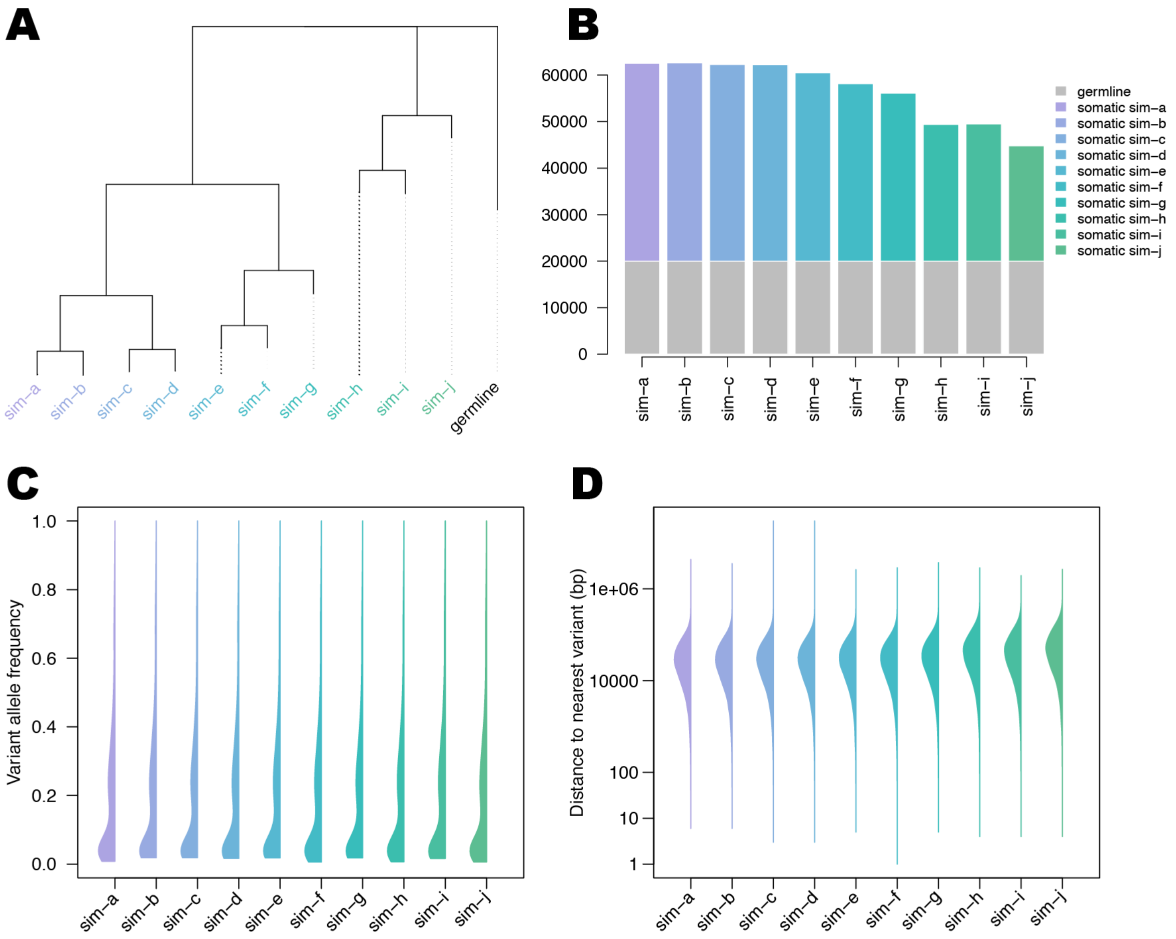
Supplementary Material for

Custom workflows to improve joint variant calling from multiple related tumour samples: FreeBayesSomatic and Strelka2pass

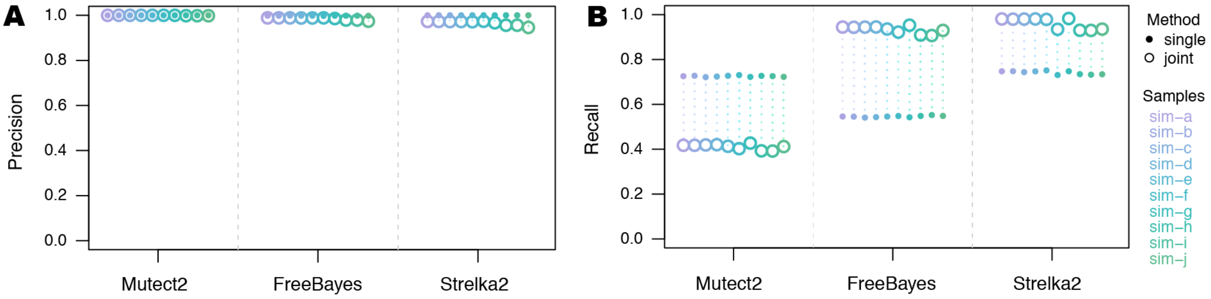
Hollizeck S.[1,2], Wong S. Q.[1,2], Solomon B.[1,2], Chandrananda D.[1,2,*], Dawson S-J.[1,2,3,*]

[1]Peter MacCallum Cancer Centre, Melbourne 3000, Victoria, Australia
[2]Sir Peter MacCallum Department of Oncology, University of Melbourne, Melbourne 3000, Victoria, Australia
[3]Centre for Cancer Research, University of Melbourne, Melbourne 3000, Victoria, Australia

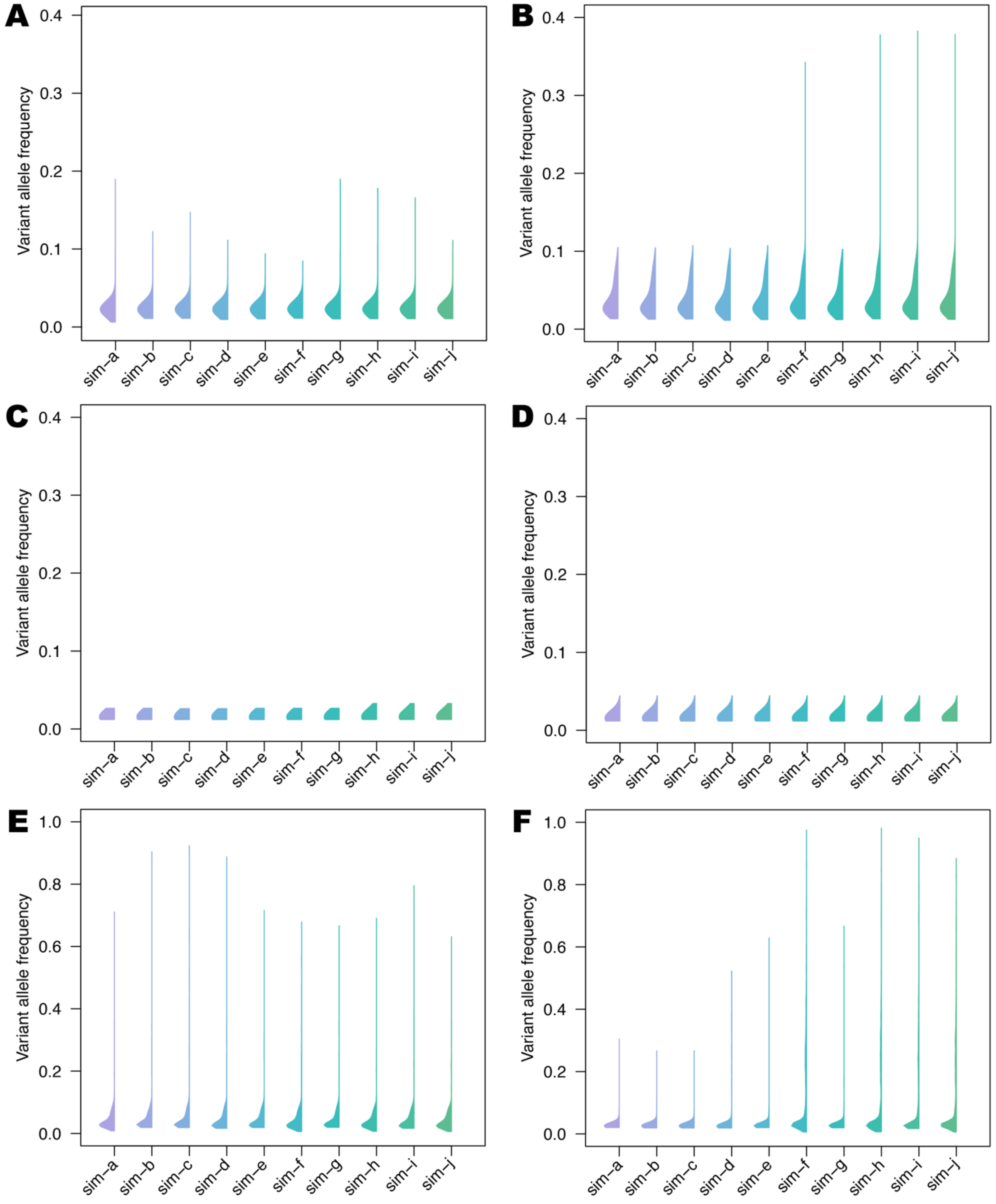
*DC and SJD are co-senior authors and contributed equally to this article



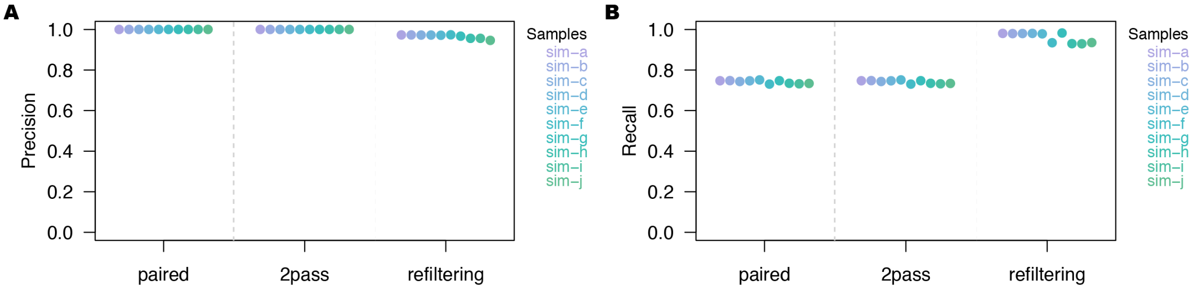
Sup Fig 1: Characteristics of simulated data; A) Simulated phylogeny of samples B) Number of simulated germline and somatic variants per sample C) Variant allele frequency distribution of simulated variants per sample D) Distance to nearest variant in each sample.



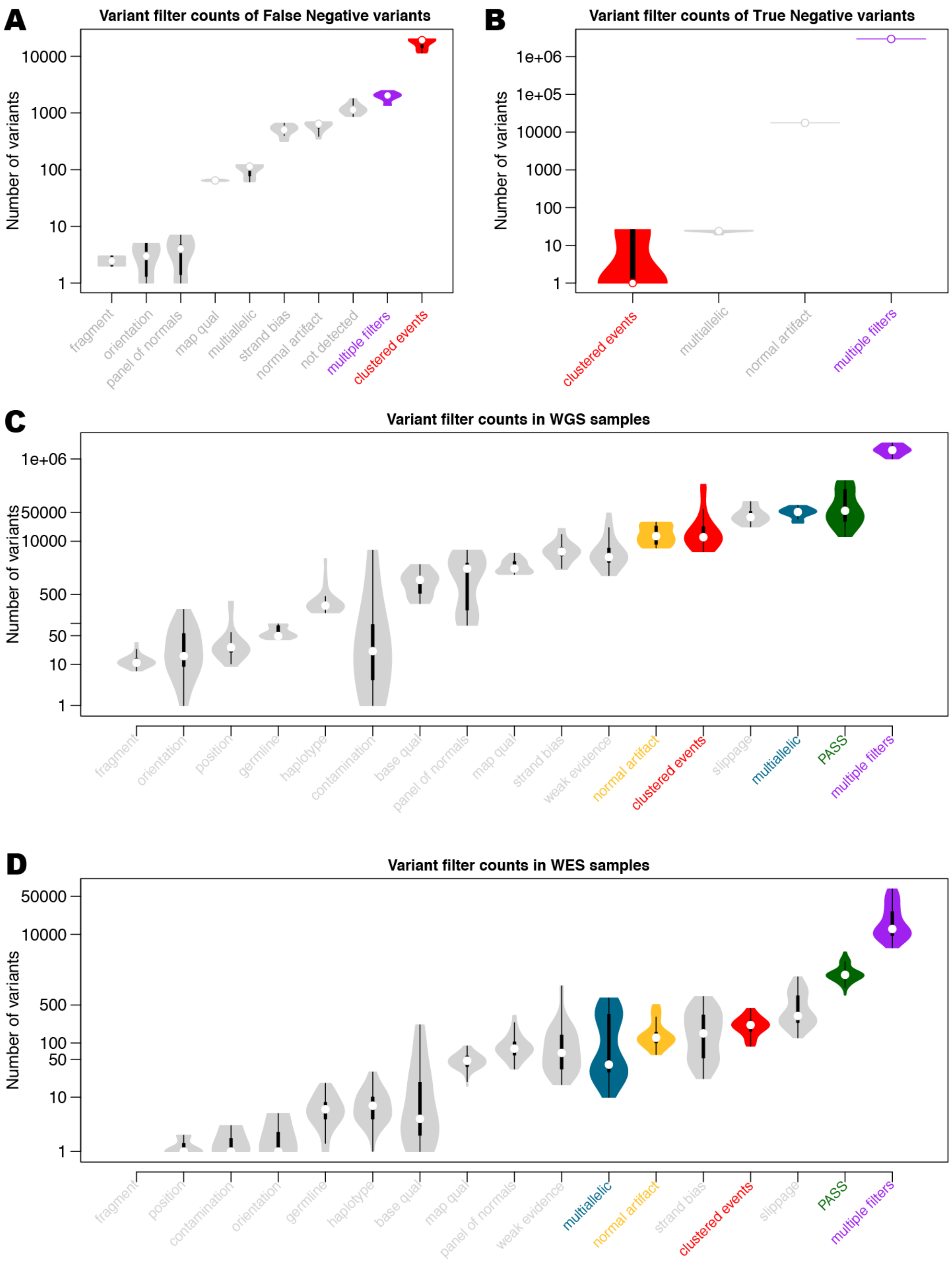
Sup Figure 2: Performance of workflows using simulated data: A) Precision and B) Recall of Mutect2, FreeBayes and Strelka2, run in single tumour-normal paired and joint calling configurations.



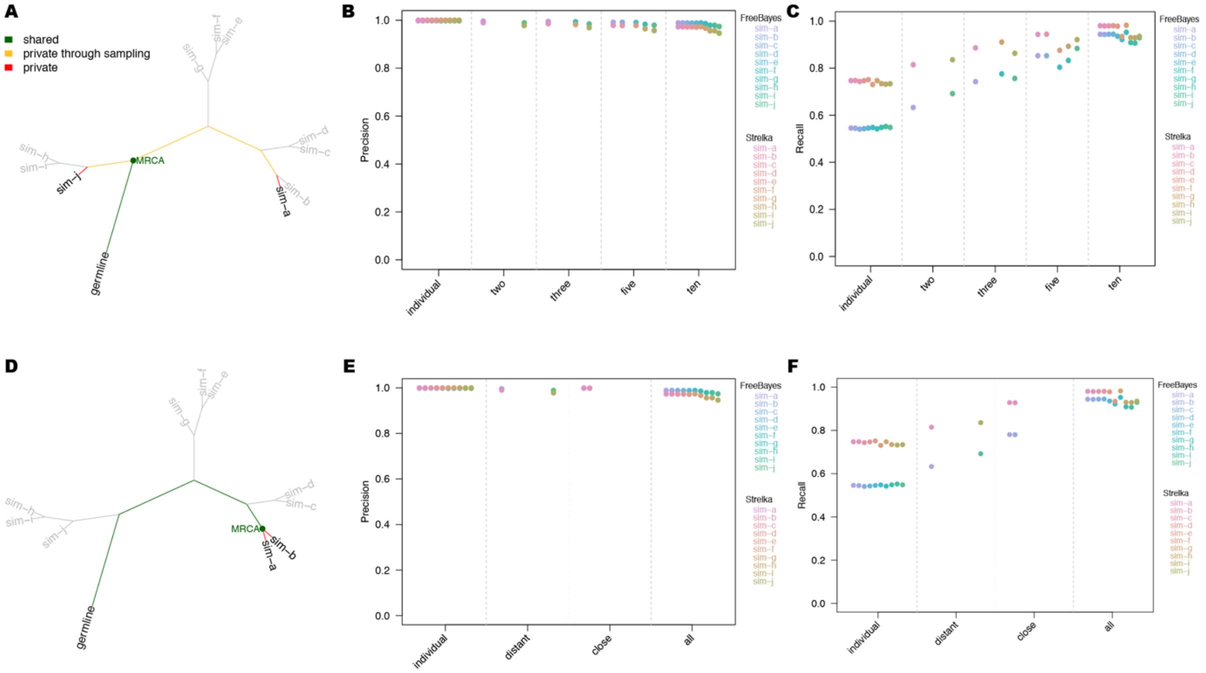
Supp Figure 3: Variant allele frequencies (VAF) of variants detected by joint sample analysis; A) VAF distribution of true positive variants additionally detected by Strelka2pass B) and FreeBayesSomatic C) VAF distribution of false positive variants additionally detected by FreeBayesSomatic D) and Strelka2pass E) VAF distribution of false negatives not called by FreeBayesSomatic F) and Strelka2pass.



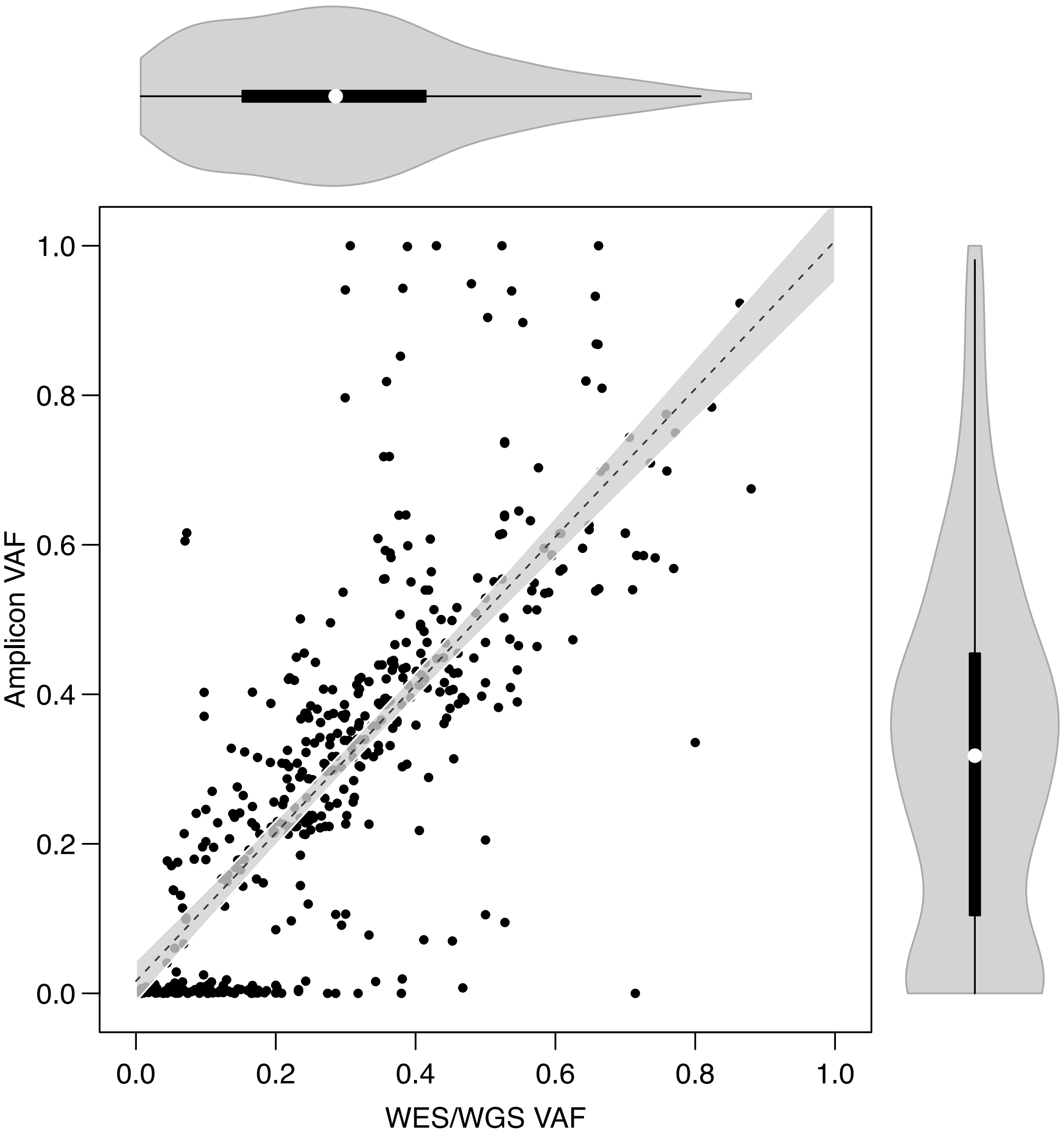
Sup Fig. 4: Performance of individual steps in the Strelka2pass workflow using the simulated data: A) Precision and B) Recall of tumour-normal paired analysis, two-pass step without refiltering (supplying variants from all tumour-normal pairs for evaluation) and two-pass step with refiltering (the final workflow).



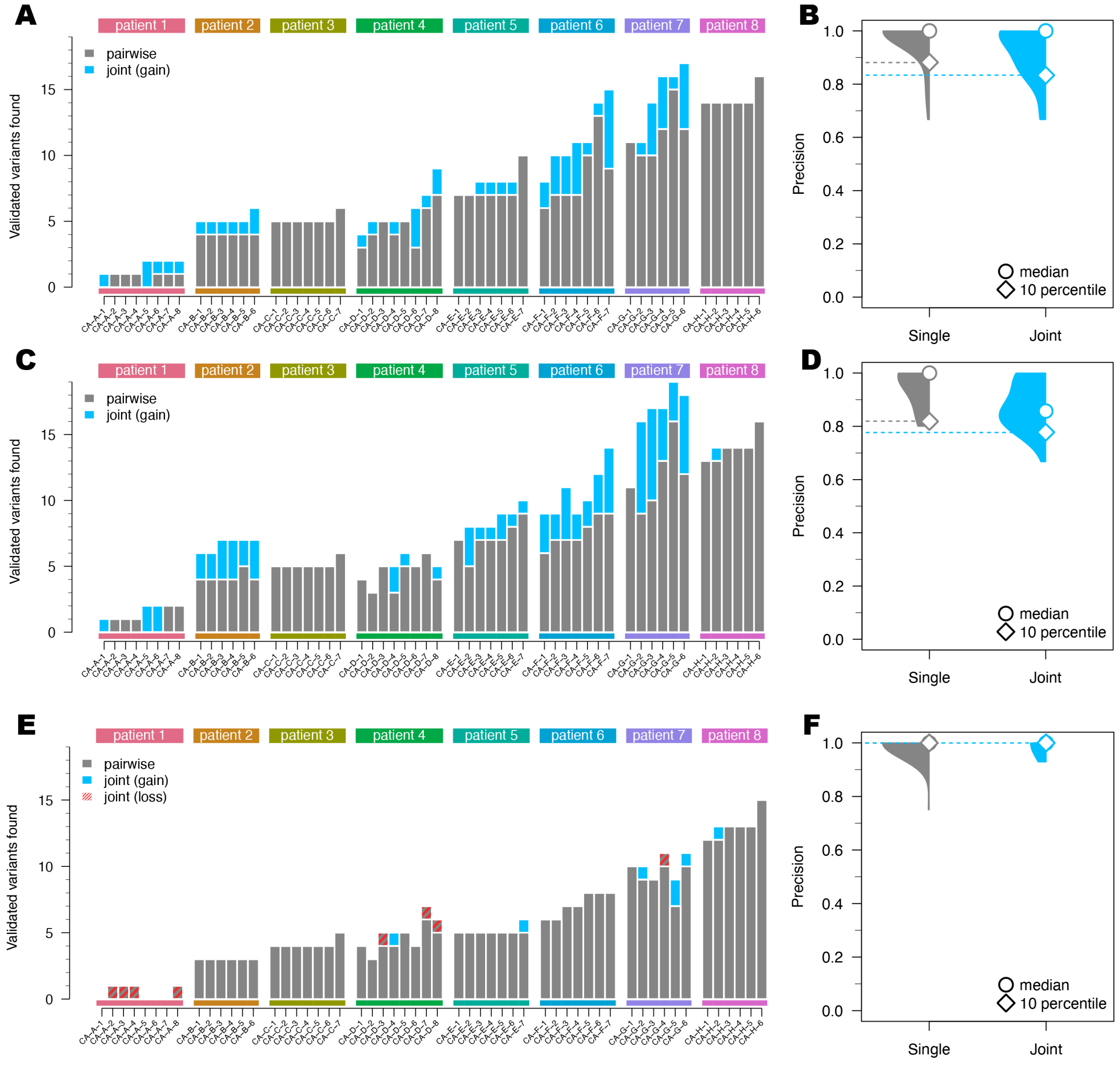
Supp Figure 5: Summary of variant filters assigned by Mutect2; The counts for each filter type are denoted by black boxplots with white circles depicting the median values. The fitted distribution of variant counts outlines each boxplot; A) Counts of filter assignments for false negative variants and B) true negative variants called by Mutect2 C) Filter assignment for all variants reported for sequenced patient data sequenced with WGS or D) WES.



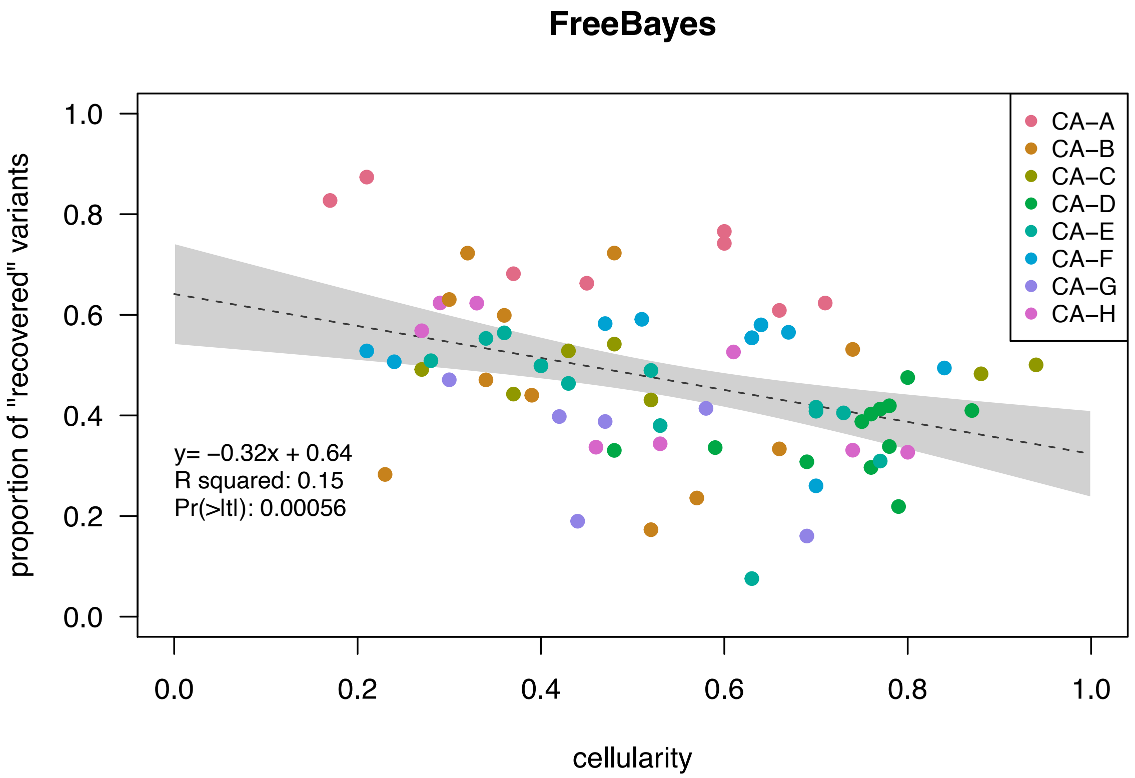
Sup Fig. 6: Assessing the performance of different workflows using tumour samples with different evolutionary relationships in the simulated data; A) Simulated phylogeny highlighting two samples with high evolutionary distance (sim-a and sim-j) where MRCA denotes the most recent common ancestor. B) Precision and C) Recall estimates of FreeBayes and Strelka, run in individual tumour-normal paired and joint calling configurations using two (sim-a and sim-j), three (sim-a, sim-g and sim-j), five (sim-a, sim-c, sim-f, sim-h and sim-j) and all ten tumour samples D) Simulated phylogeny highlighting two samples with low evolutionary distance (sim-a and sim-b). E) Precision and F) Recall estimates for FreeBayes and Strelka run in individual tumour-normal paired and joint calling configurations. The plots compare the performance of these workflows when using two evolutionary distant samples (sim-a and sim-j), two evolutionary close samples (sim-a and sim-b) and all ten tumour samples.



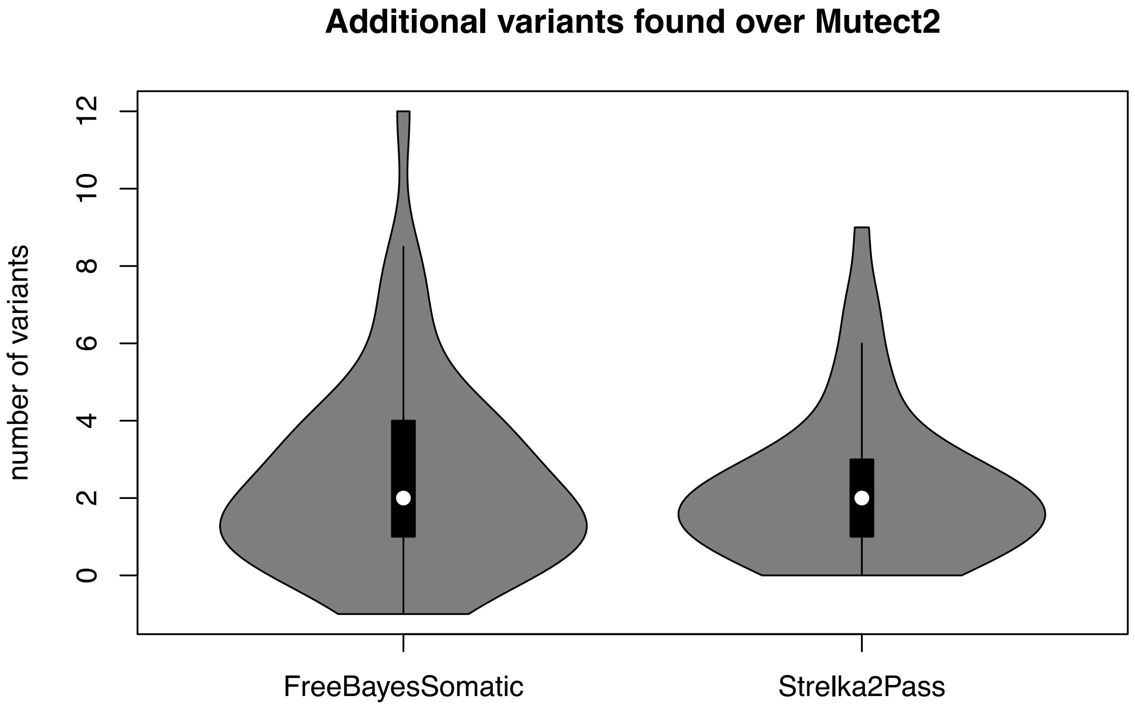
Sup. Fig. 7: Correlation of variant allele frequencies (VAF) from WES and WGS data against targeted amplicon sequencing VAF values with fitted violin plots of each individual distribution. Grey background shows 95% confidence interval for the fit of the linear model (dotted line)



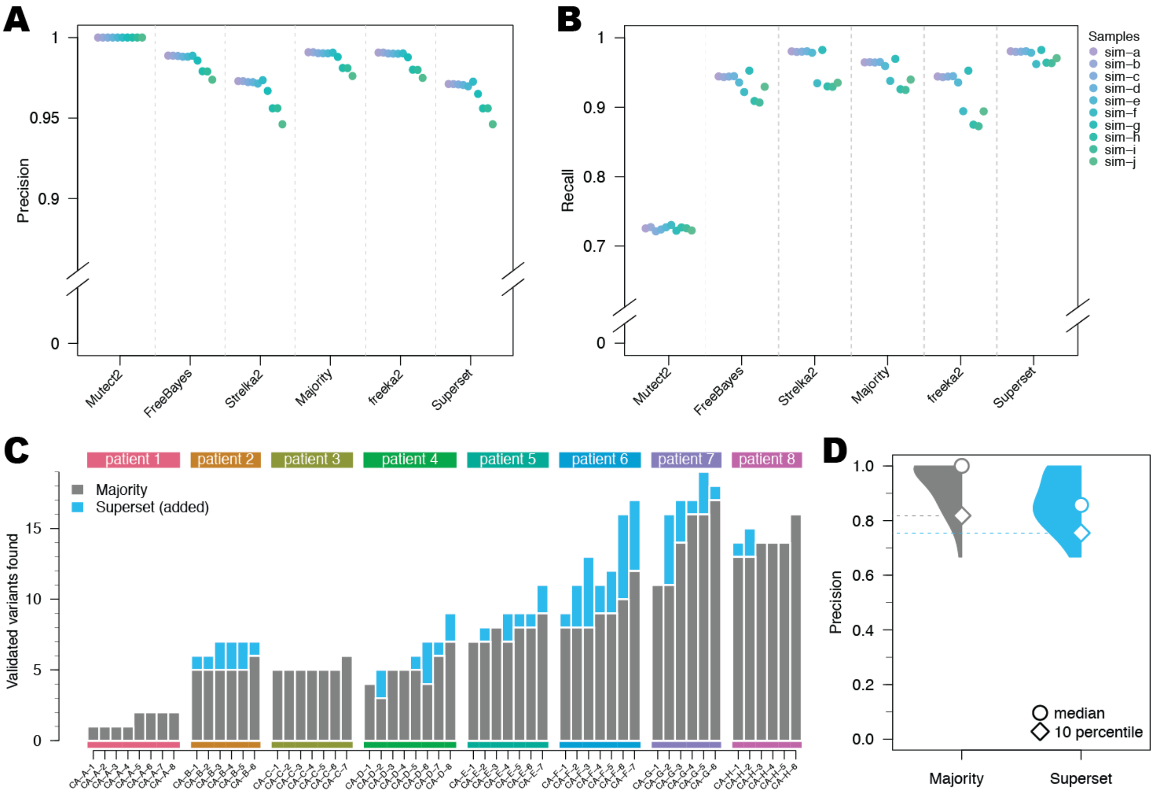
Sup Fig. 8: Performance of the different workflows using clinical samples from eight cancer patients: A) Number of variants called by Strelka2 run in the tumour-normal paired (grey) and joint calling configurations, which have been validated by targeted amplicon sequencing (TAS). The same for C) FreeBayes and E) Mutect2 workflows. Precision of tumour-normal paired and joint analysis of TAS validated clinical data for B) Strelka2, D) FreeBayes and F) Mutect2; Sup. Table 1 provides the sample naming map to the original publications.



Sup. Fig. 9: Correlation between cellularity and proportion of variants found only with joint calling using FreeBayesSomatic. Grey background shows 95% confidence interval for fit of linear model (dotted line)



Sup. Fig. 10: Improvement in recall using FreeBayesSomatic and Strelka2pass over Mutect2 in the clinical samples.



Supp Figure 11: Performance of ensemble variant calling strategies. A) Precision and B) Recall of variant detection using the joint multi-sample calling of each tool separately and compared to using Majority-vote ensemble calling (variant is called by at least two callers), Freeka2 (variant is called by both FreeBayesSomatic and Strelka2pass) and Superset (variant is called by either FreeBayesSomatic or Strelka2pass) for the simulated dataset D) Number of TAS validated variants found in the clinical samples with Majority-vote and Superset methods and the corresponding D) Precision estimates.

Sup Table 1: Sample naming map relating to previously published datasets. The first column contains sample names as they appear in this work, and the third column denotes how the samples are referred to in the original studies. Forth column shows the type of sequencing WES: whole-exome sequencing; WGS: whole genome sequencing

SAMPLE NAME	PUBLISHED STUDY	ORIGINAL NAME	SEQUENCING TYPE
CA-A-1	Solomon, et al. (2020)	Case 1 Left liver 1	WGS
CA-A-2		Case 1 Right occipital	
CA-A-3		Case 1 Right liver 2	
CA-A-4		Case 1 Right pleura	
CA-A-5		Case 1 Left lower lung lobe	
CA-A-6		Case 1 Left liver 5	
CA-A-7		Case 1 Right liver 3	
CA-A-8		Case 1 Left liver 2	
CA-B-1	Vergara, et al. (2021)	CAS-B-21-L-LUNG	WES
CA-B-2		CAS-B-22-R-LUNG	
CA-B-3		CAS-B-14B37035-1B	
CA-B-4		CAS-B-Primary-1	
CA-B-5		CAS-B-15B08317-3A	
CA-B-6		CAS-B-14B37035-1C	
CA-C-1		CAS-A-FR07935894	WGS
CA-C-2		CAS-A-FR07935905	
CA-C-3		CAS-A-FR07935906	
CA-C-4		CAS-A-FR07935907	
CA-C-5		CAS-A-FR07935908	
CA-C-6		CAS-A-FR07935916	
CA-C-7		CAS-A-FR07935918	
CA-D-1		CAS-G-91-2	WES
CA-D-2		CAS-G-75	
CA-D-3		CAS-G-74	
CA-D-4		CAS-G-71	
CA-D-5		CAS-G-91	
CA-D-6		CAS-G-76	
CA-D-7		CAS-G-94	
CA-D-8		CAS-G-72	
CA-E-1		CAS-D-70	WES
CA-E-2		CAS-D-61-3	
CA-E-3		CAS-D-66	
CA-E-4		CAS-D-68	
CA-E-5		CAS-D-64	
CA-E-6		CAS-D-61-2	
CA-E-7		CAS-D-62	
CA-F-1		CAS-C-41	WES
CA-F-2		CAS-C-40-Fresh	
CA-F-3		CAS-C-37	
CA-F-4		CAS-C-44	
CA-F-5		CAS-C-42-Fresh	
CA-F-6		CAS-C-43-Fresh	
CA-F-7		CAS-C-46-Primary	
CA-G-1		CAS-F-FR07935922	WGS
CA-G-2		CAS-F-FR07935915	
CA-G-3		CAS-F-FR07935913	
CA-G-4		CAS-F-FR07935909	
CA-G-5		CAS-F-FR07935904	
CA-G-6		CAS-F-FR07935903	
CA-H-1		CAS-E-1	WES
CA-H-2		CAS-E-3	
CA-H-3		CAS-E-4	
CA-H-4		CAS-E-10	
CA-H-5		CAS-E-6	
CA-H-6		CAS-E-8	

Sup Table 2: Runtime of different workflows on simulated data; The runtimes were generated on the Peter MacCallum Cancer Centre HPC cluster with Intel(R) Xeon(R) CPU E5-2660 v3 @ 2.60GHz. The times are displayed in single CPU runtime, but each workflow is highly parallelised, such that the user runtime is far lower.

	Number of tumour samples used for joint calling			
Method	2	3	5	10
FreeBayesSomatic	562h	811h	1185h	2292h
Strelka2pass	310h	465h	776h	1552h
Mutect2	-	-	-	28418h

Supplementary methods

Alignment of clinical data

Detailed information on processing of the clinical sequencing datasets was published previously (Solomon, et al., 2020; Vergara, et al., 2021). Briefly, reads were aligned to GRCh38 for patient CAS-A and GRCh37 for patients CAS-B through CAS-H using BWA version 0.7.17 (Li and Durbin, 2009) allowing the use of alternative contigs. Reads were then marked as duplicates with Picard software (v2.17.3).

Validation of clinical data

Detailed information on targeted amplicon sequencing of patient samples can be found in the original publications (Solomon, et al., 2020; Vergara, et al., 2021). A SNV called in WES with any workflow was considered a true positive when the adjusted p-value calculated through an exact binomial test was lower than 0.05 on the TAS data. The probability of success for this test was estimated as the number of bases different from the reference divided by the total number of sequenced bases (0.001) and the number of trials was the read depth covering the variant. For indels, a variant was considered to be validated if either of the panel variant callers *primal* (in house) or *canary* (Doig, et al., 2017) called the same variant.

Only amplicons with an average mapping rate of at least 80% over all samples, as well as an average coverage of more than 300 were considered for further analysis. WES variants were first subsetted to be within the area of the respective amplicons.

Purity estimation with sequenza

For CA-A the sequenza-utils python program was used to generate input files for the sequenza R program on the aligned BAM files (Favero, et al., 2014). Kmin and gamma were set to 100 and 500 respectively to discourage a highly fragmented result. For CA-B through -H the reported tumour purities were used from the publication (Vergara, et al., 2021).

Performance of individual steps in Strelka2Pass

As each of the three steps potentially has implications for the performance, we assessed the improvement provided by each step in the Strelka2pass workflow. Fig. S4 shows, that there is no change in either precision or recall just by supplying variants from all tumour-normal pairs for a second round of evaluation. However, there is a >20% improvement in recall when coupling this to the refiltering step that we have built into the workflow.

Ensemble workflows – user suggestions

An overall workflow can contain any number of additional variant callers, when not restricted to callers with joint analysis capability. Importantly, there is no benefit of jointly analysing samples with Mutect2, and it may decrease the performance in some cases. Each of our presented workflows outperformed Mutect2 on the data shown here, so when assembling an ensemble method, these methods, should have a higher confidence assigned to them in joint analysis cases, than tumour-normal pair approaches.

Depending on the end needs of the user, an ensemble workflow can be optimised towards precision or recall. In Sup Fig. 11 we show the performance changes improvement that can be achieved by combining Mutect2 in tumour-normal paired analysis with the two new workflows FreeBayesSomatic and Strelka2Pass. First, in a “best of three” majority vote, where the variant needs to be called by two out of three variant callers, we enhance the precision of each of the individual tools, with slightly lower recall.

On the other hand, with the super set approach, where any variant called in either FreeBayesSomatic or Strelka2Pass is included in the end result, this improves the recall even further, but slightly reduces the precision. This approach has the additional benefit of not needing to run Mutect2 which is an order of magnitude slower in our tests, than Strelka2Pass and FreeBayesSomatic (Sup. Table 2).

The usage of these workflows can be easily integrated into existing workflows and can be customised to the needs of the user.

References

Doig, K.D.*, et al.* Canary: an atomic pipeline for clinical amplicon assays. *BMC Bioinformatics* 2017;18(1).

Favero, F.*, et al.* Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Annals of Oncology* 2014;26(1):64–70-64–70.

Li, H. and Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009;25(14):1754–1760-1754–1760.

Solomon, B.J.*, et al.* RET Solvent Front Mutations Mediate Acquired Resistance to Selective RET Inhibition in RET-Driven Malignancies. *Journal of Thoracic Oncology* 2020;15(4):541–549-541–549.

Vergara, I.A.*, et al.* Evolution of late-stage metastatic melanoma is dominated by aneuploidy and whole genome doubling. *Nature Communications* 2021;12(1).