the user visualizes non-redundant recommendations for harmonizing the pipelines that minimize the effort of adding too many variant callers (see Figure S19; Table S2).

Cell Genomics, Volume *4*

Supplemental information

ONCOLINER: A new solution for monitoring, improving, and harmonizing somatic variant calling across genomic oncology centers

Rodrigo Martín, Nicolás Gaitán, Frédéric Jarlier, Lars Feuerbach, Henri de Soyres, Marc Arbonés, Tom Gutman, Montserrat Puiggròs, Alvaro Ferriz, Asier Gonzalez, Lucía Estelles, Ivo Gut, Salvador Capella-Gutierrez, Lincoln D. Stein, Benedikt Brors, Romina Royo, Philippe Hupé, and David Torrents

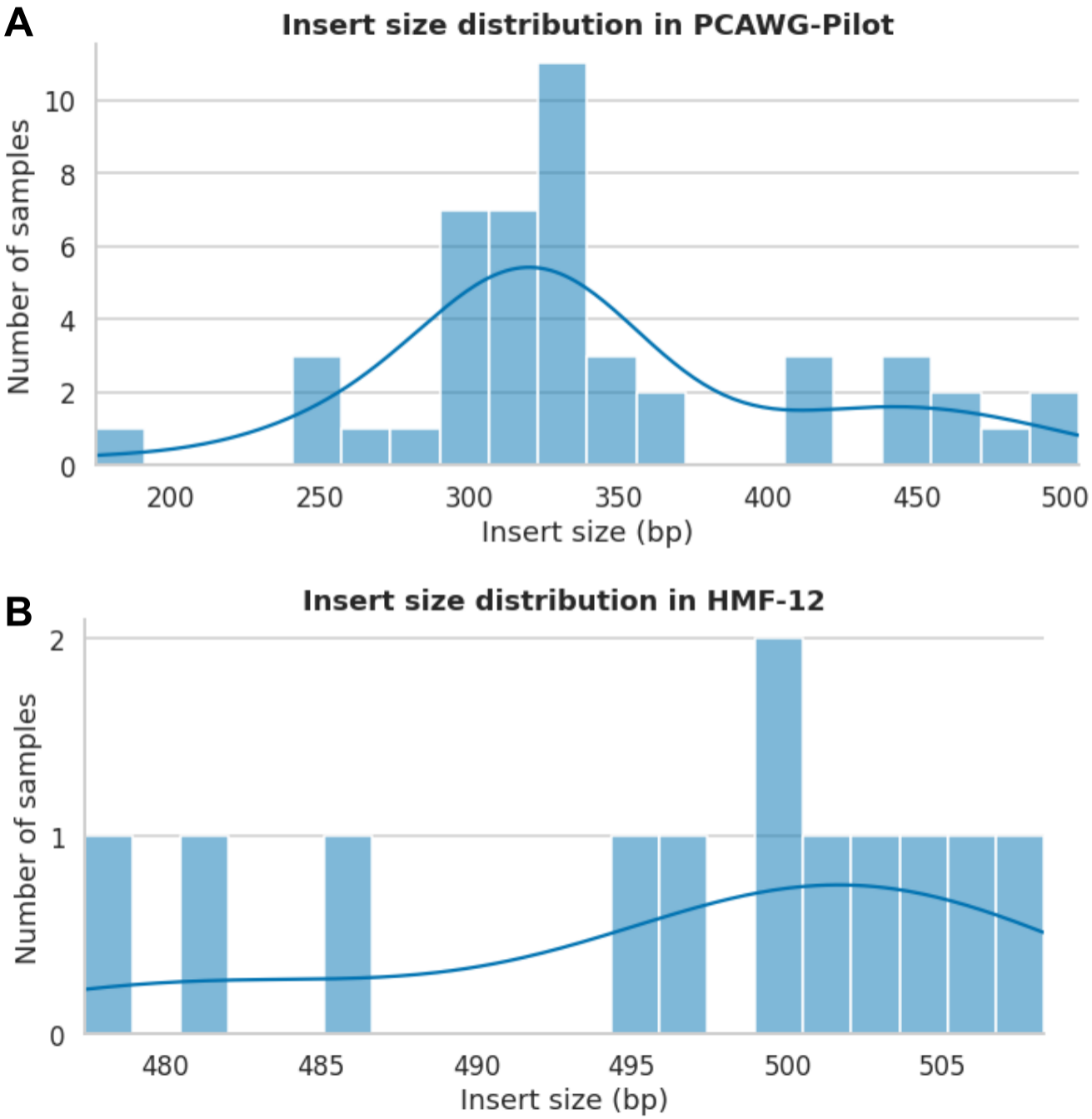


Figure S1: **Insert size distributions for all samples from PCAWG-Pilot and HMF-12, related to Figure 1.** A) PCAWG-Pilot insert sizes. Analysis of this distribution allowed us to find a bimodal sample, which was excluded from the experiments. B) HMF-12 samples show consistent insert size homogeneity.

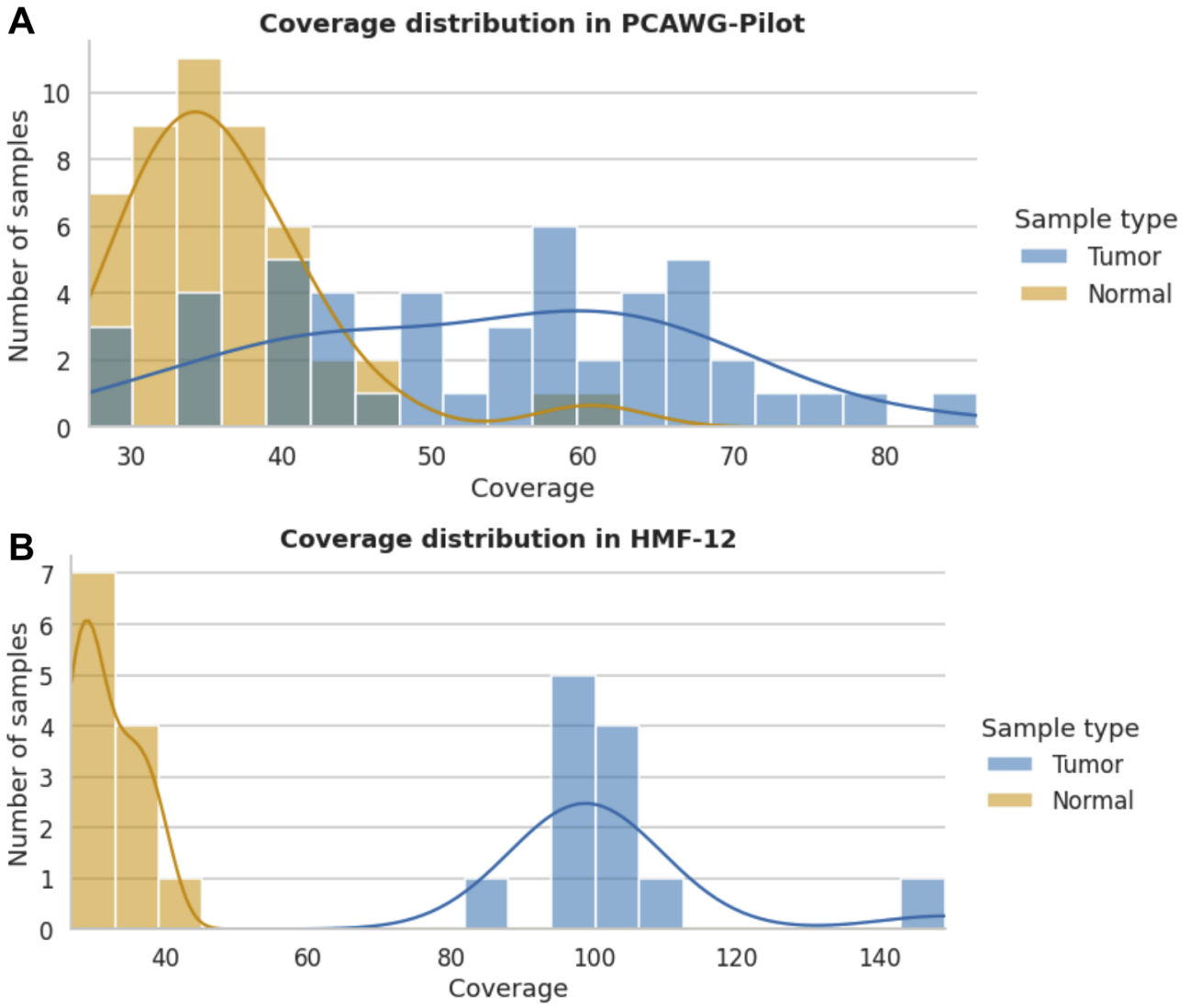


Figure S2: **Coverage distribution in the tumor-normal datasets, related to Figure 1.** A) PCAWG-Pilot samples and B) HMF-12 samples.

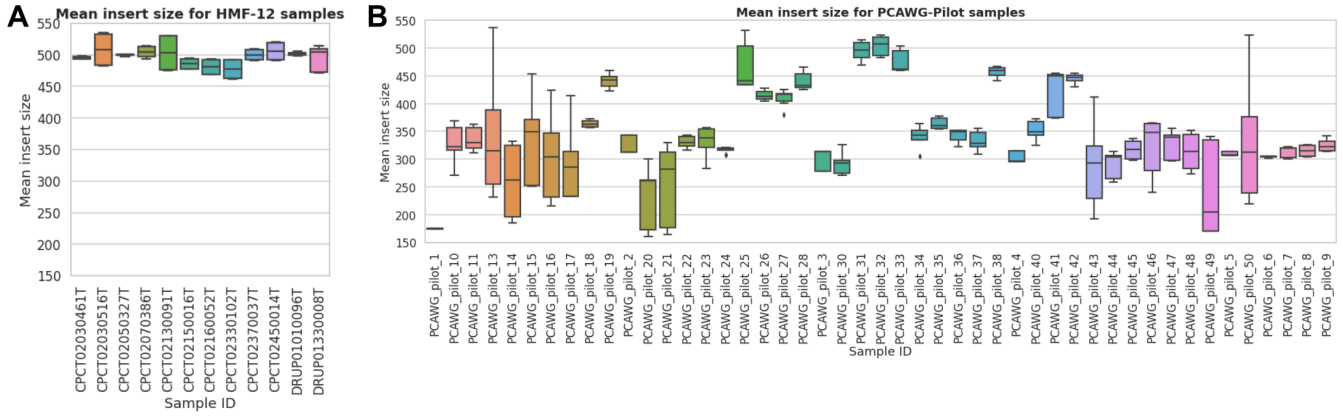


Figure S3: **Insert size boxplots samples in the HMF-12 and PCAWG-Pilot datasets (see Methods).** A) The HMF-12 samples showcase similar insert sizes where all the average values are between approximately a 50bp range. B) Conversely, there is a high level of heterogeneity between insert sizes for each sample in the PCAWG-Pilot dataset, with many even showing significantly bigger variances than others. Mean insert sizes also range from as low as 200 bp to 500 bp.

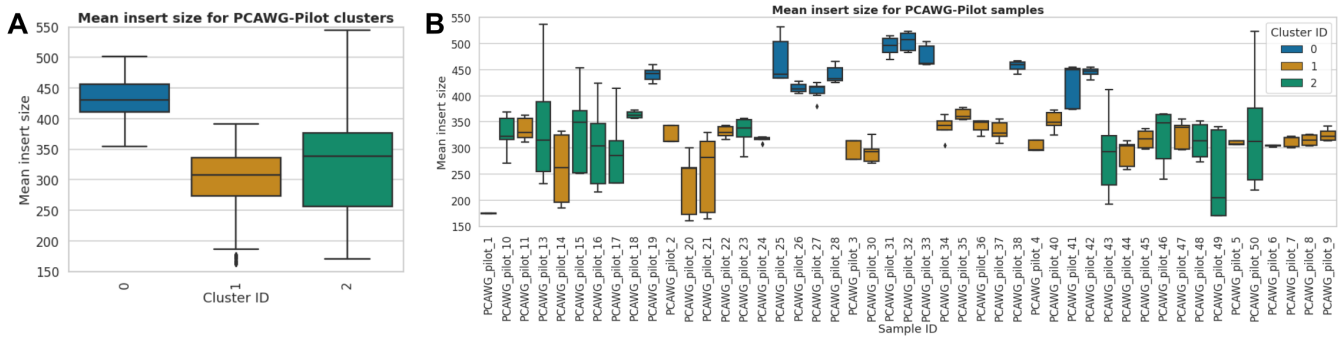


Figure S4: **Insert sizes of sample clusters from the PCAWG-Pilot datasets (see Methods).** A) The three clusters grouping the PCAWG-Pilot samples show low variance for insert sizes, where 0 and 1 range mostly between 50 bp - 100 bp, and 2 depicts a higher variance concordant with the original samples included in it. Nevertheless, the values between the first and third quartiles are limited to a range of at most 150 bp. B) The insert sizes are shown for each sample, where the color indicates the clustering conformation.

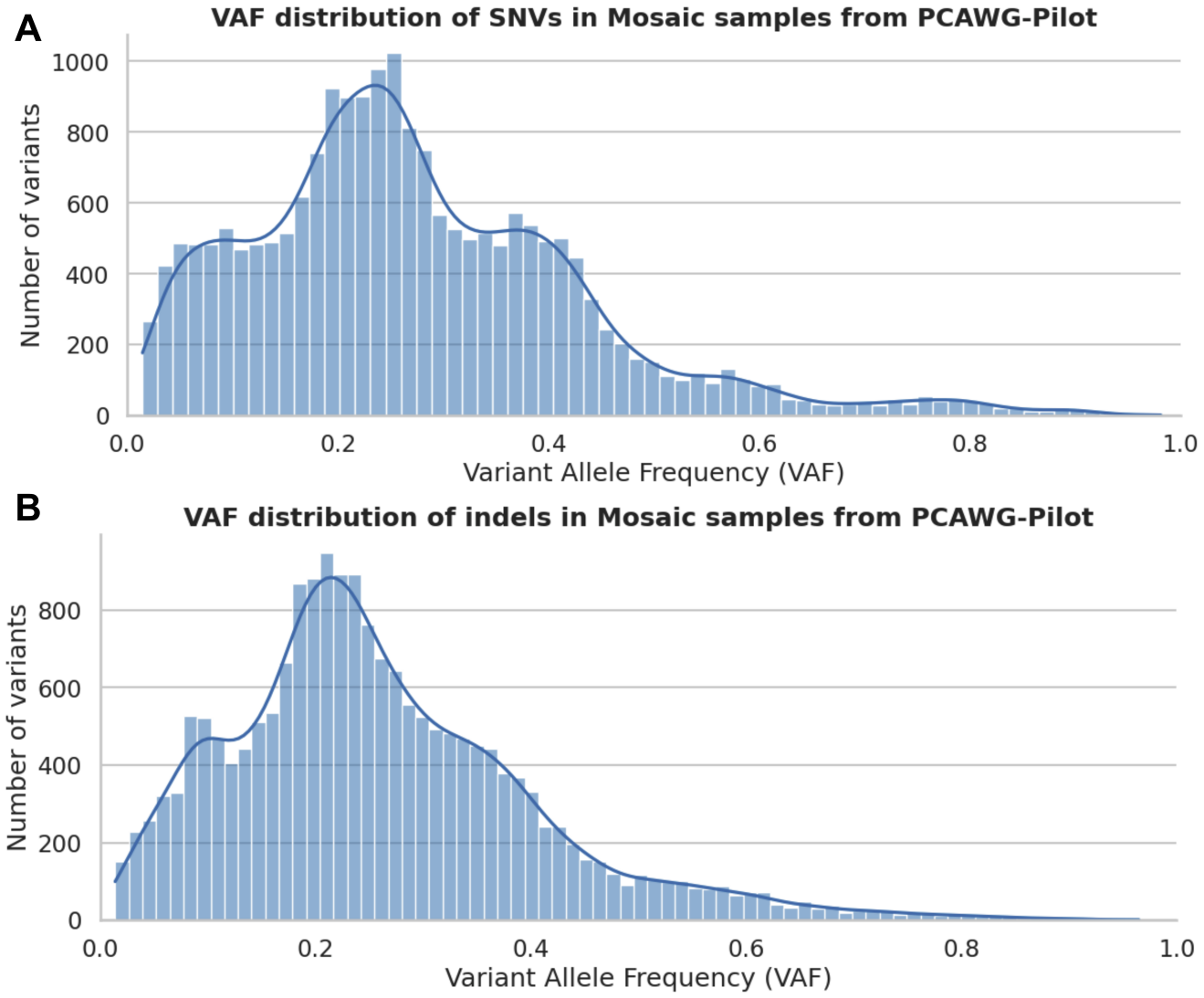


Figure S5: **Variant Allele Frequency (VAF) distributions for the PCAWG-Pilot gold standard variants included in this study, related to Figure 1.** A) Variant Allele Frequency (VAF) distributions for SNVs and B) Indels from the PCAWG-Pilot variant collection included in the mosaic genomes. VAF values shown here follow a multimodal distribution, not produced by an underlying biological cause but by the criterion for selecting these variants, which was the certainty of them being true calls.

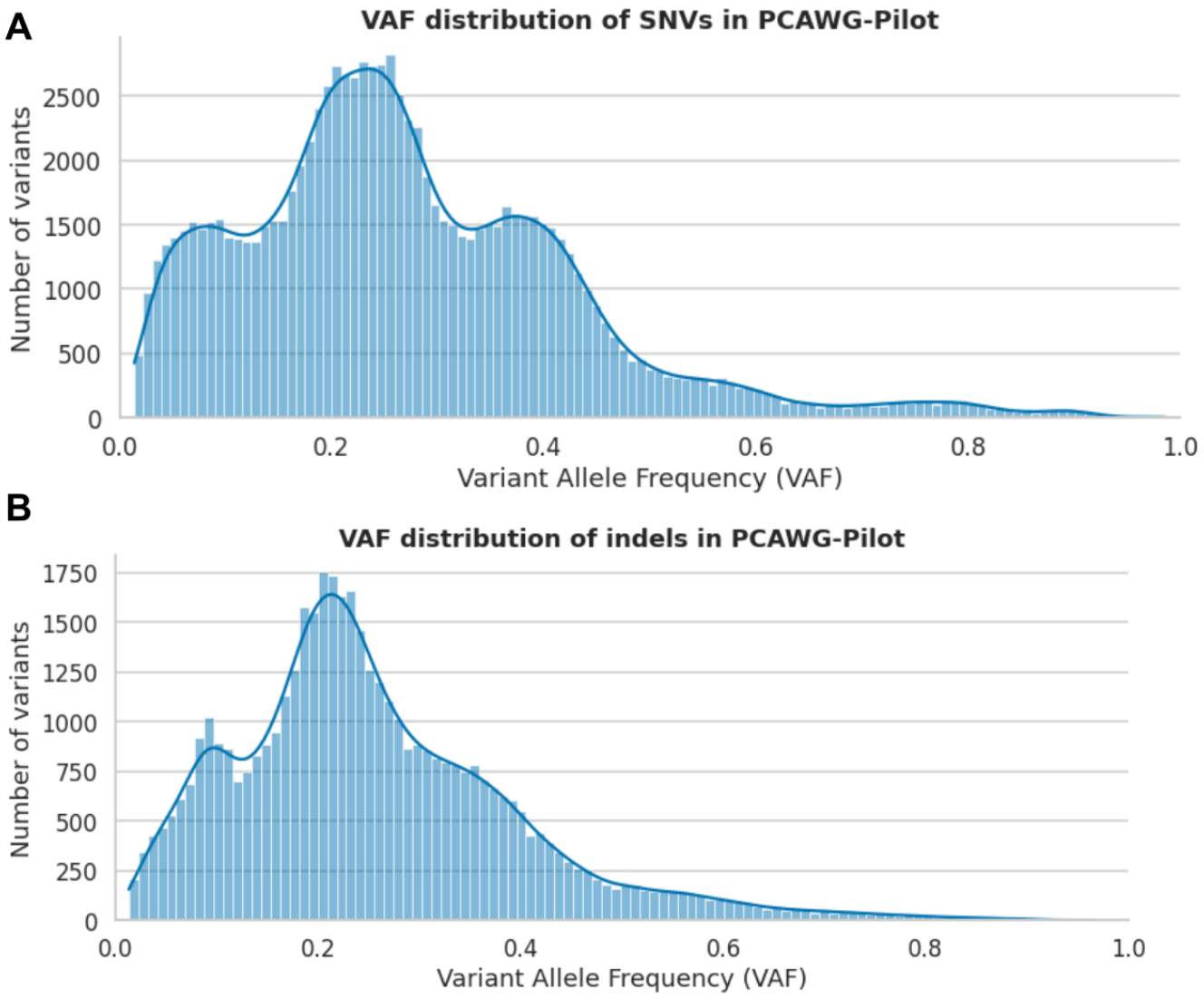


Figure S6: **Variant Allele Frequency Distributions for all variants in the PCAWG-Pilot datasets (including samples not used in the study), related to Figure 1.** A) shows the VAFs for SNVs following a multimodal distribution due to the original selection bias of the validation process, with the most values aggregated around 0.2. B) shows the distribution for indels with a similar trend to the SNVs. Due to the difficulties in SV VAF estimation, these values were not present in either the PCAWG-Pilot or the HMF-12 Gold Standard VCFs.

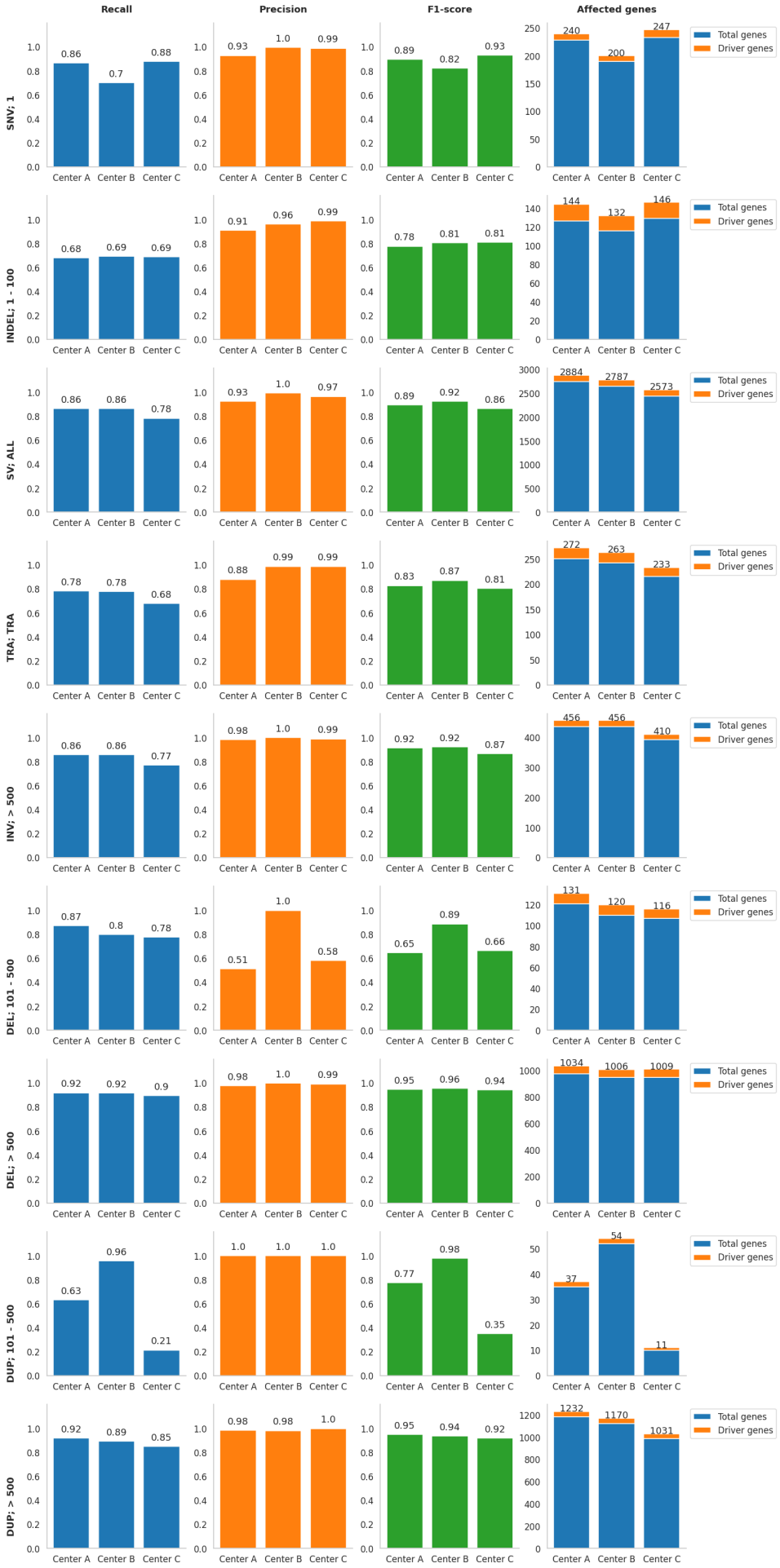

Recall
Precision
F1-score
Affected genes
SNV; 1
INDEL; 1 - 100
SV; ALL
TRA; TRA
INV; > 500
DEL; 101 - 500
DEL; > 500
DUP; 101 - 500
DUP; > 500
Center A
Center B
Center C
Total genes
Driver genes
0.86 0.7 0.88
0.93 1.0 0.99
0.89 0.82 0.93
240 200 247
0.68 0.69 0.69
0.91 0.96 0.99
0.78 0.81 0.81
144 132 146
0.86 0.86 0.78
0.93 1.0 0.97
0.89 0.92 0.86
2884 2787 2573
0.78 0.78 0.68
0.88 0.99 0.99
0.83 0.87 0.81
272 263 233
0.86 0.86 0.77
0.98 1.0 0.99
0.92 0.92 0.87
456 456 410
0.87 0.8 0.78
0.51 1.0 0.58
0.65 0.89 0.66
131 120 116
0.92 0.92 0.9
0.98 1.0 0.99
0.95 0.96 0.94
1034 1006 1009
0.63 0.96 0.21
1.0 1.0 1.0
0.77 0.98 0.35
37 54 11
0.92 0.89 0.85
0.98 0.98 1.0
0.95 0.94 0.92
1232 1170 1031

Figure S7: **Results from benchmarking on the mosaic and tumorized genomes for pipelines from centers A, B, and C (see Methods).** Rows show the specific variant type dissected by subtype and size range for SVs, and columns showcase performance and functional measures including (in order) Recall, Precision, F1 Score, and affected genes. The latter metric is additionally broken into two categories by the color code of protein-coding and cancer-driver genes. The y-axis of each panel shows the respective measure unit of the metric (proportions for performance and integer counts for genes) and the x-axis displays the centers. Results for SNVs show how different research centers adapted to their specific needs. Center B produced the lowest recall value of the three (70%) but did not generate false positives. Center C shows the best performance in this category with a better recall (88%) while maintaining a close precision (99%). This recall improvement translates into functional impact, as C also captures the most gene-altering SNVs. A produced lower recall and precision than C, showing it has an important margin for improvement. These results attest to the recall-precision trade-off, especially for B, where a lower detection threshold is likely to produce more false positive calls. Still, a conservative method should produce better precision at the cost of variant discovery. Short indel calling showed more homogeneous results for the three centers with lower performances from center A. Although it had a slightly worse recall than B (A: 68%, B: 69%), it detected more indels that functionally affected genes (A: 148, B: 136) which evidences that some pipelines may better identify variants in non-repeat regions. Overall SV-calling heterogeneity does not differ greatly in comparison to the experiments on SNVs and indels, with exceptions in certain size ranges of specific SV types. Detecting all SVs, A presents the lowest precision (93%) while C suffers from the lowest recall (78%). In turn, B shows the best results overall achieving near-perfect precision (100%) while keeping the highest recall (86%) which translates into an F1 score of 92%. The two most radical examples of differing results are deletions and duplications with lengths in the 100 - 500 bp range. For these deletions, the maximum difference of recall is only 10%, but precision varies from 51% for A to a value of 100% for B. C presents the lowest recall (87%) and a low precision (58%) closer to A (51%). Pipeline B had substantially better results for detecting these deletions with an F1 score of 89%. Although A shows poor performance on precision, it had the highest recall (87%) which allows it to capture 11 more deletion-affected genes than the closest center B. For 100-500bp duplications, no pipeline reports false positives, but recall values differ substantially. Similar to the observations from deletions, B reported a much bigger recall (96%) compared to A (63%) and to the worst-performing C (21%). This significantly affected the detection of duplication-affected genes where the difference in the raw counts from B to C is 43, meaning C was not able to capture most of the functional impact of these mid-sized duplications. Both centers A and B detected most of the inversions (86% recall) while center C underperformed in this category, as evidenced by a 9% lower recall (77%). This difference in recall could be evidenced by C missing 46 genes affected by inversion breakpoints in comparison to A and B. Nevertheless, the three pipelines proved to be highly precise, shown by values close to 100% precision. Translocation calling is fairly similar across all centers, but none was achieved with a recall bigger than 78%. Finally, deletions and duplications with lengths above 500bp show homogeneous results, where differences in recall are consistent with those in detected SV-affected genes.

Table S1: **Optimal pipelines generated by PipelineDesigner, related to Figure 4.**

Variant type	Optimal combination	F1-score
SNV	[mutect2 (from GATK 4.2.6.1) ∩ Strelka (v22.9.10)] ∪ SAGE (v3.0)	93%
Indel	[mutect2 (from GATK 4.2.6.1) ∩ Strelka (v22.9.10)] ∪ SAGE (v3.0)	87%
SV	[Delly (v1.1.6) ∩ Manta (v1.6.0)] ∪ GRIDSS (v2.13.2)	93%

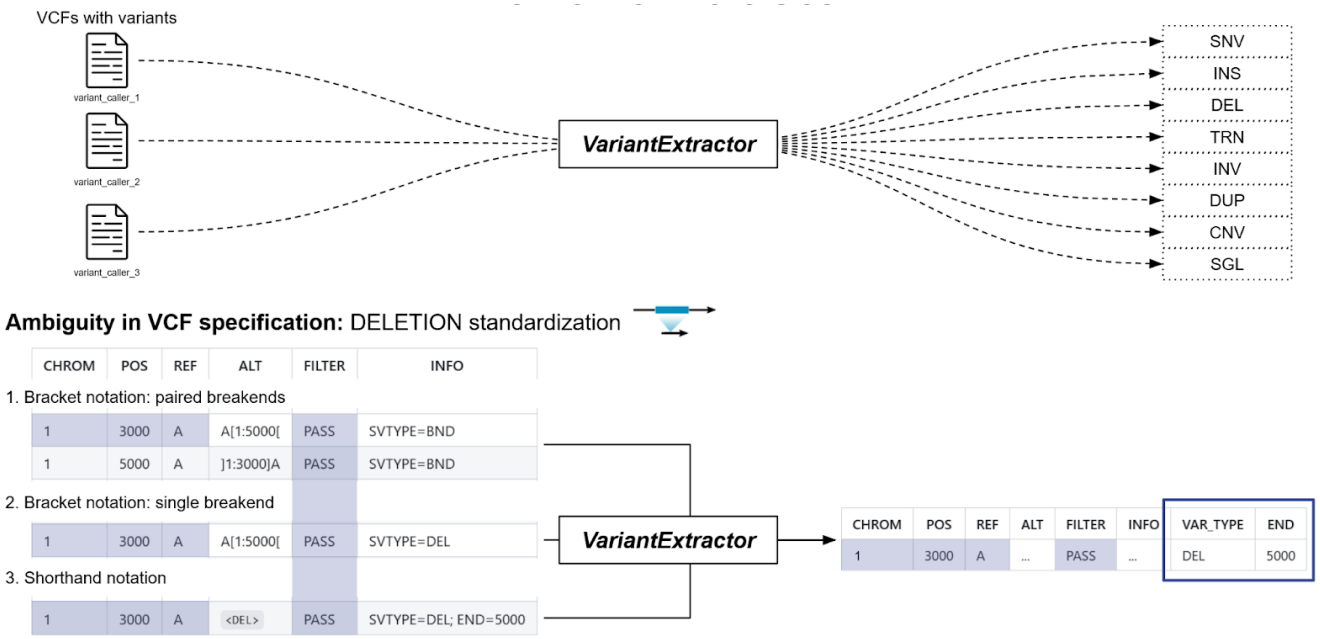


Figure S8: **VariantExtractor module functionality diagram, related to Figure 4.** The figure shows the functionality of VariantExtractor. VariantExtractor takes input VCF files and consumes VCF Records as multiple types of variants interpreting them as specific objects keeping all their information fields. One of the main advantages of this utility is the interpretation of breakends (BND) into comprehensive SV records with type-specific representations. BNDs are the most common way to represent SVs in short-read-based variant calling, due to the inherent limitations of this sequencing technology. VariantExtractor solves this issue by applying homogenization rules to the input BND records thereby removing ambiguity in the interpretation of these variants. This is especially useful for variants recorded in the bracket notation. This advantage is shown in the example of the figure. This deletion is represented in multiple ways from different variant calling pipelines. VariantExtractor standardizes the deletion information from bracket notations (paired or single), and shorthand notations into a comprehensive representation that can be used for variant analysis, or to produce a new VCF with a different representation.

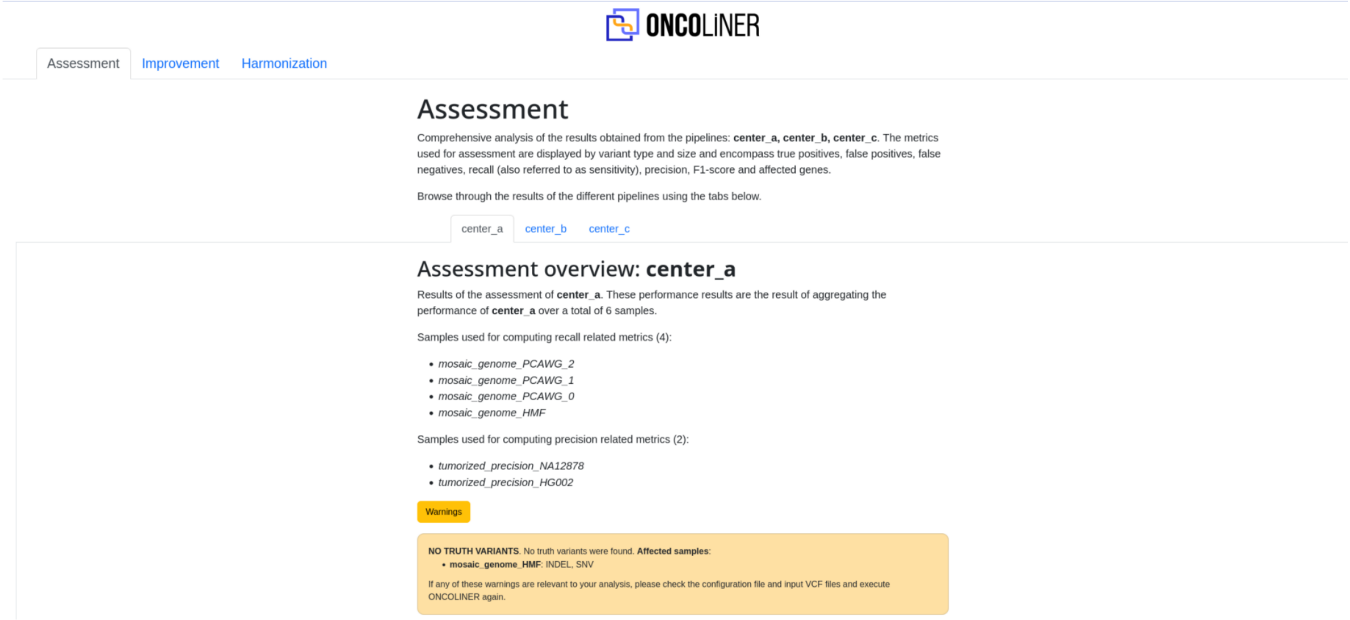
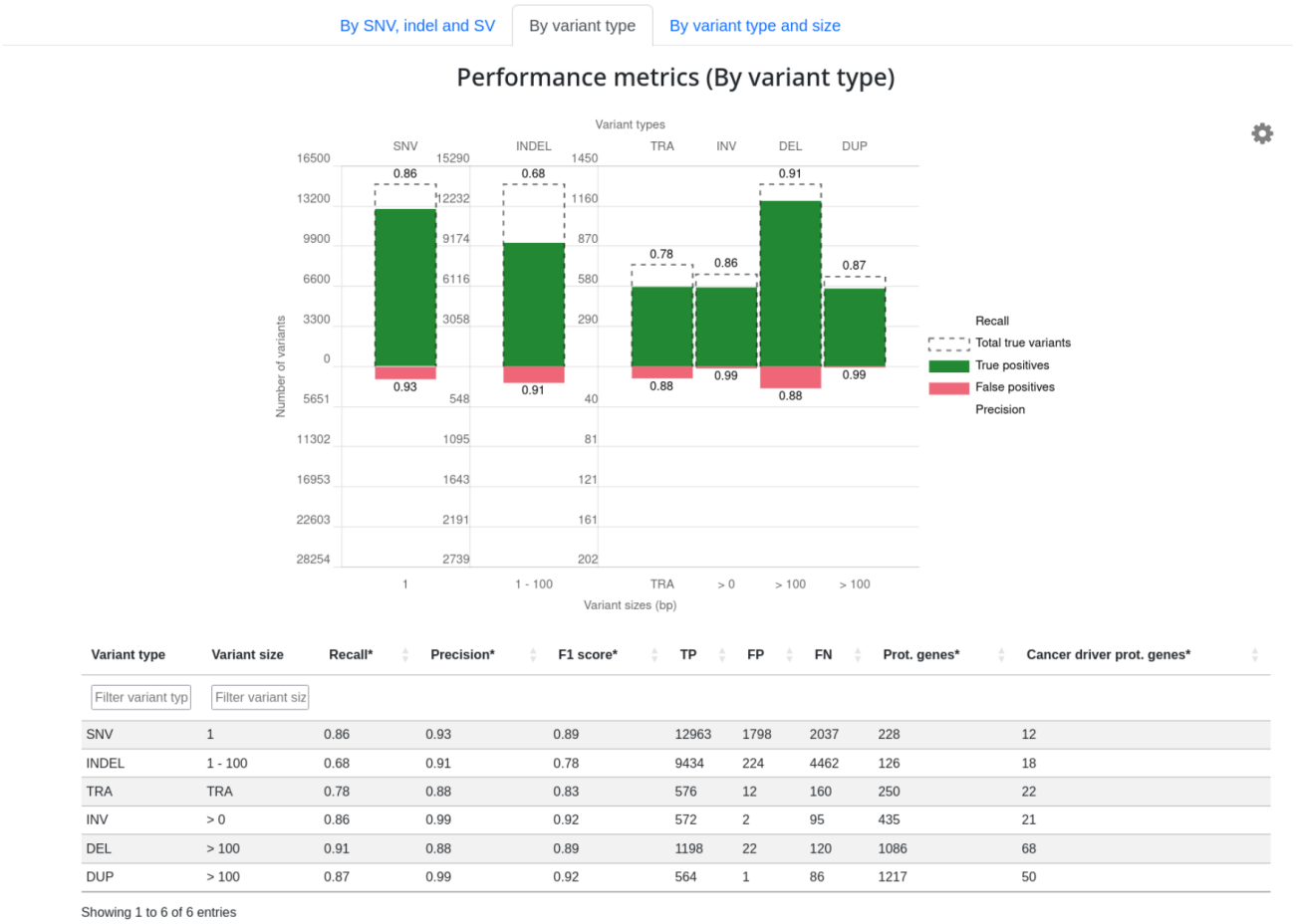


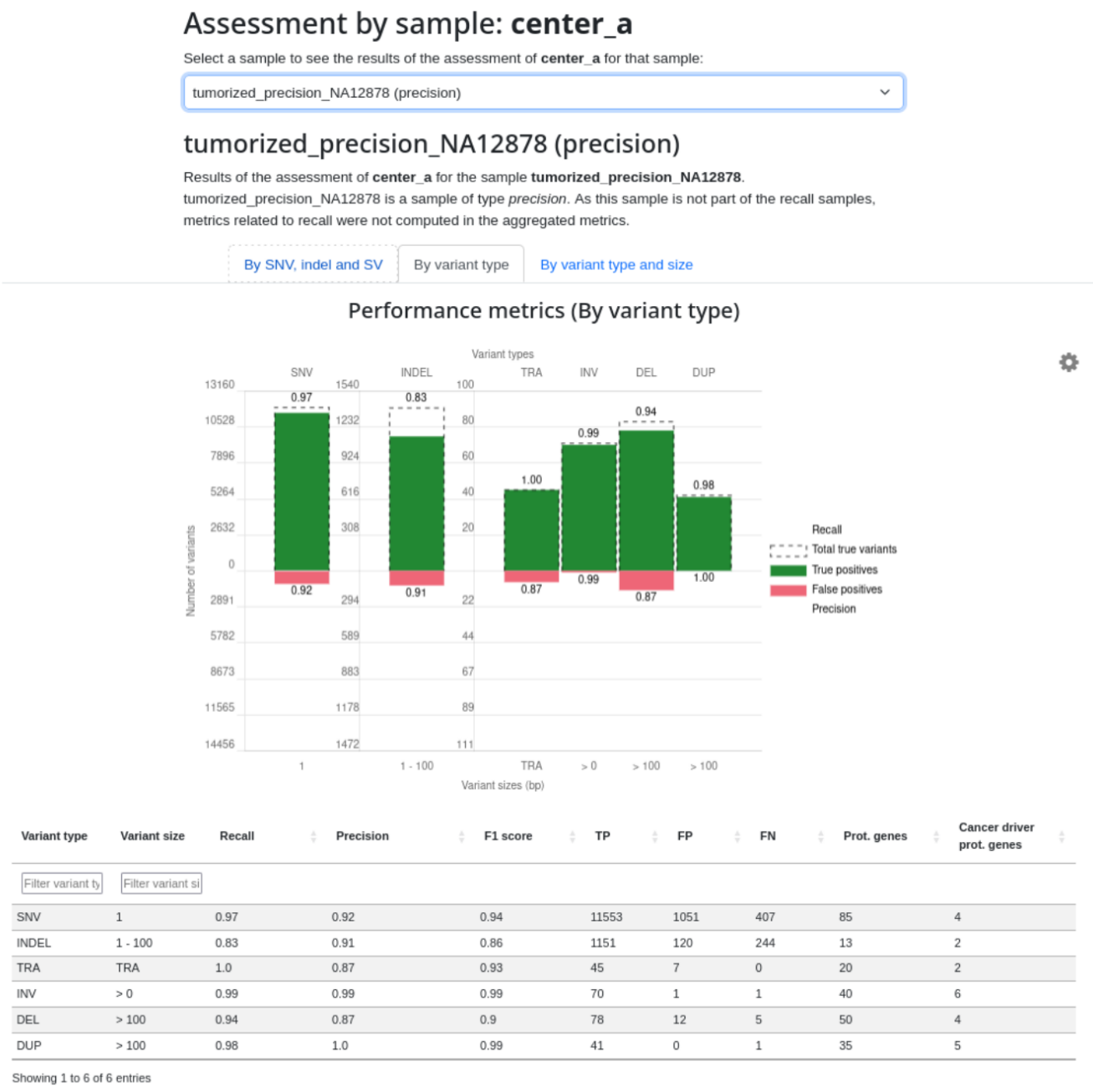
Figure S9: **Screenshot of the assessment module description section (see Methods).** Interactive results are provided in HTML format. This report is composed of different sections. First, the visualization of the results is described by enumerating the input pipelines and the benchmarking genomes. As can be seen in the image, the warning button provides the user with useful information when discrepancies or lacking information are found in the VCF inputs. The samples used as the gold standard to calculate precision and recall are also displayed.



Variant type	Variant size	Recall*	Precision*	F1 score*	TP	FP	FN	Prot. genes*	Cancer driver prot. genes*
Filter variant typ	Filter variant siz								
SNV	1	0.86	0.93	0.89	12963	1798	2037	228	12
INDEL	1 - 100	0.68	0.91	0.78	9434	224	4462	126	18
TRA	TRA	0.78	0.88	0.83	576	12	160	250	22
INV	> 0	0.86	0.99	0.92	572	2	95	435	21
DEL	> 100	0.91	0.88	0.89	1198	22	120	1086	68
DUP	> 100	0.87	0.99	0.92	564	1	86	1217	50

Showing 1 to 6 of 6 entries

Figure S10: **Graphic section of the assessment module, showcasing performance metrics (see Methods).** This section displays the results in terms of performance metrics (recall, precision), and true positive-false positive variant counts, as a cohesive figure that can be exported and used for further publications by the user. These results can be plotted by variant types in SNVs, Indels, and SVs, broken by SV types, or further dissected in these categories plus SV size ranges. These options also consistently modify a table that shows all of these values accordingly, found just below the figure panel. Additionally, this table showcases the counts for protein-coding and cancer-driver genes affected by true positive variants. The display can be modified according to the criteria shown in the tabs. The table can be filtered or sorted according to the column variables.



Variant type	Variant size	Recall	Precision	F1 score	TP	FP	FN	Prot. genes	Cancer driver prot. genes
SNV	1	0.97	0.92	0.94	11553	1051	407	85	4
INDEL	1 - 100	0.83	0.91	0.86	1151	120	244	13	2
TRA	TRA	1.0	0.87	0.93	45	7	0	20	2
INV	> 0	0.99	0.99	0.99	70	1	1	40	6
DEL	> 100	0.94	0.87	0.9	78	12	5	50	4
DUP	> 100	0.98	1.0	0.99	41	0	1	35	5

Figure S11: **Graphics of the assessment on individual benchmarking samples (see Methods).** The last panel of this section displays similar figures plotted according to the selected benchmarking dataset, in this case, the results for precision assessment over only one of the tumorized genomes. This may also be useful for a hypothetical user who wants to evaluate if using one benchmarking genome over another impacts the quality of variant calling of their pipelines.

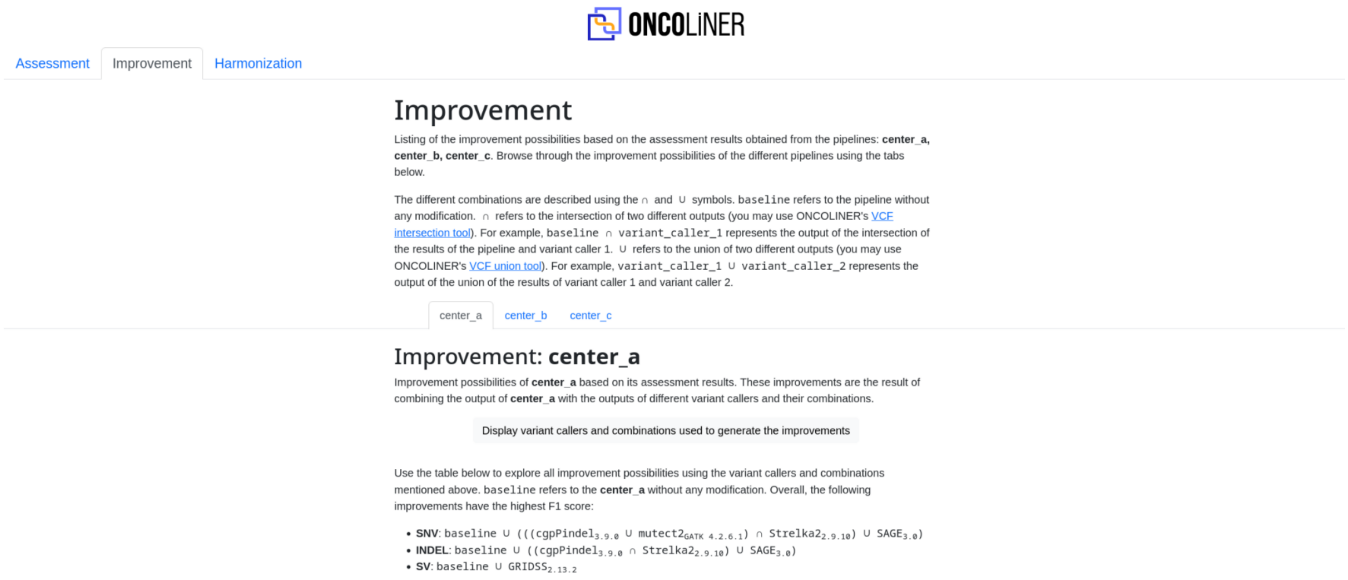


Figure S12: **Description of the improvement module (see Methods).** To facilitate the rapid improvement of a specific pipeline, the best recommendations according to the F1 score by variant category are highlighted here.

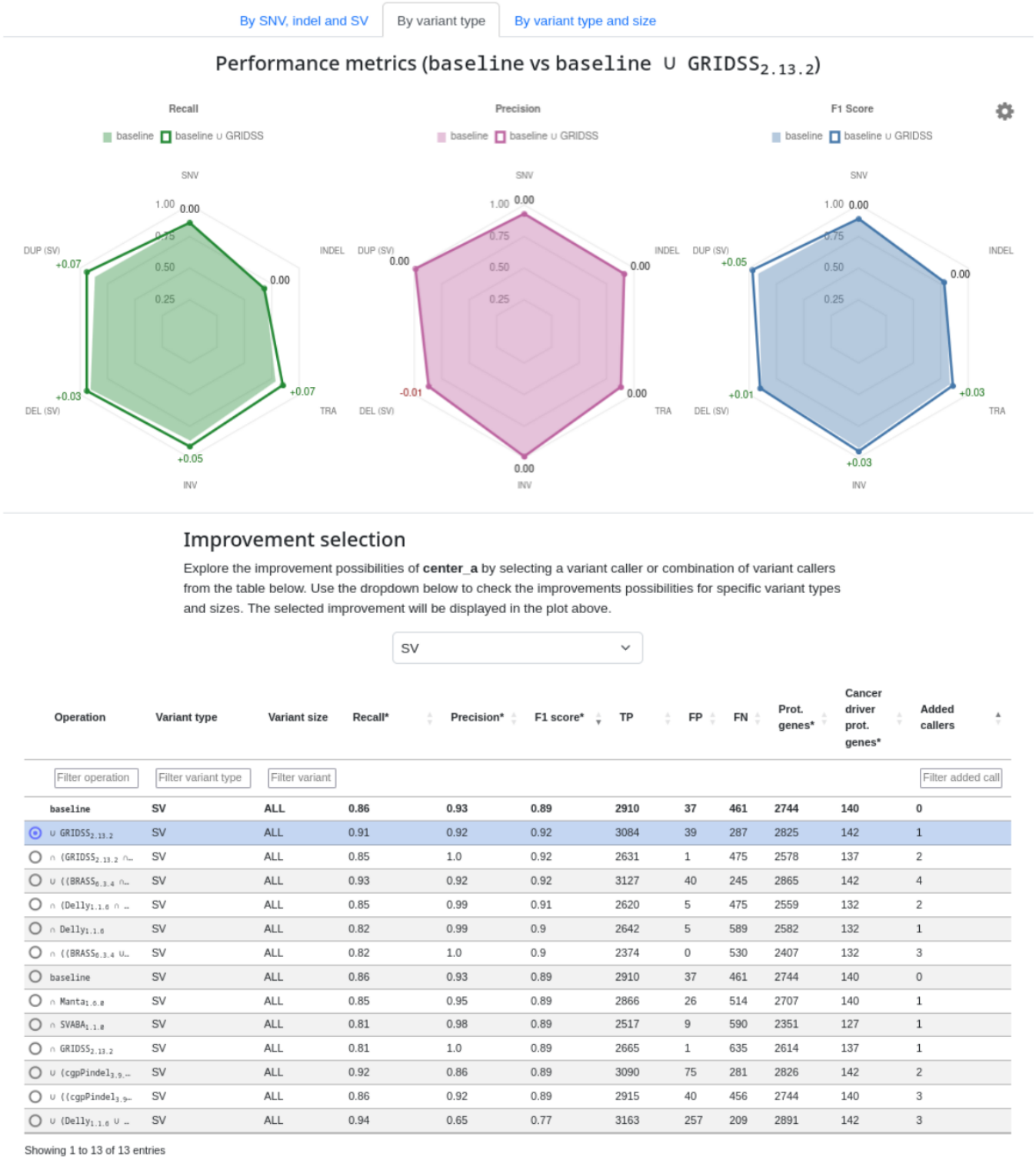


Figure S13: **Graphic results of the improvement solutions (see Methods).** The graphical results of the improvement recommendations are displayed for each one of the individual input pipelines as performance figures, showing the baseline performance, the improved value, and the difference in percentage for recall, precision, and F1-Score. Recommendations can be categorized by variant type, SV type, and size by choosing one of these options. The user chooses which recommendation is displayed in the result figures. The results table can be sorted by any criteria between the performance metrics, counts, affected protein-coding or cancer-driver genes, or even by the number of added variant callers.

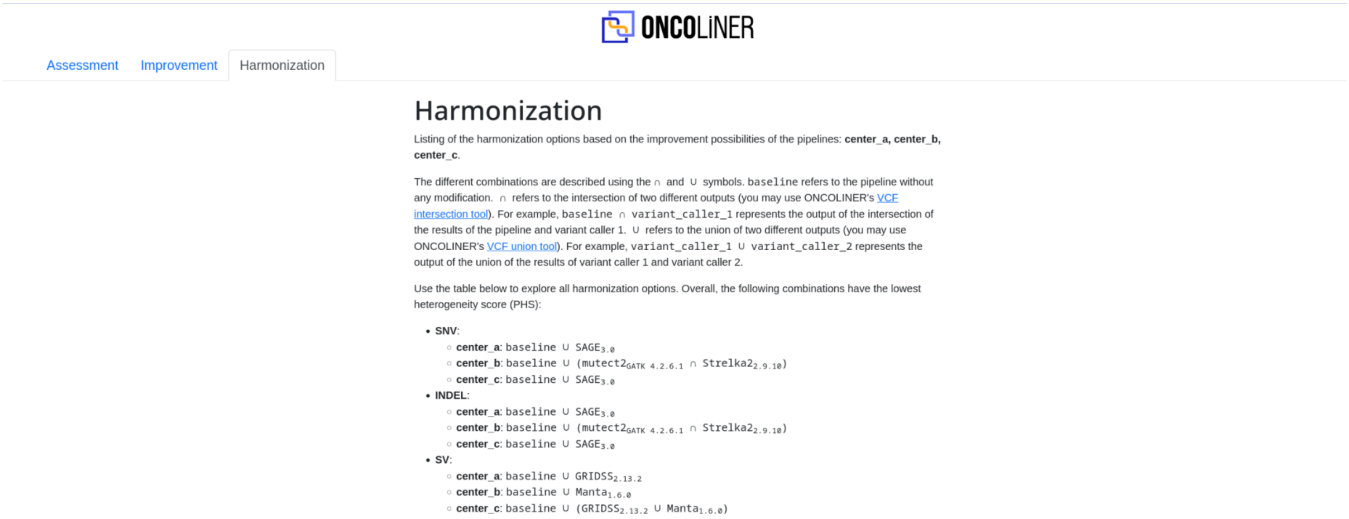
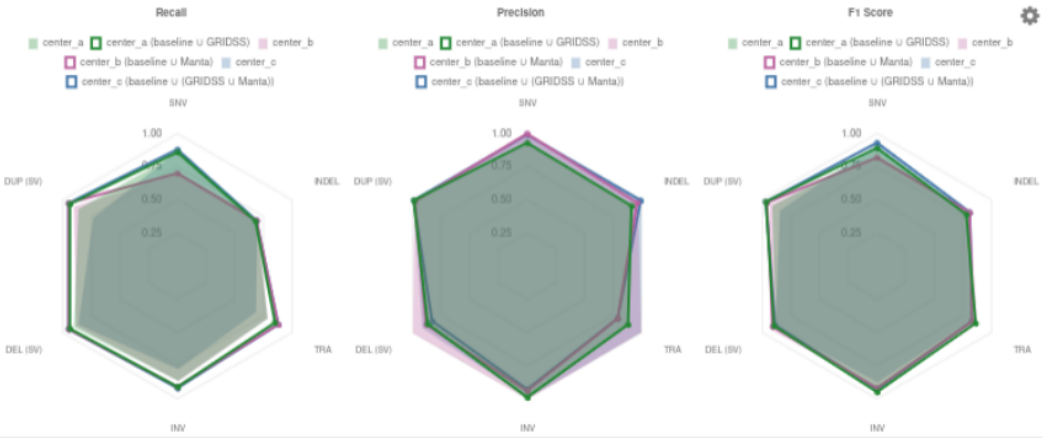


Figure S14: **Screenshot of the description section for the harmonization tab (see Methods).** The harmonization tab is the last element of the output report and follows a similar structure to the improvement tab. The description found first describes the top recommendation for harmonizing each type of variant for the input pipelines according to the lowest Performance Heterogeneity Score (PHS), in addition to the standard description of the results. For each variant category, the best combinations between the user pipelines and the recommended callers are shown to allow easy access to the best results of the harmonization functionality, based on improving accuracy and minimizing heterogeneity.

By SNV, indel and SV | By variant type | By variant type and size

Performance metrics (baseline vs harmonization)



Harmonization selection

Explore the harmonization options by selecting any row from the table below. Use the dropdown below to check the harmonization options for specific variant types and sizes. The selected harmonization will be displayed in the plot above.

SV

center_a	center_b	center_c	Variant type	Variant size	PHS	Recall avg.*	Precision avg.*	F1 score avg.	GDR	Prot. genes avg.*	Cancer driver prot. genes avg.*	Total added callers
Filter center_a	Filter center_b	Filter center_c	Filter variant t	Filter v								Filter tota
baseline	**baseline**	**baseline**	**SV**	**ALL**	**0.07**	**0.84**	**0.96**	**0.89**	**0.18**	**2613.33**	**134.67**	**0**
∪ GRIDSS$_{2.13.2}$	∪ Manta$_{1.6.0}$	∪ (GRIDSS$_{2.13.2}$ ∪…	SV	ALL	0.03	0.92	0.9	0.91	0.02	2837.67	142.0	4
∪ GRIDSS$_{2.13.2}$	∪ Manta$_{1.6.0}$	∪ Manta$_{1.6.0}$	SV	ALL	0.03	0.91	0.9	0.9	0.04	2812.33	142.0	3
∪ GRIDSS$_{2.13.2}$	∪ (∣BRASS$_{6.3.4}$ ∪…	∪ GRIDSS$_{2.13.2}$	SV	ALL	0.03	0.91	0.95	0.93	0.05	2802.33	141.0	6
∪ GRIDSS$_{2.13.2}$	∪ (∣BRASS$_{6.3.4}$ ∩…	∪ GRIDSS$_{2.13.2}$	SV	ALL	0.04	0.9	0.96	0.93	0.03	2806.0	141.67	5
∪ GRIDSS$_{2.13.2}$	∪ Manta$_{1.6.0}$	∪ GRIDSS$_{2.13.2}$	SV	ALL	0.04	0.91	0.93	0.92	0.03	2818.0	141.67	3
∪ GRIDSS$_{2.13.2}$	∪ [GRIDSS$_{2.13.2}$ ∪…	∪ GRIDSS$_{2.13.2}$	SV	ALL	0.04	0.91	0.92	0.92	0.03	2824.67	141.67	4
baseline	∪ Manta$_{1.6.0}$	∪ Manta$_{1.6.0}$	SV	ALL	0.04	0.89	0.9	0.9	0.05	2785.33	141.33	2
∪ GRIDSS$_{2.13.2}$	baseline	∪ GRIDSS$_{2.13.2}$	SV	ALL	0.04	0.89	0.96	0.92	0.07	2758.67	139.67	2
baseline	baseline	∪ GRIDSS$_{2.13.2}$	SV	ALL	0.04	0.87	0.96	0.91	0.1	2731.67	139.0	1
∩ Manta$_{1.6.0}$	baseline	∪ GRIDSS$_{2.13.2}$	SV	ALL	0.04	0.87	0.97	0.91	0.1	2719.33	139.0	2
∩ Delly$_{1.1.6}$	baseline	∪ GRIDSS$_{2.13.2}$	SV	ALL	0.04	0.86	0.98	0.91	0.11	2677.67	136.33	2
∩ Delly$_{1.1.6}$	baseline	baseline	SV	ALL	0.04	0.82	0.98	0.89	0.16	2559.33	132.0	1
baseline	∪ Manta$_{1.6.0}$	∪ GRIDSS$_{2.13.2}$	SV	ALL	0.05	0.89	0.93	0.91	0.06	2791.0	141.0	2
baseline	∪ Manta$_{1.6.0}$	∪ (GRIDSS$_{2.13.2}$ ∪…	SV	ALL	0.05	0.9	0.92	0.91	0.06	2793.67	141.33	4
∩ GRIDSS$_{2.13.2}$	baseline	∪ GRIDSS$_{2.13.2}$	SV	ALL	0.05	0.85	0.99	0.91	0.1	2688.33	138.0	2
∩ [Delly$_{1.1.6}$ ∩ …	baseline	baseline	SV	ALL	0.05	0.83	0.98	0.9	0.16	2551.67	132.0	2
∩ Delly$_{1.1.6}$	∪ [cgpPindel$_{3.9.…}$	∩ (Delly$_{1.1.6}$ ∩ …	SV	ALL	0.05	0.82	0.99	0.89	0.19	2511.67	130.0	6
∪ GRIDSS$_{2.13.2}$	baseline	∪ (cgpPindel$_{3.9.…}$	SV	ALL	0.06	0.89	0.94	0.91	0.07	2759.67	139.67	3
baseline	baseline	∪ Manta$_{1.6.0}$	SV	ALL	0.06	0.87	0.93	0.9	0.1	2726.0	139.33	1
baseline	baseline	∪ (cgpPindel$_{3.9.…}$	SV	ALL	0.06	0.87	0.94	0.9	0.1	2732.67	139.0	2
∩ Manta$_{1.6.0}$	baseline	baseline	SV	ALL	0.06	0.83	0.97	0.89	0.17	2601.0	134.67	1
baseline	∪ [cgpPindel$_{3.9.…}$	baseline	SV	ALL	0.06	0.84	0.96	0.89	0.17	2615.33	134.67	2
∩ GRIDSS$_{2.13.2}$	baseline	∩ (GRIDSS$_{2.13.2}$ ∩…	SV	ALL	0.06	0.8	1.0	0.89	0.2	2503.67	131.67	3
∩ (∣BRASS$_{6.3.4}$ ∪…	baseline	∩ (GRIDSS$_{2.13.2}$ ∩…	SV	ALL	0.06	0.81	1.0	0.89	0.22	2434.67	130.0	5
∪ GRIDSS$_{2.13.2}$	baseline	∪ (GRIDSS$_{2.13.2}$ ∪…	SV	ALL	0.07	0.9	0.93	0.91	0.08	2778.33	140.0	3
baseline	baseline	baseline	SV	ALL	0.07	0.84	0.96	0.89	0.18	2613.33	134.67	0
∩ GRIDSS$_{2.13.2}$	∪ (∣BRASS$_{6.3.4}$ ∩…	∩ (∣cgpPindel$_{3.9.…}$	SV	ALL	0.07	0.82	0.99	0.9	0.21	2553.33	133.0	8
∩ GRIDSS$_{2.13.2}$	∪ (∣BRASS$_{6.3.4}$ ∩…	∩ (Delly$_{1.1.6}$ ∩ …	SV	ALL	0.07	0.82	0.99	0.9	0.22	2567.67	133.67	7
baseline	∪ Manta$_{1.6.0}$	baseline	SV	ALL	0.08	0.86	0.93	0.89	0.15	2672.67	136.67	1

Figure S15: **Graphs of the harmonization options for different variant types, displaying the performance metrics for each harmonized pipeline according to the selected recommendations (see Methods).** The figures show the performance metrics from the baseline assessment and the improvements from the chosen harmonization strategy located in the table. To avoid an uninformative and difficult-to-visualize display, performance values and counts are shown as the average between the results of assessing all input pipelines. Two important columns are included in this harmonization tab showing the PHS and Gene Discordance Ratio (GDR) achieved by each recommendation row, which can also be used to sort recommendations

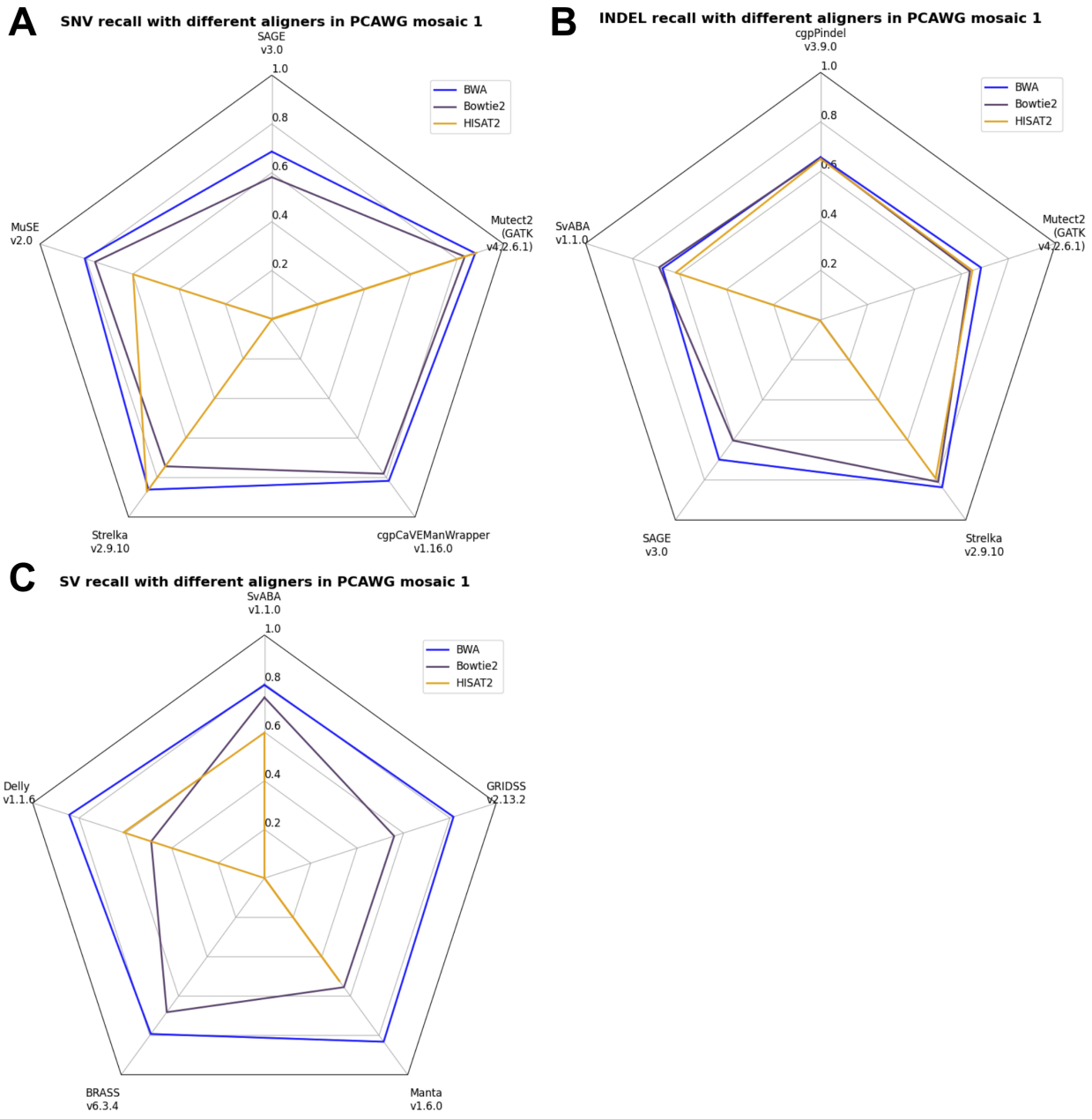


Figure S16: **Performance of all callers over a mosaic genome mapped with different read aligners, per variant type (see Methods).** Color codes portray the recall values for callers on the samples by read-mapper software. A) Shows SNVs, B) indels, and C) SVs. Panel A) and B) show that for SNVs and indels, BWA and Bowtie2 achieved similar sensitivities, but the former allowed callers to further improve on this performance metric. Panel C) shows how all callers benefit from using BWA for calling SVs since they rely on supplementary mappings provided by this tool for SV discovery. Panels A), B), and C) show how multiple tools across all variation types had technical issues when running on Hisat2 alignment inputs and, thus, are shown to have a zero recall value because they could not be evaluated.

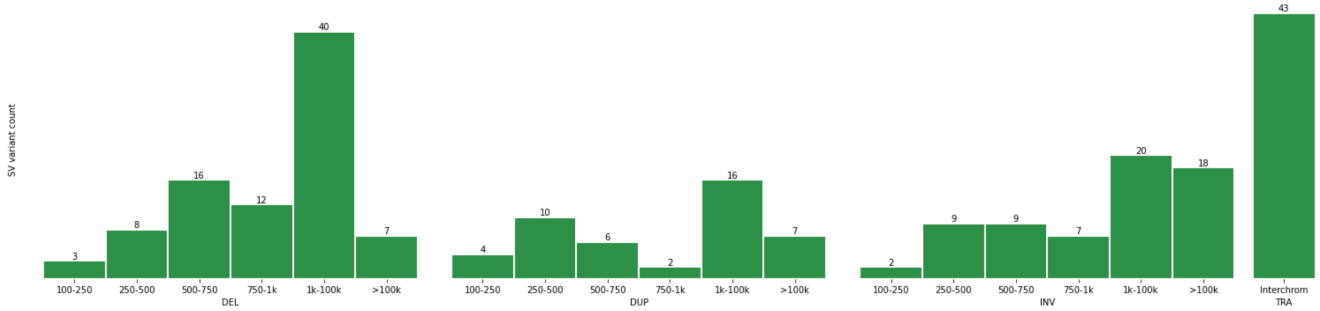


Figure S17: **Sizes of the selected variants to produce the tumorized genome, related to Figure 2.** SVs of the shown types and sizes were selected by excluding overlapping variants in this order: translocation, inversion, duplication, and deletion.

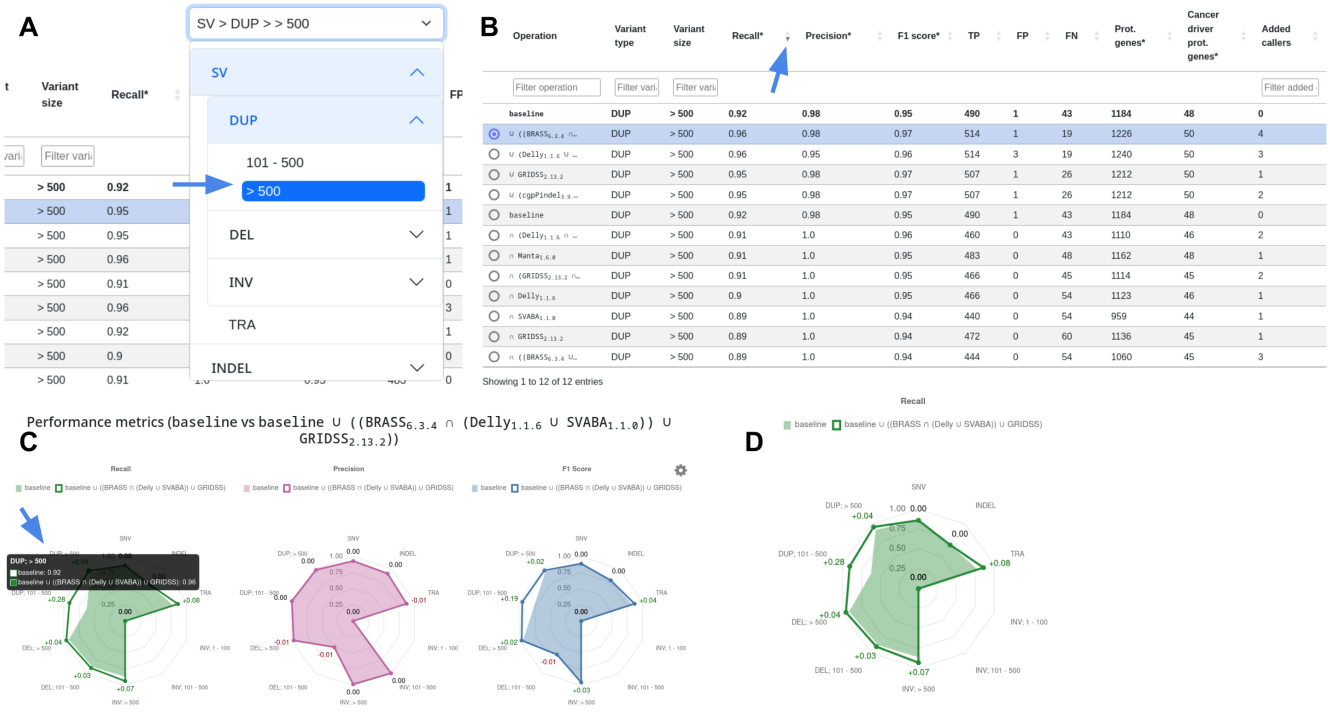


Figure S18: **User case example for improving the recall of their pipeline for large duplications and interpreting the recommendations (see Methods).** A) First, they choose to plot improvement recommendations for duplications bigger than 500bp. B) To choose based on the highest recall they sort them the recommendations in the respective column, using the symbol. C) Panel showing the figures that appear for each performance metric based on the user filtering. D) Figure downloaded as a PNG image that the user will use in their reports.

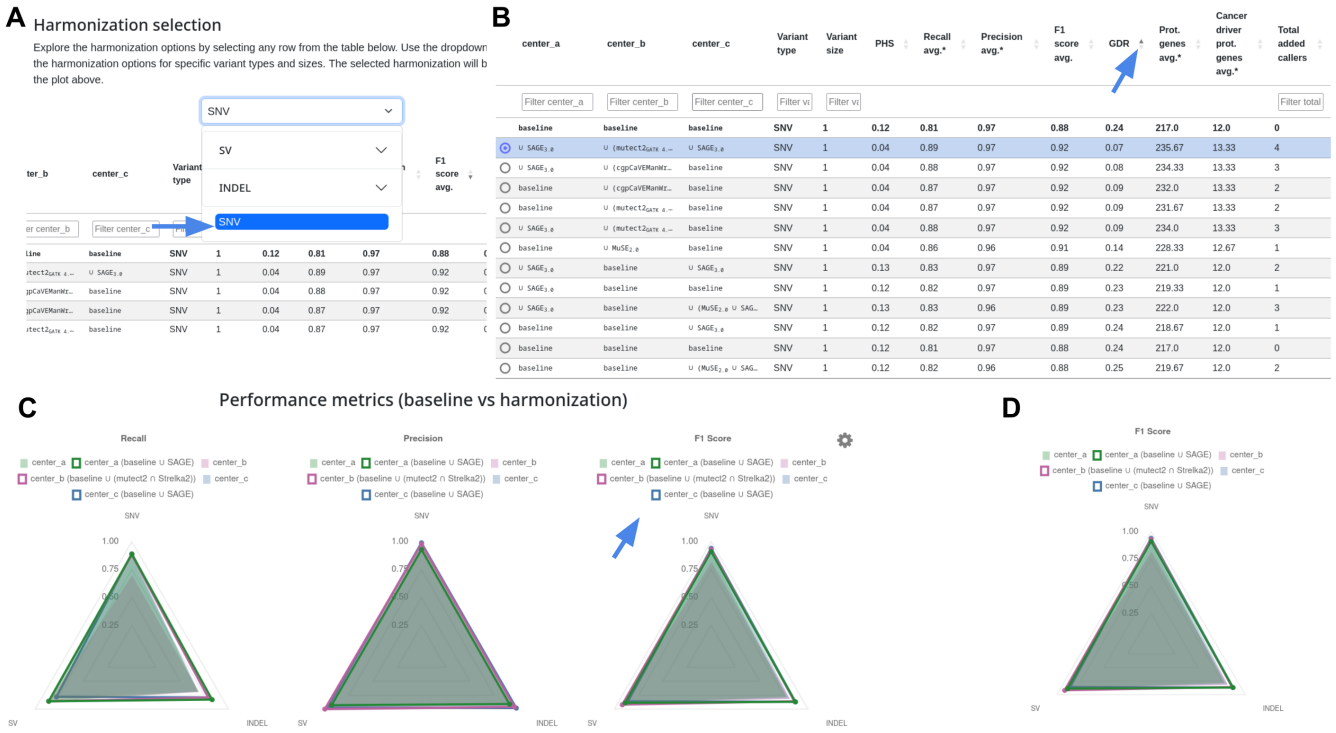


Figure S19: **User example for choosing their appropriate harmonization recommendation (see Methods).** A hypothetical user wants to harmonize three input pipelines for SNV discovery, but their use case is focused on making functionally relevant discovered variants as consistent as possible between the SNV calling pipelines. Also, he must generate a report to show graphically how this harmonization process would improve or worsen discovery performance metrics. A) First, this user would begin by filtering the table and choosing only SNVs. B) Next step, they would sort the list by clicking on the arrow on the GDR header and choose the first recommendation by clicking the button on the left. Panels C) and D) show how this element would display the figures for this recommendation, and the user would be able to save the performance figures as individual images. This chosen recommendation shows that although the main election criterion was decreasing GDR, recall and precision improved or were equal for the input pipelines. Additionally, they now know which callers they have to add and how to do it, which in this case would mean adding 4 variant callers, which is not a problem since in this use case the main need is to decrease inconsistencies in calling of gene-affecting variants.

Table S2: **Optimal strategies for improvement and harmonization (see Methods).**

Variant type	Center A	Center B	Center C
SNV	baseline	∪ [cgpCaVEManWrapper (v1.16.0) ∩ mutect2 (GATK 4.2.6.1)]	baseline
Indel	∪ SAGE (v3.0)	∪ [mutect2 (GATK 4.2.6.1) ∩ Strelka (v2.9.10)]	∪ SAGE (v3.0)
SV	∪ GRIDSS (v2.13.2)	∪ Manta (v1.6.0)	∪ [Manta (v. . .) ∪ GRIDSS (v. . .)]