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

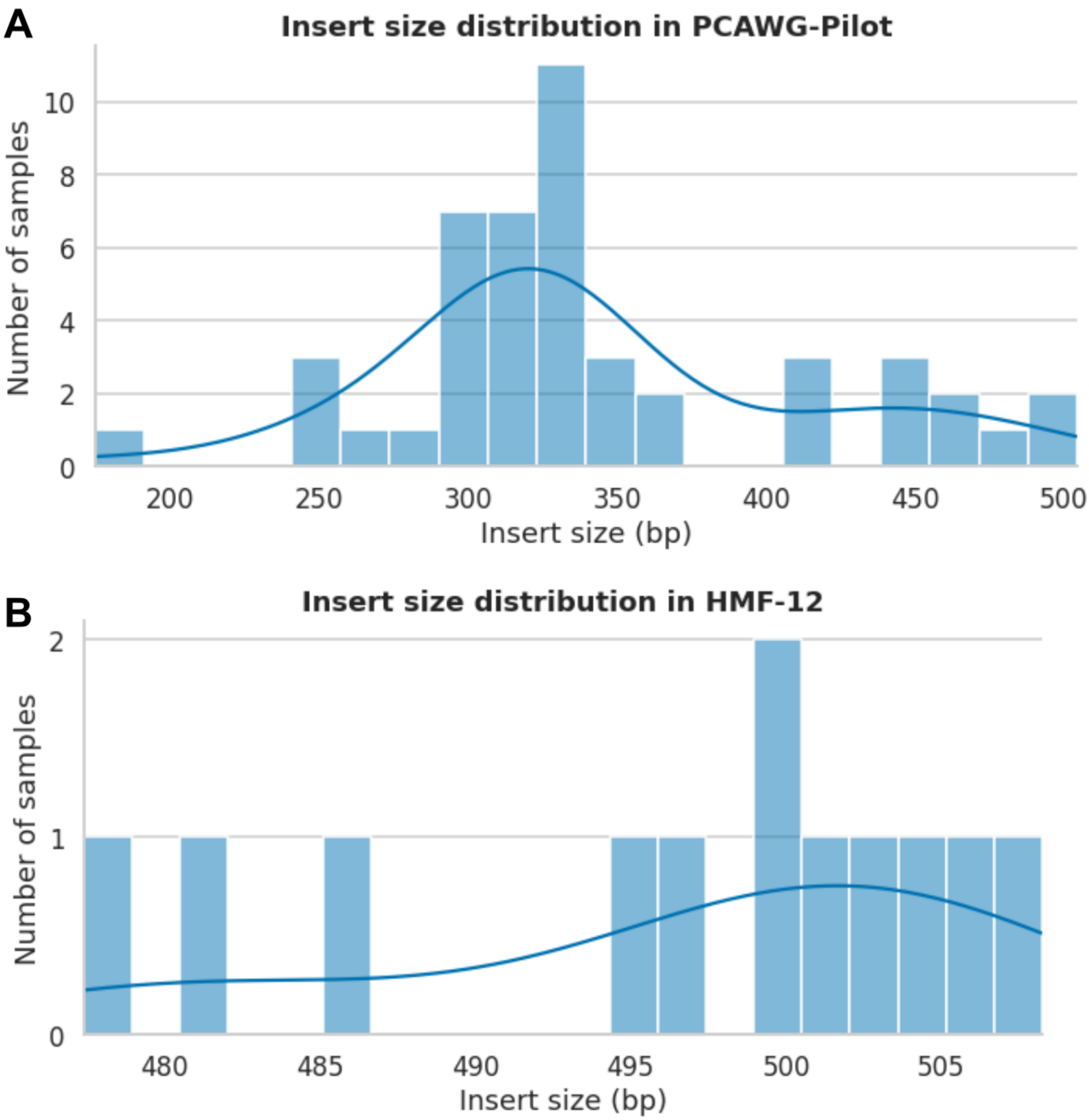


Figure S1: **Insert size distributions for all samples from PCAWG-Pilot and HMF-12, related to Figure 1.** A) PCAWG-Pilot insert sizes. Analysis of this distribution allowed us to find a bimodal sample, which was excluded from the experiments. B) HMF-12 samples show consistent insert size homogeneity.

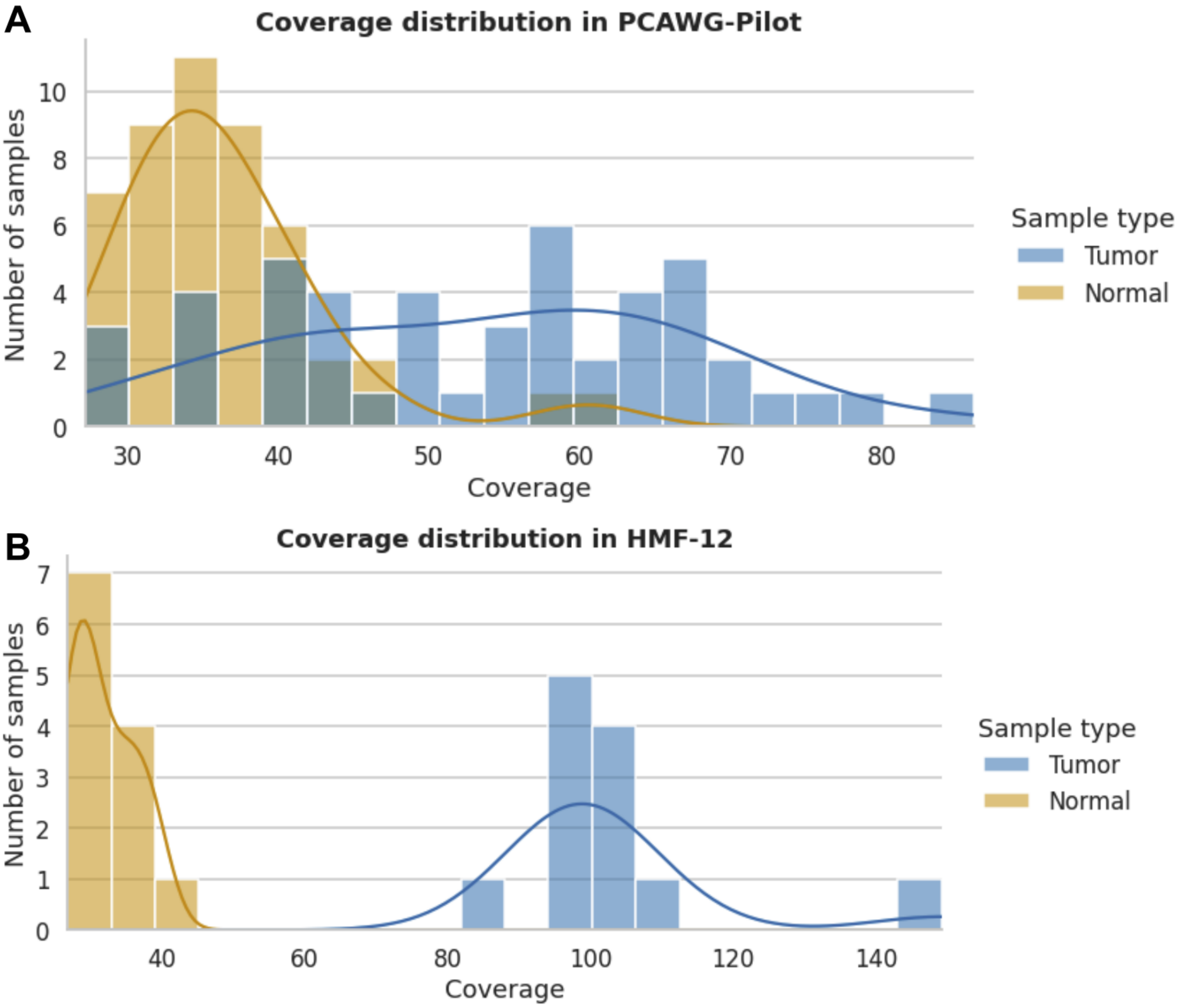


Figure S2: **Coverage distribution in the tumor-normal datasets, related to Figure 1.** A) PCAWG-Pilot samples and B) HMF-12 samples.

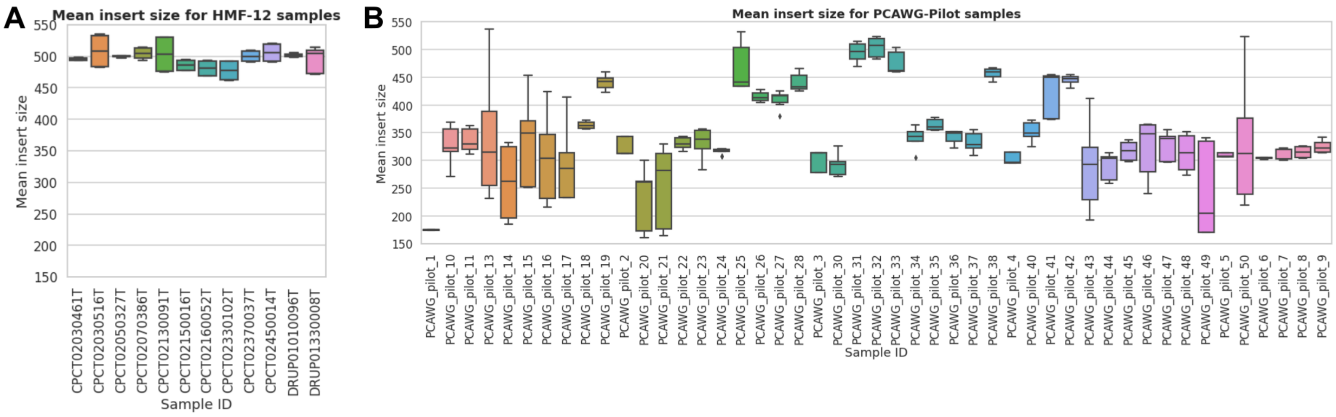


Figure S3: **Insert size boxplots samples in the HMF-12 and PCAWG-Pilot datasets (see Methods).** A) The HMF-12 samples showcase similar insert sizes where all the average values are between approximately a 50bp range. B) Conversely, there is a high level of heterogeneity between insert sizes for each 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.

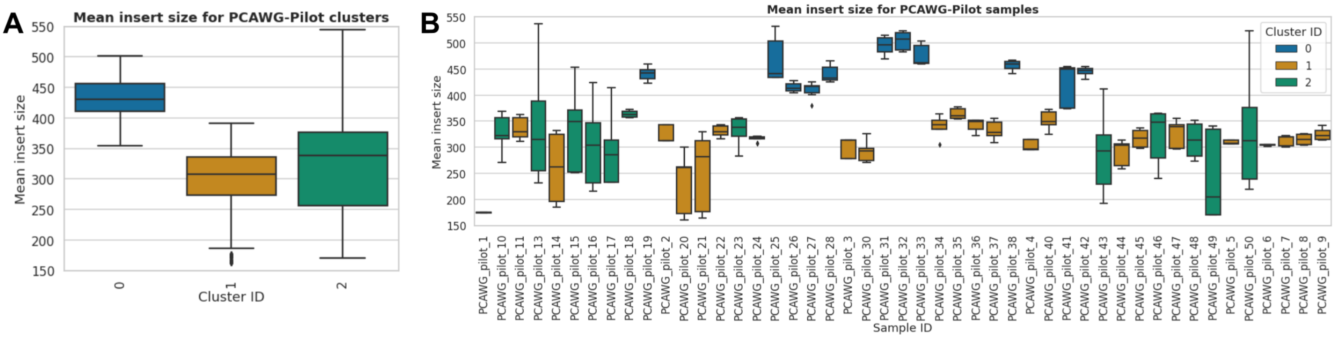


Figure S4: **Insert sizes of sample clusters from the PCAWG-Pilot datasets (see Methods).** A) The three clusters grouping the PCAWG-Pilot samples show low variance for insert sizes, where 0 and 1 range mostly between 50 bp - 100 bp, and 2 depicts a higher variance concordant with the original samples 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.

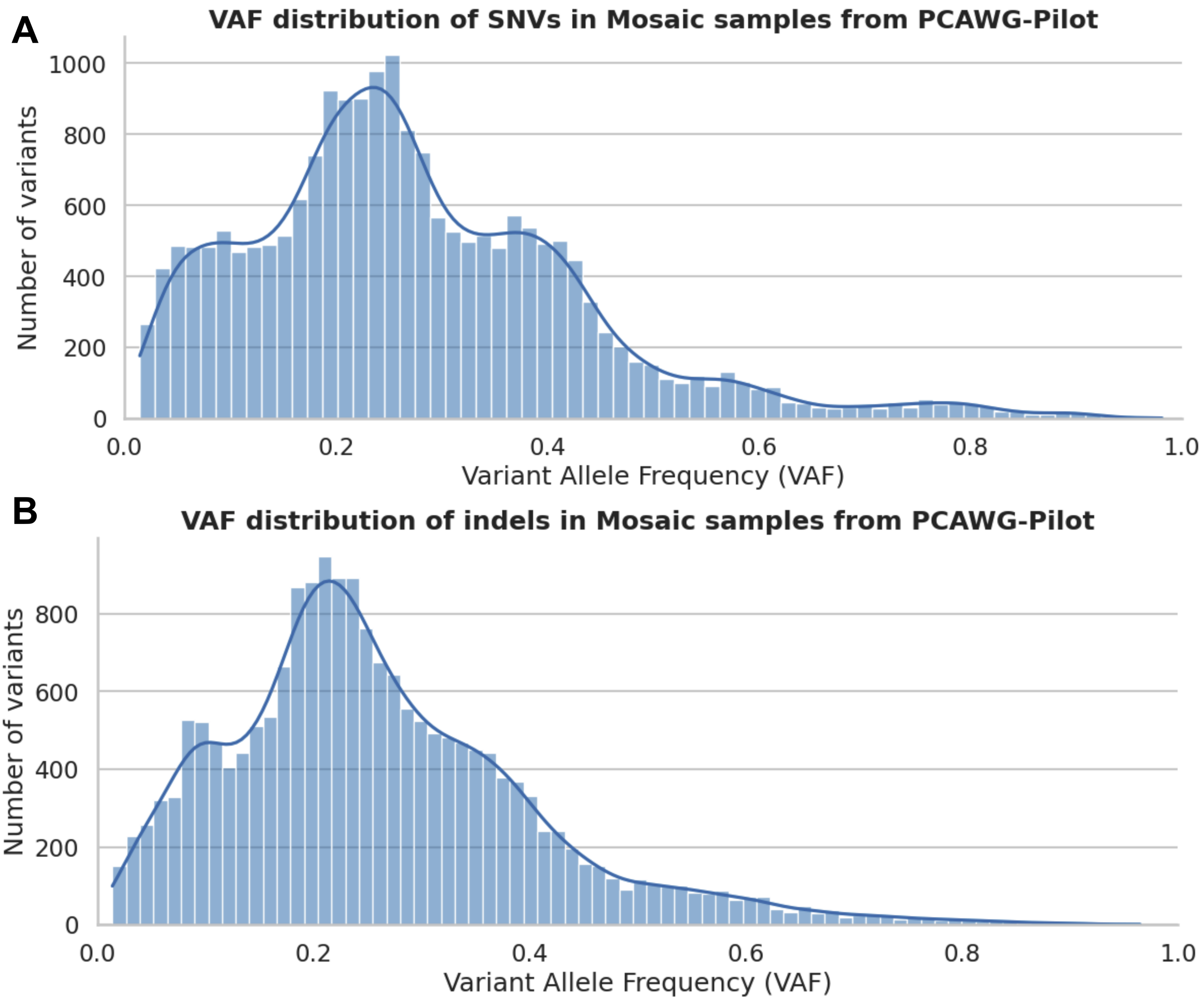


Figure S5: **Variant Allele Frequency (VAF) distributions for the PCAWG-Pilot gold standard variants included in this study, related to Figure 1.** A) Variant Allele Frequency (VAF) distributions for SNVs and B) Indels from the PCAWG-Pilot variant collection included in the mosaic genomes. VAF values 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.

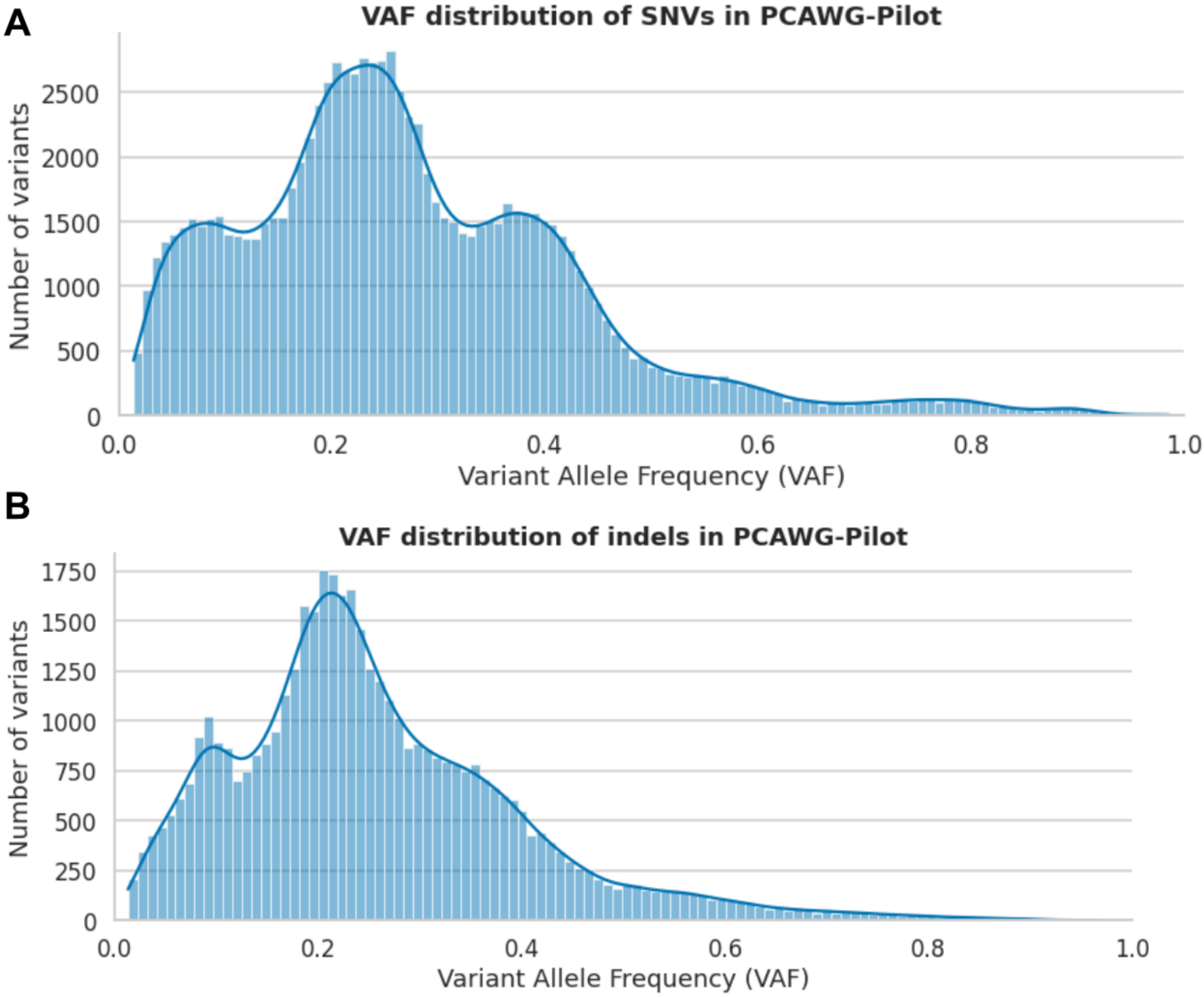


Figure S6: **Variant Allele Frequency Distributions for all variants in the PCAWG-Pilot datasets (including samples not used in the study), related to Figure 1.** A) shows the VAFs for SNVs following a multimodal distribution due to the original selection bias of the validation process, with the most 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.

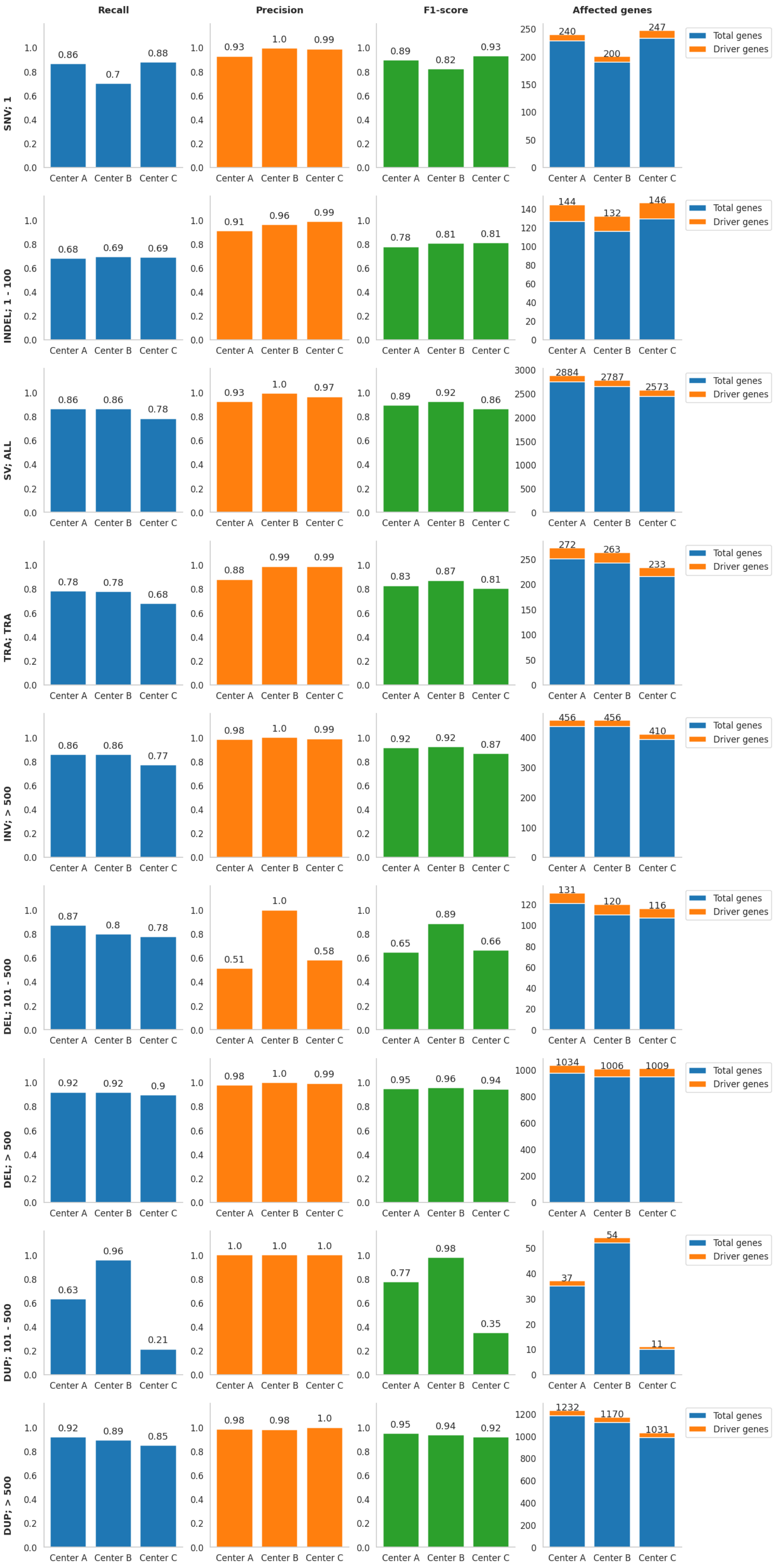
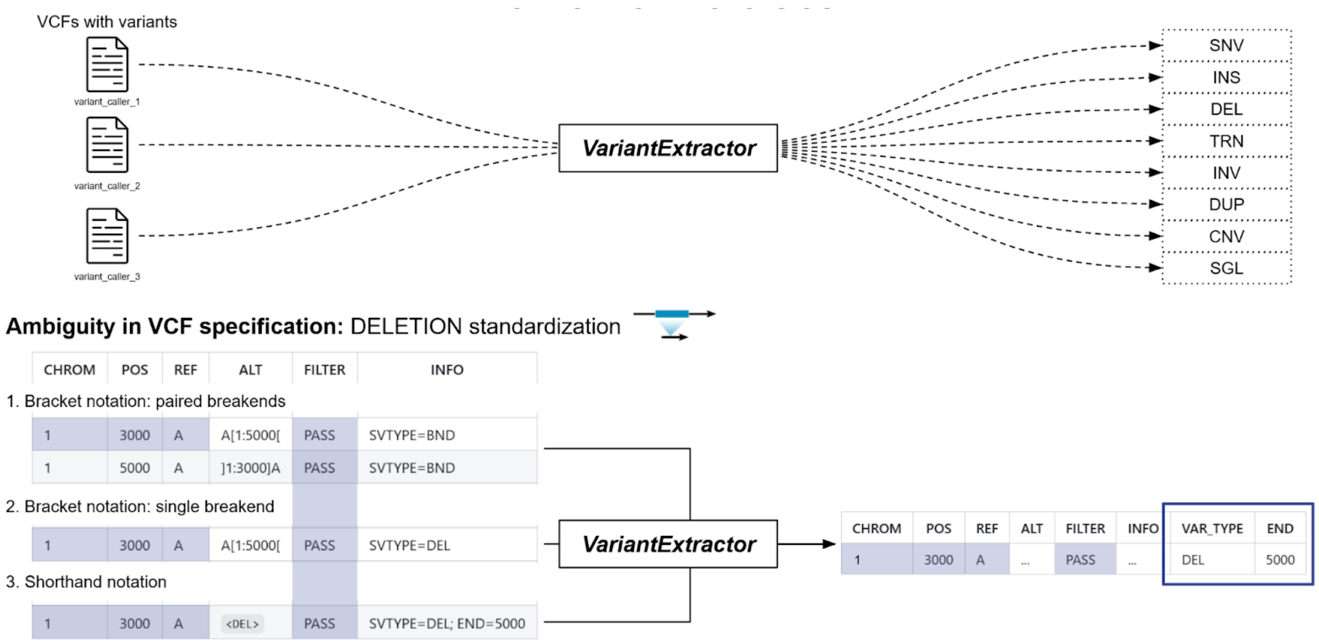

Recall
Precision
F1-score
Affected genes
SNV; 1
0.86 0.7 0.88
0.93 1.0 0.99
0.89 0.82 0.93
240 200 247
INDEL; 1 - 100
0.68 0.69 0.69
0.91 0.96 0.99
0.78 0.81 0.81
144 132 146
SV; ALL
0.86 0.86 0.78
0.93 1.0 0.97
0.89 0.92 0.86
2884 2787 2573
TRA; TRA
0.78 0.78 0.68
0.88 0.99 0.99
0.83 0.87 0.81
272 263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

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.**

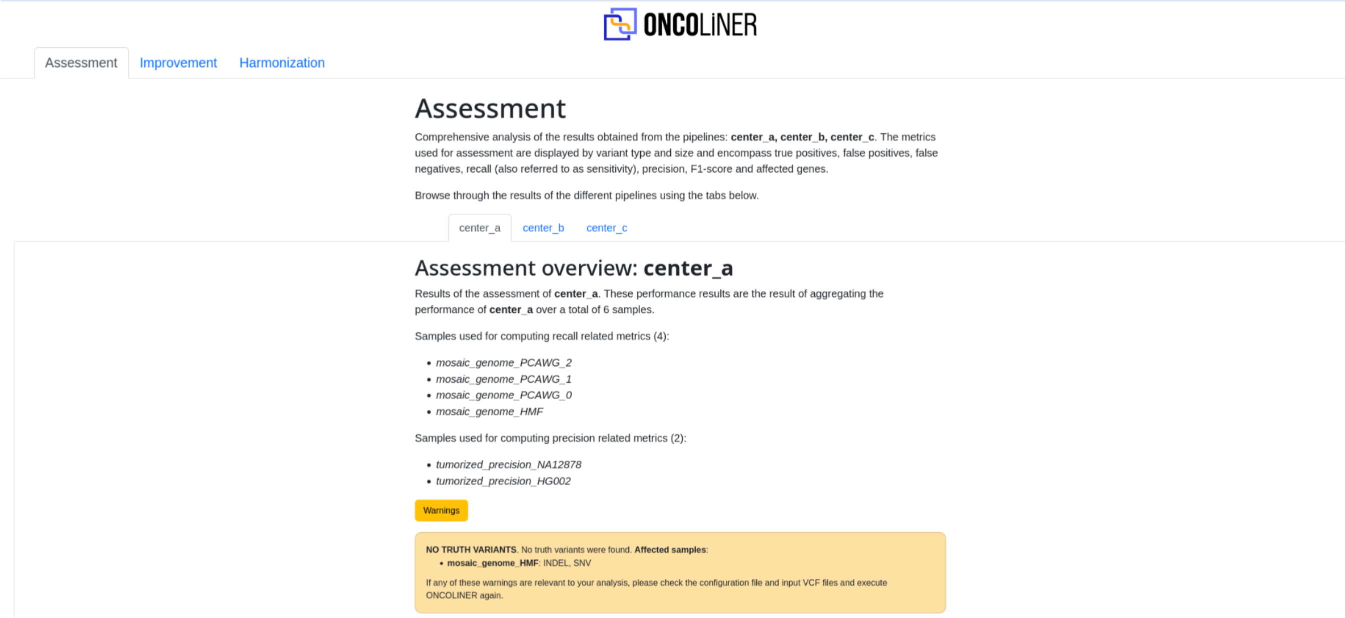
Variant type	Optimal combination	F1-score
SNV	[mutect2 (from GATK 4.2.6.1) ∩ Strelka (v22.9.10)] ∪ SAGE (v3.0)	93%
Indel	[mutect2 (from GATK 4.2.6.1) ∩ Strelka (v22.9.10)] ∪ SAGE (v3.0)	87%
SV	[Delly (v1.1.6) ∩ Manta (v1.6.0)] ∪ GRIDSS (v2.13.2)	93%



CHROM	POS	REF	ALT	FILTER	INFO
1. Bracket notation: paired breakends					
1	3000	A	A[1:5000[	PASS	SVTYPE=BND
1	5000	A	]1:3000]A	PASS	SVTYPE=BND
2. Bracket notation: single breakend					
1	3000	A	A[1:5000[	PASS	SVTYPE=DEL
3. Shorthand notation					
1	3000	A	<DEL>	PASS	SVTYPE=DEL; END=5000

CHROM	POS	REF	ALT	FILTER	INFO	VAR_TYPE	END
1	3000	A	...	PASS	...	DEL	5000

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.



Assessment

Comprehensive analysis of the results obtained from the pipelines: **center_a, center_b, center_c**. The metrics used for assessment are displayed by variant type and size and encompass true positives, false positives, false negatives, recall (also referred to as sensitivity), precision, F1-score and affected genes.

Browse through the results of the different pipelines using the tabs below.

Assessment overview: **center_a**

Results of the assessment of **center_a**. These performance results are the result of aggregating the performance of **center_a** over a total of 6 samples.

Samples used for computing recall related metrics (4):

- *mosaic_genome_PCAWG_2*
- *mosaic_genome_PCAWG_1*
- *mosaic_genome_PCAWG_0*
- *mosaic_genome_HMF*

Samples used for computing precision related metrics (2):

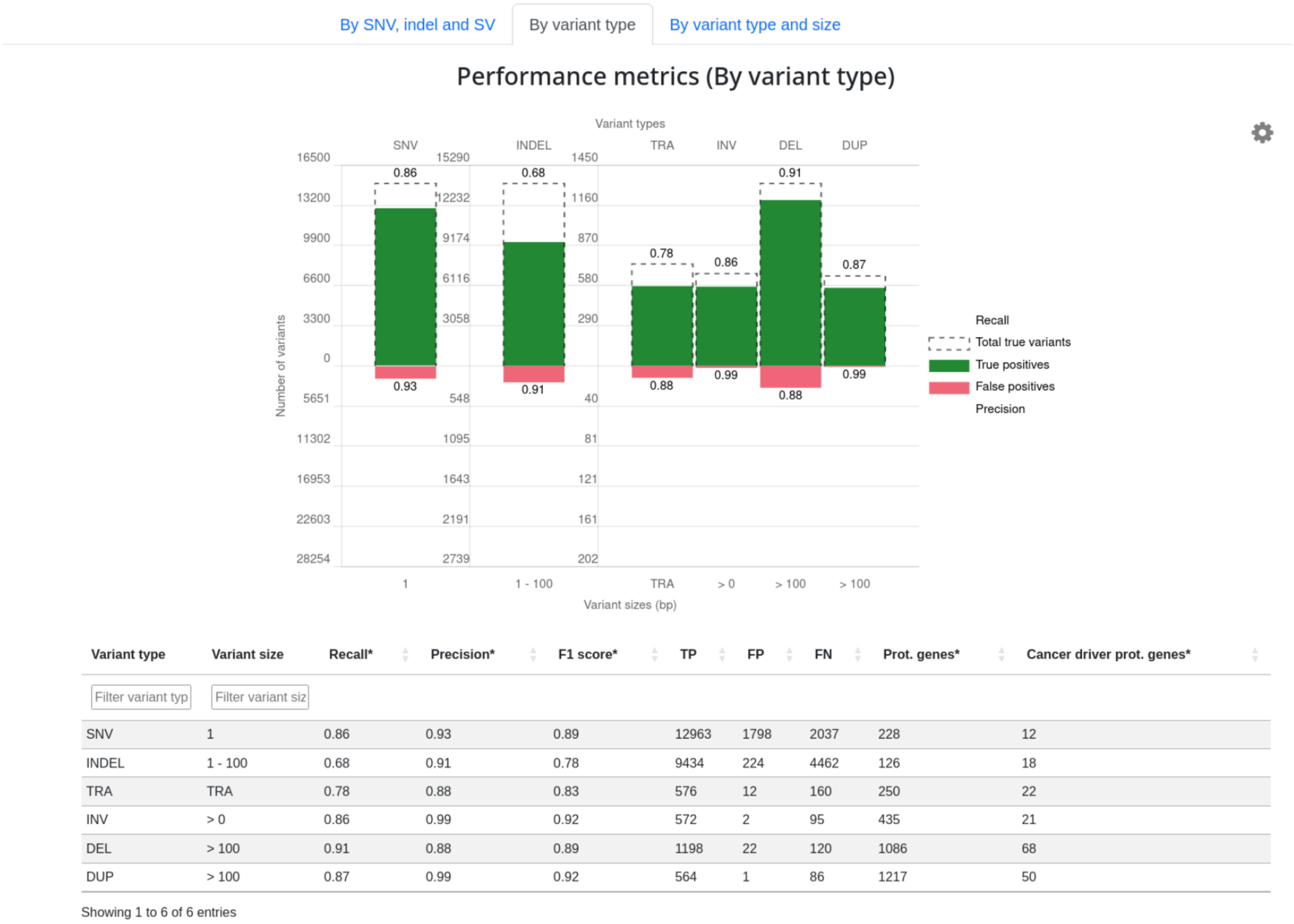
- *tumorized_precision_NA12878*
- *tumorized_precision_HG002*

NO TRUTH VARIANTS. No truth variants were found. **Affected samples**:

- **mosaic_genome_HMF**: INDEL, SNV

If any of these warnings are relevant to your analysis, please check the configuration file and input VCF files and execute ONCOLINER again.

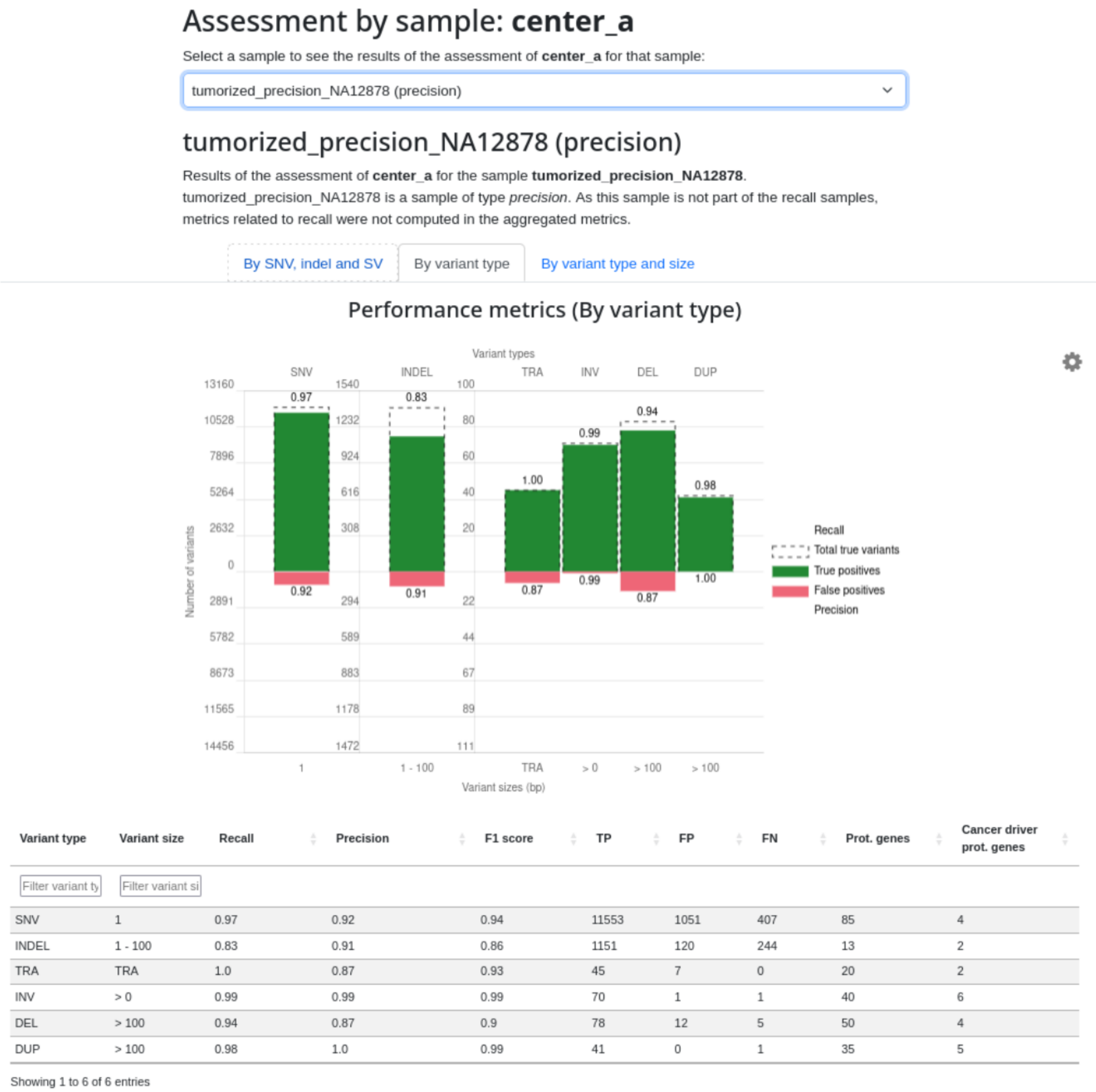
Figure S9: **Screenshot of the assessment module description section (see Methods).** Interactive results are provided in HTML format. This report is composed of different sections. First, the visualization of the results is described by enumerating the input pipelines and the benchmarking genomes. As 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.



Assessment by sample: **center_a**

Select a sample to see the results of the assessment of **center_a** for that sample:

tumorized_precision_NA12878 (precision)

tumorized_precision_NA12878 (precision)

Results of the assessment of **center_a** for the sample **tumorized_precision_NA12878**. tumorized_precision_NA12878 is a sample of type *precision*. As this sample is not part of the recall samples, metrics related to recall were not computed in the aggregated metrics.

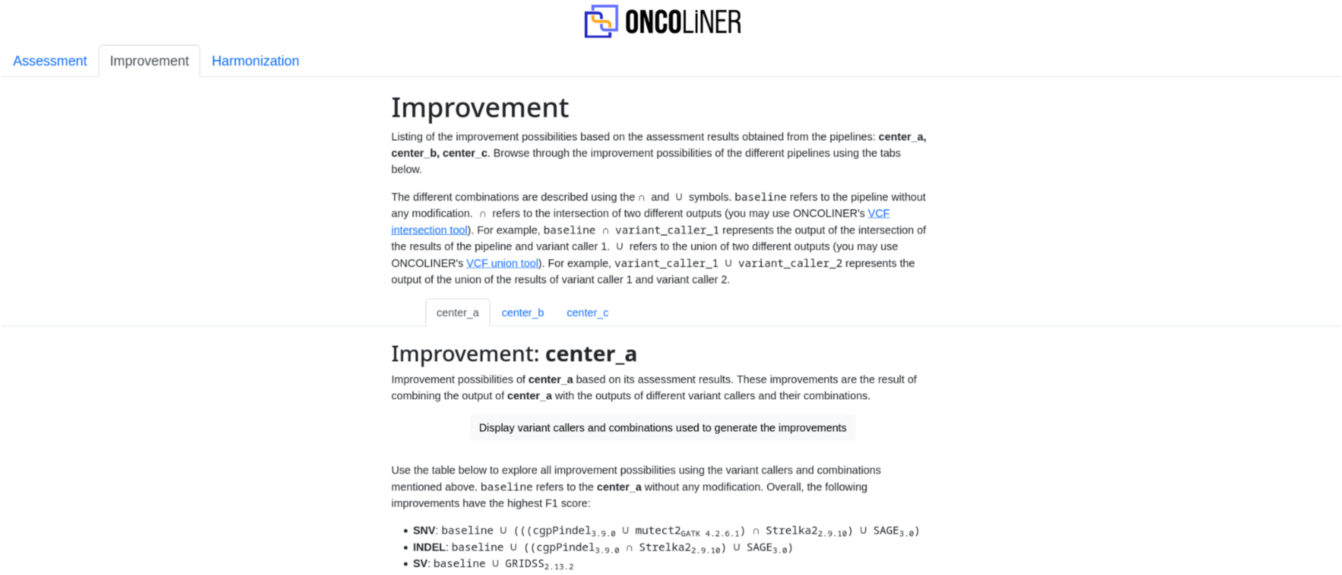
By SNV, indel and SV | By variant type | By variant type and size

Performance metrics (By variant type)

Variant type	Variant size	Recall	Precision	F1 score	TP	FP	FN	Prot. genes	Cancer driver prot. genes
Filter variant ty	Filter variant si								
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

Showing 1 to 6 of 6 entries

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.


ONCOLiNER
Assessment
Improvement
Harmonization
center_a
center_b
center_c
Display variant callers and combinations used to generate the improvements

Improvement

Listing of the improvement possibilities based on the assessment results obtained from the pipelines: **center_a, center_b, center_c**. Browse through the improvement possibilities of the different pipelines using the tabs below.

The different combinations are described using the ∩ and ∪ symbols. `baseline` refers to the pipeline without any modification. ∩ refers to the intersection of two different outputs (you may use ONCOLINER's VCF intersection tool). For example, `baseline ∩ variant_caller_1` represents the output of the intersection of the results of the pipeline and variant caller 1. ∪ refers to the union of two different outputs (you may use ONCOLINER's VCF union tool). For example, `variant_caller_1 ∪ variant_caller_2` represents the output of the union of the results of variant caller 1 and variant caller 2.

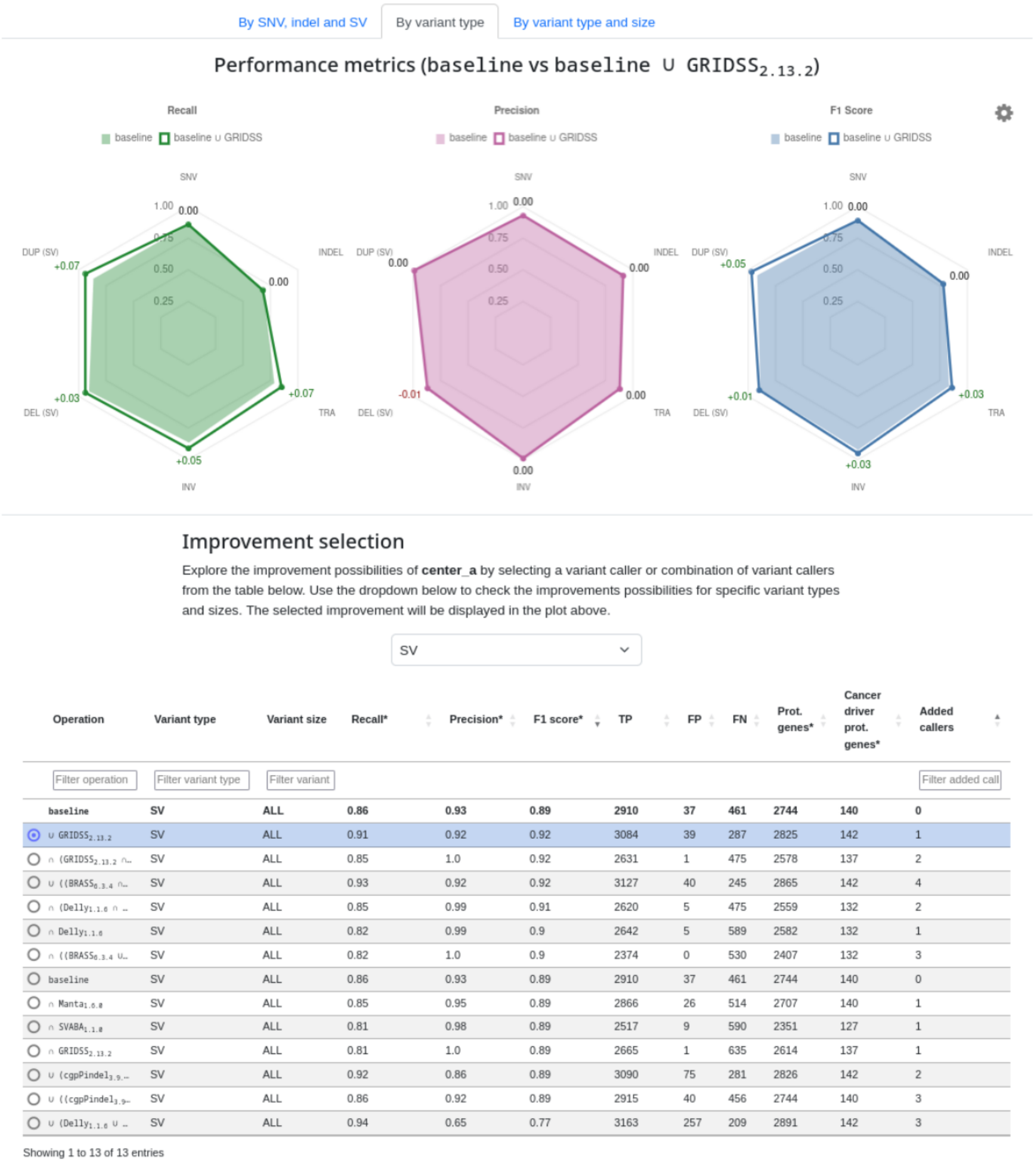
Improvement: **center_a**

Improvement possibilities of **center_a** based on its assessment results. These improvements are the result of combining the output of **center_a** with the outputs of different variant callers and their combinations.

Use the table below to explore all improvement possibilities using the variant callers and combinations mentioned above. `baseline` refers to the **center_a** without any modification. Overall, the following improvements have the highest F1 score:

- **SNV**: $\texttt{baseline} \cup (((\texttt{cgpPindel}_{3.9.0} \cup \texttt{mutect2}_{\texttt{GATK 4.2.6.1}}) \cap \texttt{Strelka2}_{2.9.10}) \cup \texttt{SAGE}_{3.0})$
- **INDEL**: $\texttt{baseline} \cup ((\texttt{cgpPindel}_{3.9.0} \cap \texttt{Strelka2}_{2.9.10}) \cup \texttt{SAGE}_{3.0})$
- **SV**: $\texttt{baseline} \cup \texttt{GRIDSS}_{2.13.2}$

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.



Performance metrics (baseline vs baseline $\cup$ GRIDSS$_{2.13.2}$)

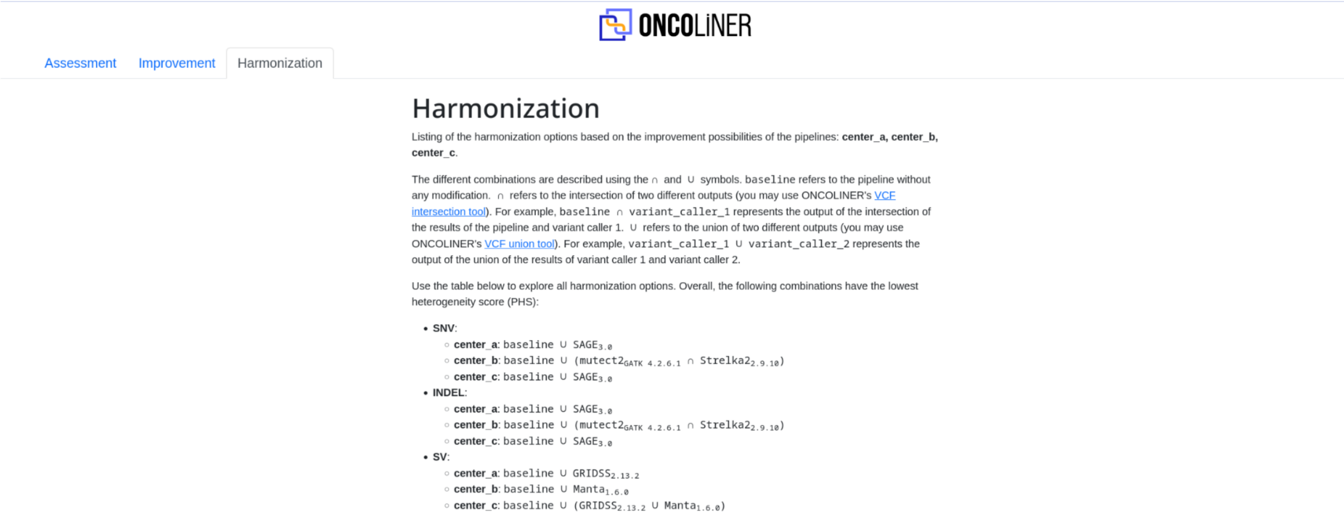
Improvement selection

Explore the improvement possibilities of **center_a** by selecting a variant caller or combination of variant callers from the table below. Use the dropdown below to check the improvements possibilities for specific variant types and sizes. The selected improvement will be displayed in the plot above.

	Operation	Variant type	Variant size	Recall*	Precision*	F1 score*	TP	FP	FN	Prot. genes*	Cancer driver prot. genes*	Added callers
	Filter operation	Filter variant type	Filter variant									Filter added call
	baseline	**SV**	**ALL**	**0.86**	**0.93**	**0.89**	**2910**	**37**	**461**	**2744**	**140**	**0**
●	$\cup$ GRIDSS$_{2.13.2}$	SV	ALL	0.91	0.92	0.92	3084	39	287	2825	142	1
○	$\cap$ (GRIDSS$_{2.13.2}$ $\cap$…	SV	ALL	0.85	1.0	0.92	2631	1	475	2578	137	2
○	$\cup$ ((BRASS$_{6.3.4}$ $\cap$…	SV	ALL	0.93	0.92	0.92	3127	40	245	2865	142	4
○	$\cap$ (Delly$_{1.1.6}$ $\cap$ …	SV	ALL	0.85	0.99	0.91	2620	5	475	2559	132	2
○	$\cap$ Delly$_{1.1.6}$	SV	ALL	0.82	0.99	0.9	2642	5	589	2582	132	1
○	$\cap$ ((BRASS$_{6.3.4}$ $\cup$…	SV	ALL	0.82	1.0	0.9	2374	0	530	2407	132	3
○	baseline	SV	ALL	0.86	0.93	0.89	2910	37	461	2744	140	0
○	$\cap$ Manta$_{1.6.0}$	SV	ALL	0.85	0.95	0.89	2866	26	514	2707	140	1
○	$\cap$ SVABA$_{1.1.0}$	SV	ALL	0.81	0.98	0.89	2517	9	590	2351	127	1
○	$\cap$ GRIDSS$_{2.13.2}$	SV	ALL	0.81	1.0	0.89	2665	1	635	2614	137	1
○	$\cup$ (cgpPindel$_{3.9.}$…	SV	ALL	0.92	0.86	0.89	3090	75	281	2826	142	2
○	$\cup$ ((cgpPindel$_{3.9}$…	SV	ALL	0.86	0.92	0.89	2915	40	456	2744	140	3
○	$\cup$ (Delly$_{1.1.6}$ $\cup$ …	SV	ALL	0.94	0.65	0.77	3163	257	209	2891	142	3

Showing 1 to 13 of 13 entries

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.



Harmonization

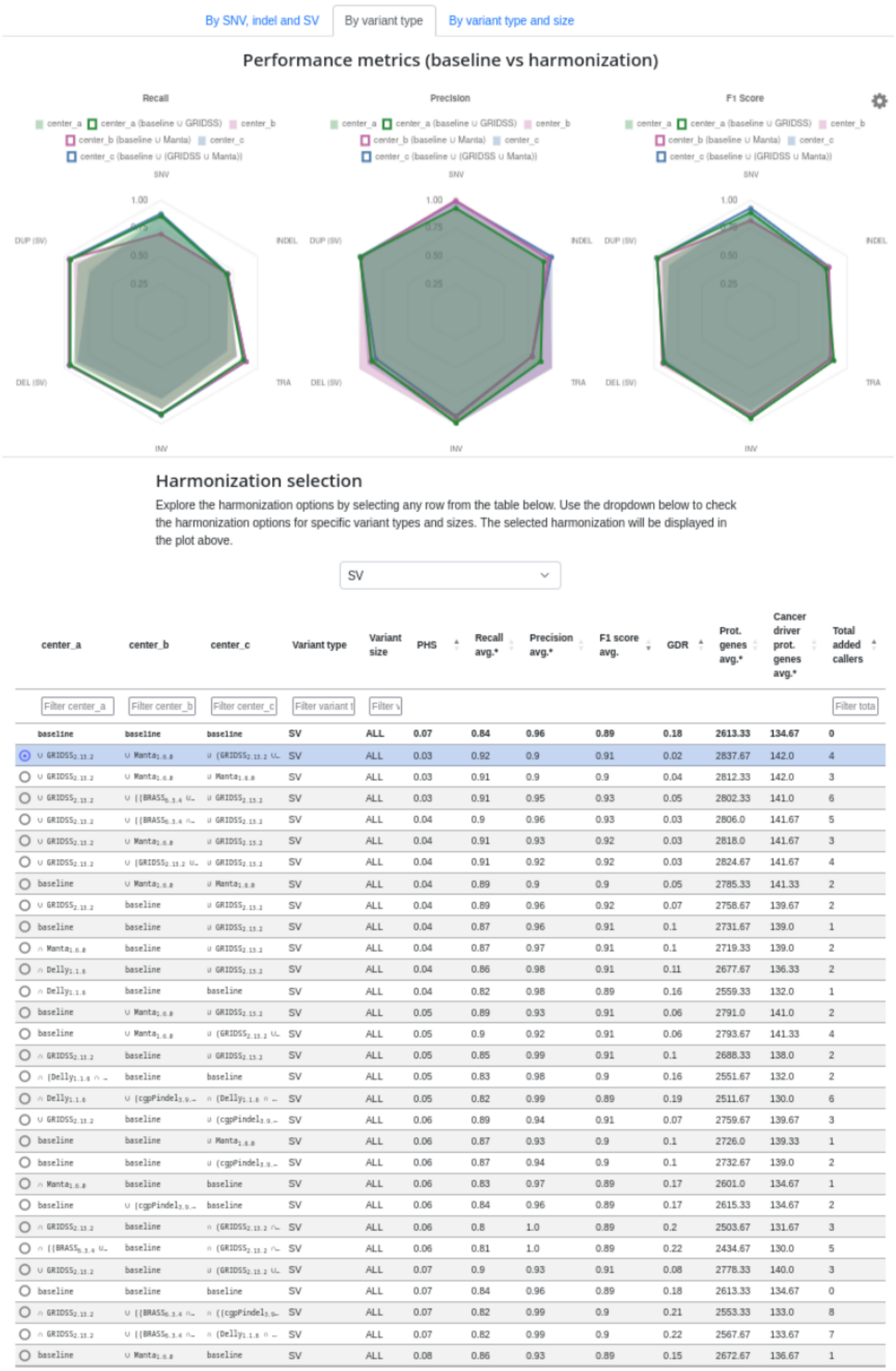
Listing of the harmonization options based on the improvement possibilities of the pipelines: **center_a, center_b, center_c**.

The different combinations are described using the ∩ and ∪ symbols. `baseline` refers to the pipeline without any modification. ∩ refers to the intersection of two different outputs (you may use ONCOLINER's VCF intersection tool). For example, `baseline ∩ variant_caller_1` represents the output of the intersection of the results of the pipeline and variant caller 1. ∪ refers to the union of two different outputs (you may use ONCOLINER's VCF union tool). For example, `variant_caller_1 ∪ variant_caller_2` represents the output of the union of the results of variant caller 1 and variant caller 2.

Use the table below to explore all harmonization options. Overall, the following combinations have the lowest heterogeneity score (PHS):

- **SNV**:
  - **center_a**: baseline $\cup$ $\text{SAGE}_{3.0}$
  - **center_b**: baseline $\cup$ $(\text{mutect2}_{\text{GATK 4.2.6.1}} \cap \text{Strelka2}_{2.9.10})$
  - **center_c**: baseline $\cup$ $\text{SAGE}_{3.0}$
- **INDEL**:
  - **center_a**: baseline $\cup$ $\text{SAGE}_{3.0}$
  - **center_b**: baseline $\cup$ $(\text{mutect2}_{\text{GATK 4.2.6.1}} \cap \text{Strelka2}_{2.9.10})$
  - **center_c**: baseline $\cup$ $\text{SAGE}_{3.0}$
- **SV**:
  - **center_a**: baseline $\cup$ $\text{GRIDSS}_{2.13.2}$
  - **center_b**: baseline $\cup$ $\text{Manta}_{1.6.0}$
  - **center_c**: baseline $\cup$ $(\text{GRIDSS}_{2.13.2} \cup \text{Manta}_{1.6.0})$

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.



Performance metrics (baseline vs harmonization)

Recall
Precision
F1 Score
center_a
center_a (baseline ∪ GRIDSS)
center_b
center_b (baseline ∪ Manta)
center_c
center_c (baseline ∪ (GRIDSS ∪ Manta))
SNV
INDEL
TRA
INV
DEL (SV)
DUP (SV)
1.00
0.75
0.50
0.25

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

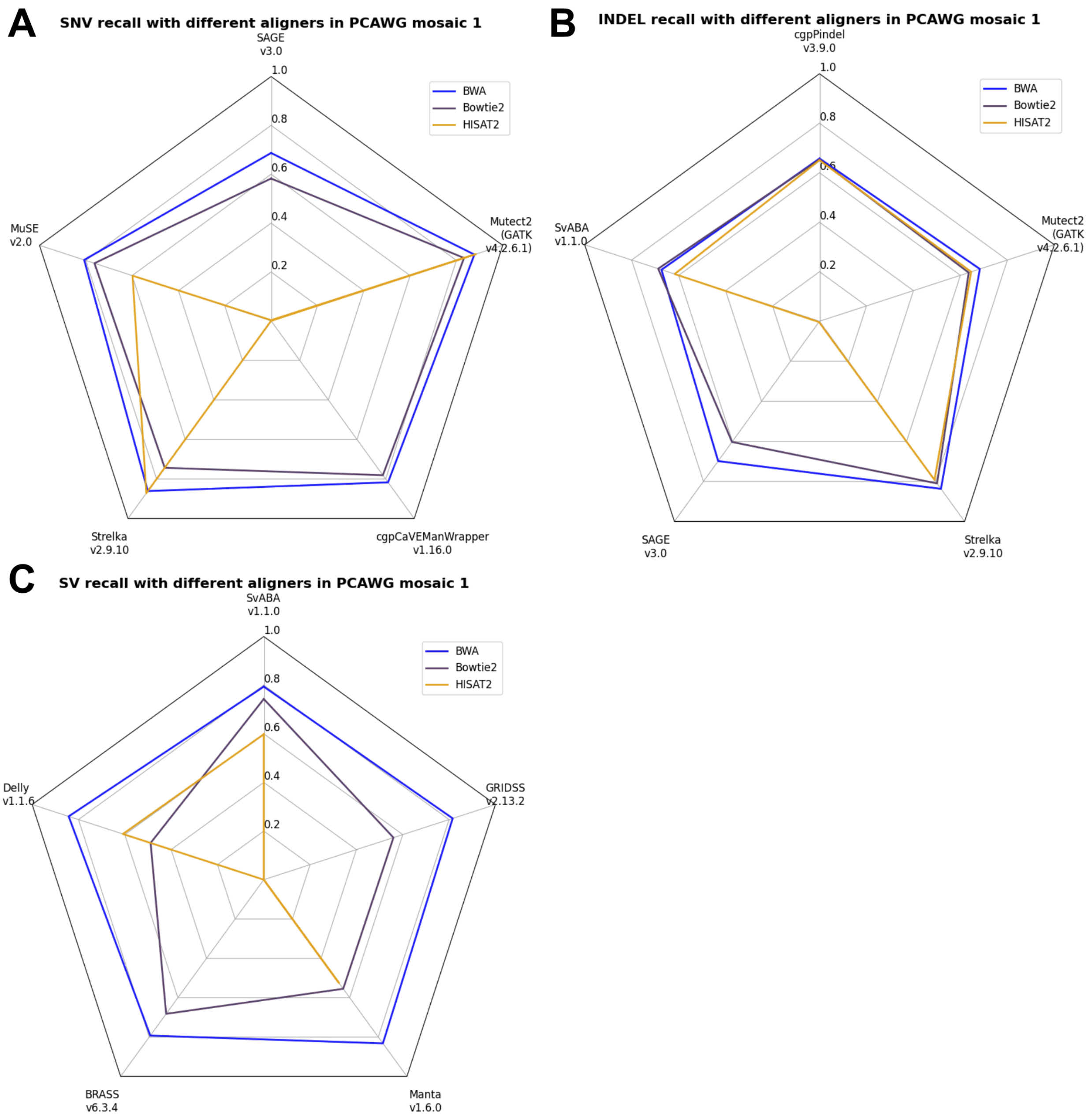


Figure S16: **Performance of all callers over a mosaic genome mapped with different read aligners, per variant type (see Methods).** Color codes portray the recall values for callers on the samples by read-mapper software. A) Shows SNVs, B) indels, and C) SVs. Panel A) and B) show that for SNVs and indels, 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.

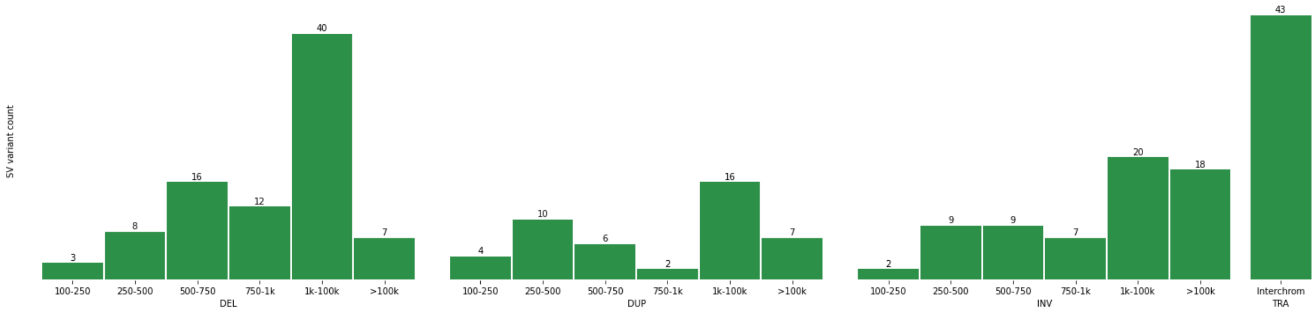
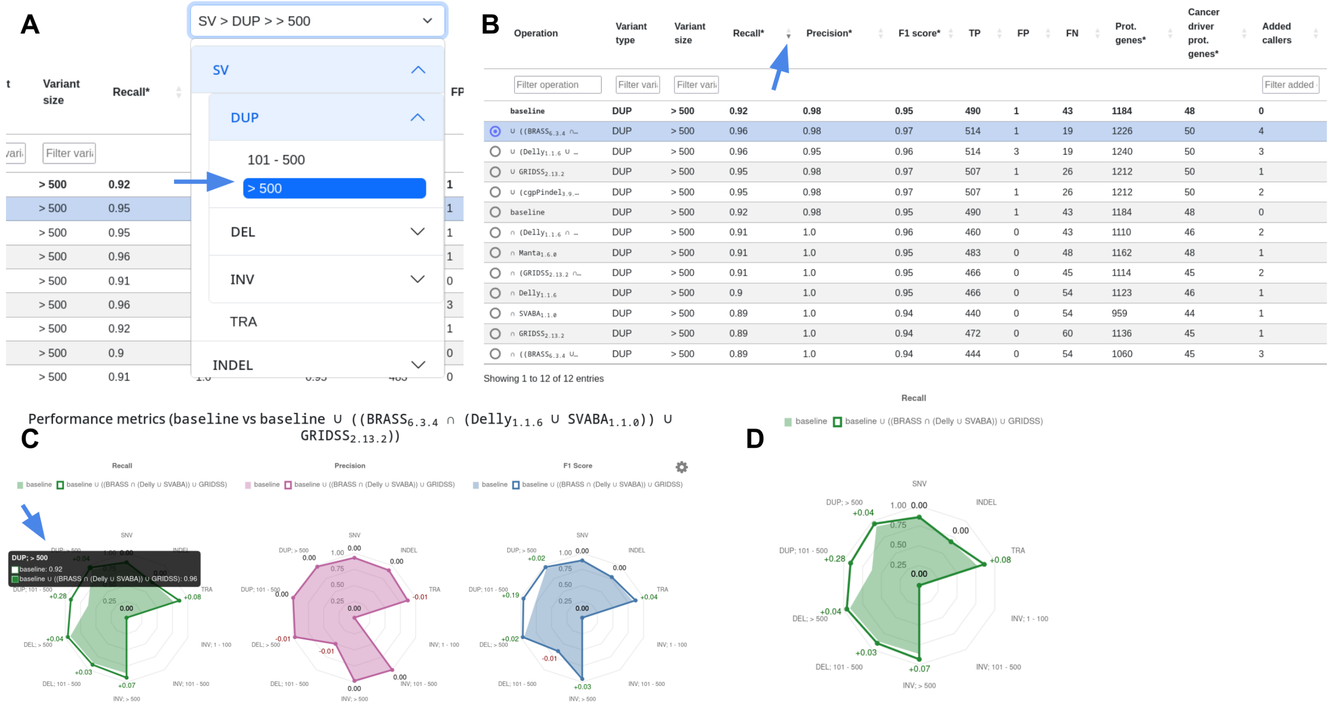


Figure S17: **Sizes of the selected variants to produce the tumorized genome, related to Figure 2.** SVs of the shown types and sizes were selected by excluding overlapping variants in this order: translocation, inversion, duplication, and deletion.



Variant size	Recall*	FP
> 500	0.92	1
> 500	0.95	1
> 500	0.95	1
> 500	0.96	1
> 500	0.91	0
> 500	0.96	3
> 500	0.92	1
> 500	0.9	0
> 500	0.91	0

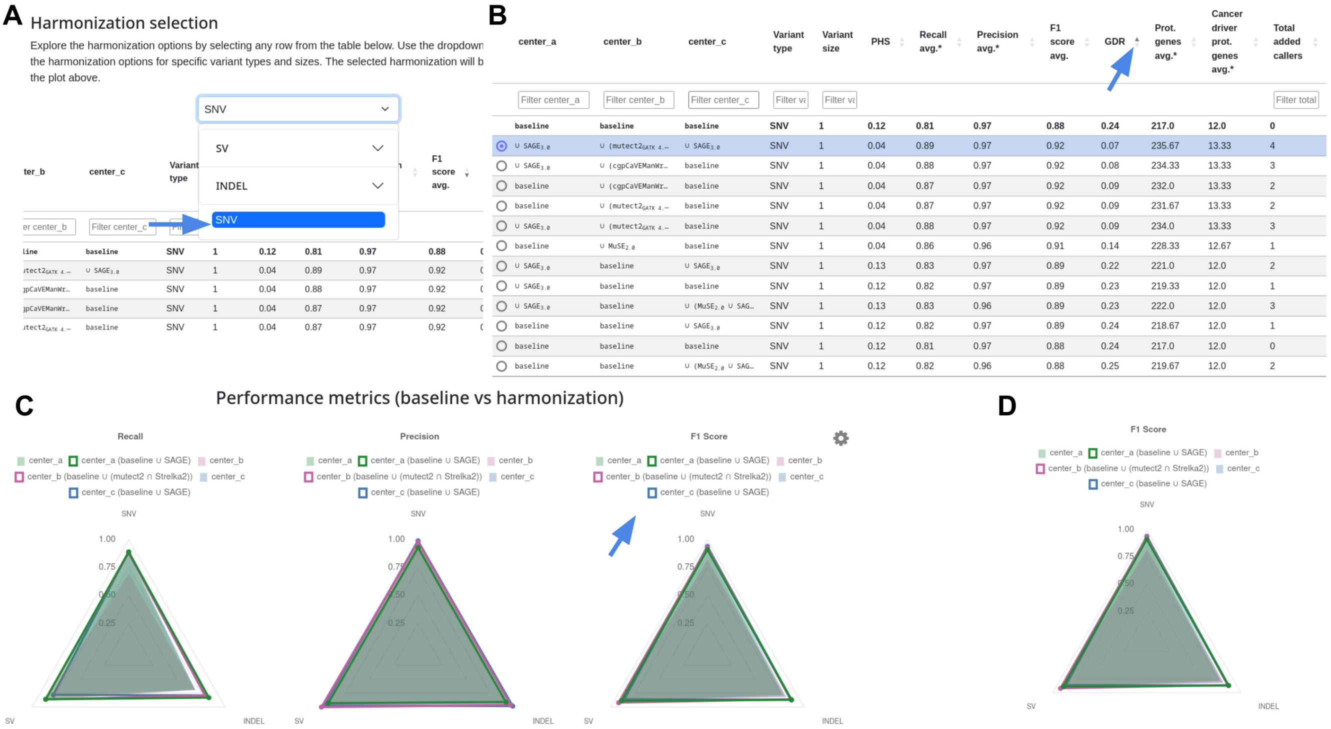
	Operation	Variant type	Variant size	Recall*	Precision*	F1 score*	TP	FP	FN	Prot. genes*	Cancer driver prot. genes*	Added callers
	Filter operation	Filter vari	Filter vari									Filter added
	baseline	**DUP**	**> 500**	**0.92**	**0.98**	**0.95**	**490**	**1**	**43**	**1184**	**48**	**0**
●	∪ ((BRASS$_{6.3.4}$ ∩…	DUP	> 500	0.96	0.98	0.97	514	1	19	1226	50	4
○	∪ (Delly$_{1.1.6}$ ∪ …	DUP	> 500	0.96	0.95	0.96	514	3	19	1240	50	3
○	∪ GRIDSS$_{2.13.2}$	DUP	> 500	0.95	0.98	0.97	507	1	26	1212	50	1
○	∪ (cgpPindel$_{3.9.…}$	DUP	> 500	0.95	0.98	0.97	507	1	26	1212	50	2
○	baseline	DUP	> 500	0.92	0.98	0.95	490	1	43	1184	48	0
○	∩ (Delly$_{1.1.6}$ ∩ …	DUP	> 500	0.91	1.0	0.96	460	0	43	1110	46	2
○	∩ Manta$_{1.6.0}$	DUP	> 500	0.91	1.0	0.95	483	0	48	1162	48	1
○	∩ (GRIDSS$_{2.13.2}$ ∩…	DUP	> 500	0.91	1.0	0.95	466	0	45	1114	45	2
○	∩ Delly$_{1.1.6}$	DUP	> 500	0.9	1.0	0.95	466	0	54	1123	46	1
○	∩ SVABA$_{1.1.0}$	DUP	> 500	0.89	1.0	0.94	440	0	54	959	44	1
○	∩ GRIDSS$_{2.13.2}$	DUP	> 500	0.89	1.0	0.94	472	0	60	1136	45	1
○	∩ ((BRASS$_{6.3.4}$ ∪…	DUP	> 500	0.89	1.0	0.94	444	0	54	1060	45	3

Showing 1 to 12 of 12 entries

Performance metrics (baseline vs baseline ∪ ((BRASS$_{6.3.4}$ ∩ (Delly$_{1.1.6}$ ∪ SVABA$_{1.1.0}$)) ∪ GRIDSS$_{2.13.2}$))

DUP; > 500
baseline: 0.92
baseline ∪ ((BRASS ∩ (Delly ∪ SVABA)) ∪ GRIDSS): 0.96

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.



Harmonization selection

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

ter_b	center_c	Variant type					F1 score avg.
er center_b	Filter center_c	Fi					
line	baseline	SNV	1	0.12	0.81	0.97	0.88
ıtect2$_{\text{GATK 4.…}}$	∪ SAGE$_{3.0}$	SNV	1	0.04	0.89	0.97	0.92
ȷpCaVEManWr…	baseline	SNV	1	0.04	0.88	0.97	0.92
ȷpCaVEManWr…	baseline	SNV	1	0.04	0.87	0.97	0.92
ıtect2$_{\text{GATK 4.…}}$	baseline	SNV	1	0.04	0.87	0.97	0.92

	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 va	Filter va								Filter total
	baseline	**baseline**	**baseline**	**SNV**	**1**	**0.12**	**0.81**	**0.97**	**0.88**	**0.24**	**217.0**	**12.0**	**0**
●	∪ SAGE$_{3.0}$	∪ (mutect2$_{\text{GATK 4.…}}$	∪ SAGE$_{3.0}$	SNV	1	0.04	0.89	0.97	0.92	0.07	235.67	13.33	4
○	∪ SAGE$_{3.0}$	∪ (cgpCaVEManWr…	baseline	SNV	1	0.04	0.88	0.97	0.92	0.08	234.33	13.33	3
○	baseline	∪ (cgpCaVEManWr…	baseline	SNV	1	0.04	0.87	0.97	0.92	0.09	232.0	13.33	2
○	baseline	∪ (mutect2$_{\text{GATK 4.…}}$	baseline	SNV	1	0.04	0.87	0.97	0.92	0.09	231.67	13.33	2
○	∪ SAGE$_{3.0}$	∪ (mutect2$_{\text{GATK 4.…}}$	baseline	SNV	1	0.04	0.88	0.97	0.92	0.09	234.0	13.33	3
○	baseline	∪ MuSE$_{2.0}$	baseline	SNV	1	0.04	0.86	0.96	0.91	0.14	228.33	12.67	1
○	∪ SAGE$_{3.0}$	baseline	∪ SAGE$_{3.0}$	SNV	1	0.13	0.83	0.97	0.89	0.22	221.0	12.0	2
○	∪ SAGE$_{3.0}$	baseline	baseline	SNV	1	0.12	0.82	0.97	0.89	0.23	219.33	12.0	1
○	∪ SAGE$_{3.0}$	baseline	∪ (MuSE$_{2.0}$ ∪ SAG…	SNV	1	0.13	0.83	0.96	0.89	0.23	222.0	12.0	3
○	baseline	baseline	∪ SAGE$_{3.0}$	SNV	1	0.12	0.82	0.97	0.89	0.24	218.67	12.0	1
○	baseline	baseline	baseline	SNV	1	0.12	0.81	0.97	0.88	0.24	217.0	12.0	0
○	baseline	baseline	∪ (MuSE$_{2.0}$ ∪ SAG…	SNV	1	0.12	0.82	0.96	0.88	0.25	219.67	12.0	2

C
Performance metrics (baseline vs harmonization)
Recall
center_a
center_a (baseline ∪ SAGE)
center_b
center_b (baseline ∪ (mutect2 ∩ Strelka2))
center_c
center_c (baseline ∪ SAGE)
SNV
1.00
0.75
0.50
0.25
SV
INDEL
Precision
center_a
center_a (baseline ∪ SAGE)
center_b
center_b (baseline ∪ (mutect2 ∩ Strelka2))
center_c
center_c (baseline ∪ SAGE)
SNV
1.00
0.75
0.50
0.25
SV
INDEL
F1 Score
center_a
center_a (baseline ∪ SAGE)
center_b
center_b (baseline ∪ (mutect2 ∩ Strelka2))
center_c
center_c (baseline ∪ SAGE)
SNV
1.00
0.75
0.50
0.25
SV
INDEL
D
F1 Score
center_a
center_a (baseline ∪ SAGE)
center_b
center_b (baseline ∪ (mutect2 ∩ Strelka2))
center_c
center_c (baseline ∪ SAGE)
SNV
1.00
0.75
0.50
0.25
SV
INDEL

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. . .)]