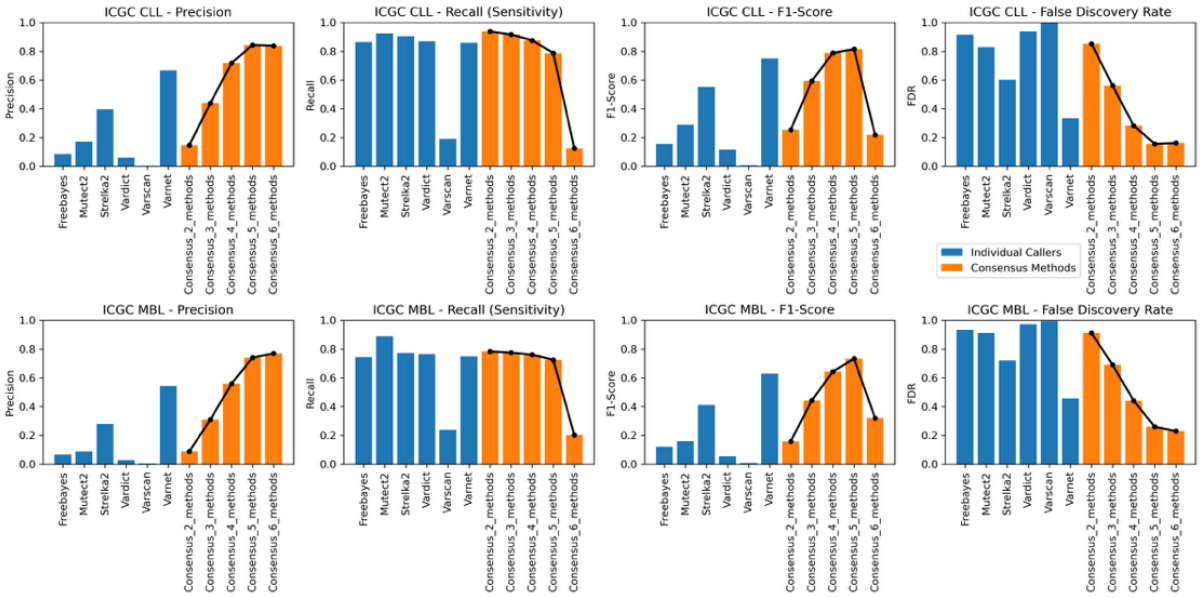
Comprehensive benchmark of somatic mutation callers on cell-free DNA Next-Generation Sequencing data in cancer liquid biopsy

Hanaé Carrié[1,2,3], Ngak Leng Sim[1], Pui Mun Wong[1], Anna Gan[1], Yi Ting Lau[1], Polly Poon[1], Saranya Thangaraju[1] Iain Tan[1,4], Yoon Sim Yap[4], Kiran Krishnamachari[1,3], Limsoon Wong[3], and Anders Skanderup[1,]*

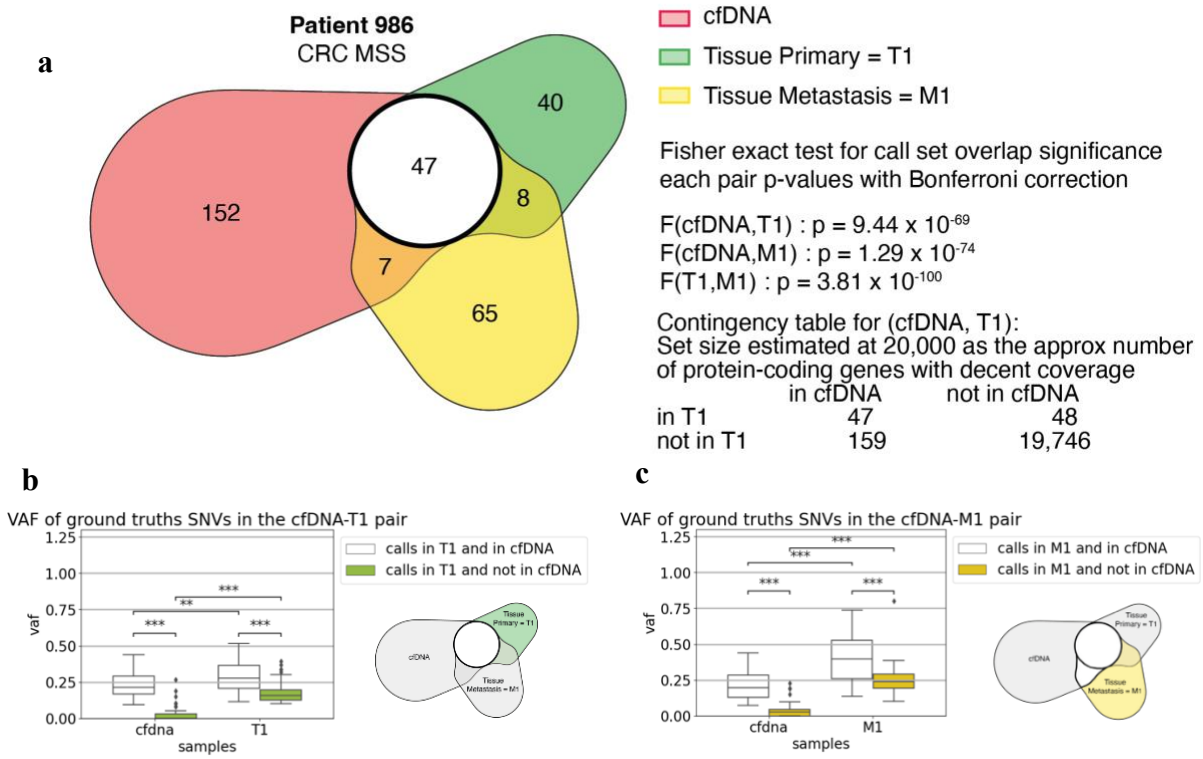
[1] Genome Institute of Singapore, Agency for Science Technology and Research
[2] Institute of Data Science, National University of Singapore
[3] Department of Computer Science, National University of Singapore
[4] National Cancer Center Singapore, Singapore
** Corresponding author*

Supplementary Information



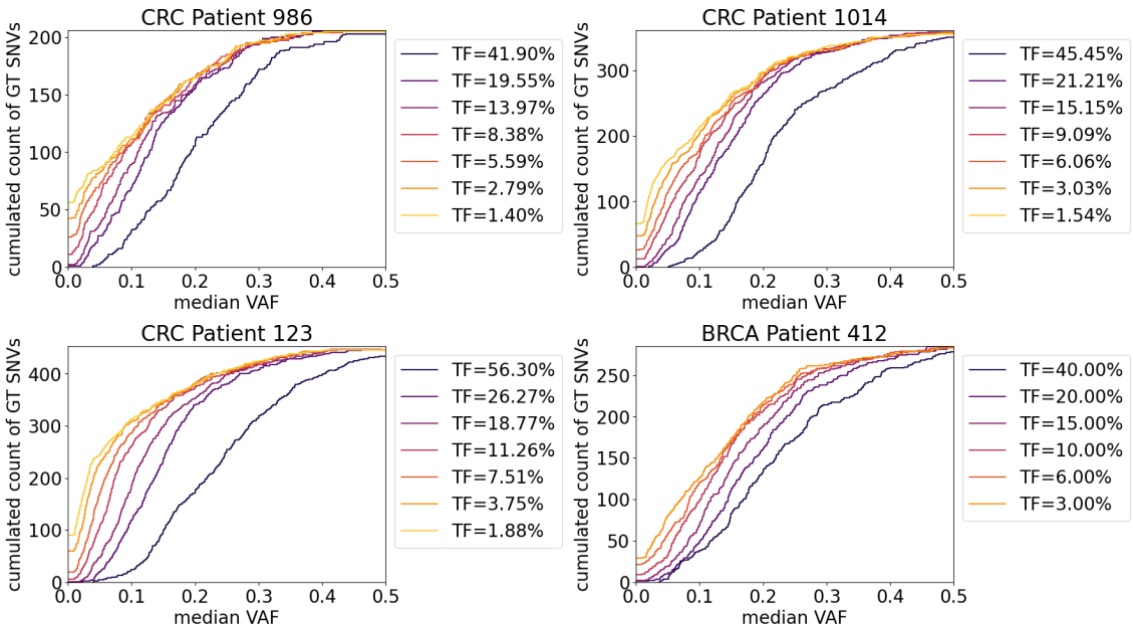
Supplementary Figure 1: Evaluation of k-of-N consensus calling in ICGC benchmark datasets

Precision/Recall/F1/FDR for individual callers (blue) and consensus approaches (orange, k-of-N); elbow at 4/6 and 5/6 (F1) supports a majority voting scheme.



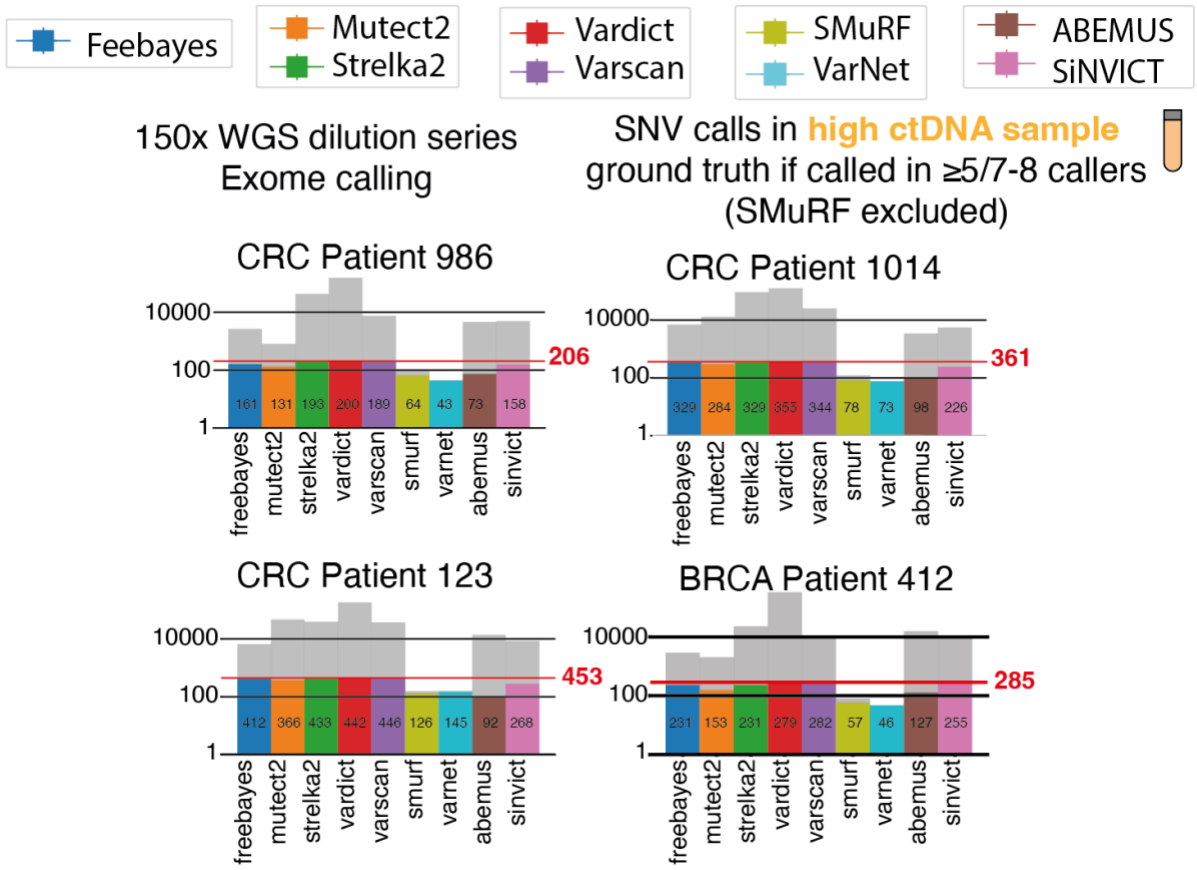
Supplementary Figure 2: Pseudo ground truth concordance between cell-free DNA (100x, WGS) and tumor tissue (100x, WGS), restricted to the exome

(a) Venn diagram showing the overlap in consensus exomic calls using the high ctDNA sample (K = 5 callers out of 8) vs. the tumor - primary and/or metastasis - tissue samples (K = 3 callers out of 6) for colorectal cancer patient 986. SMuRF is excluded from ground truth generation as it is an ensemble caller. For tissue samples, cfDNA- specific methods ABEMUS and SiNVICT are not applied. Fisher exact test (two-sided, with Bonferroni correction) based on contingency tables is applied on each pair of call sets to quantify overlap significance. (b-c) VAF distributions for tissue-specific ground-truth SNVs in T1 **(**b**)** and M1 **(**c**)**. White boxplots denote variants also detected in high-ctDNA plasma; colored boxplots denote variants absent from plasma. Tissue-only variants show lower VAFs in tissue - consistent with subclonality - and ultra-low/zero VAFs in plasma, explaining non-detection. (Boxes: median/IQR; whiskers: 1.5×IQR.) The statistical tests performed are two-sided two-sample independent t-tests without correction for multiple testing. ctDNA: circulating tumor DNA, VAF: Variant Allele Frequency.



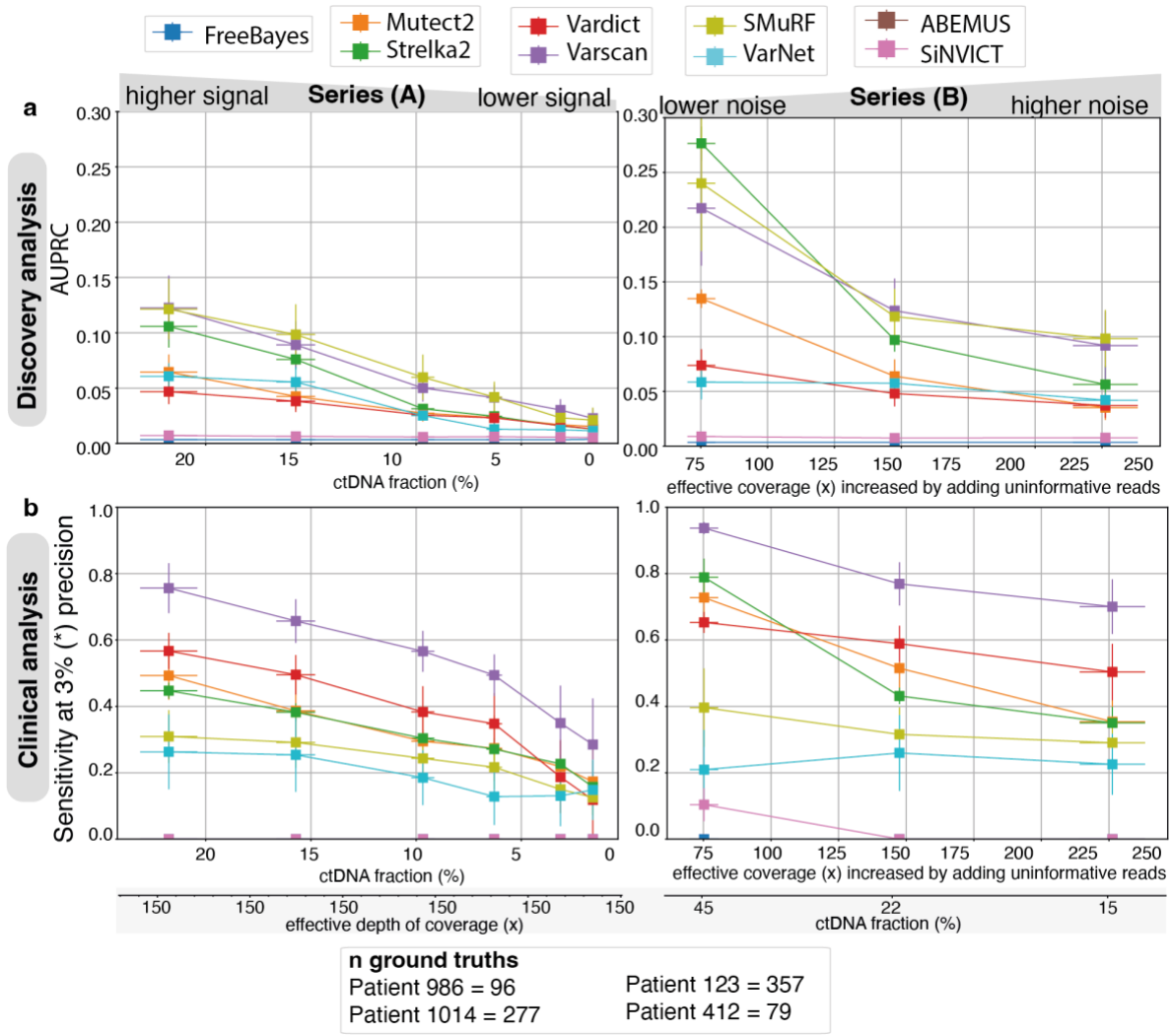
Supplementary Figure 3: Cumulative variant allele frequency of pseudo ground truths Single Nucleotide Variants in each dilution series (A) for the four cancer patients

For each ground truth, the median variant allele frequency value reported by all nine studied callers is considered. The sample with the highest TF for each patient corresponds to the undiluted high ctDNA sample at 70x coverage depth taken as reference to build consensus ground truths. The other samples are diluted while maintaining a fixed 150x coverage depth forming the decreasing signal series (A). TF: Tumor Fraction. ctDNA: circulating tumor DNA.



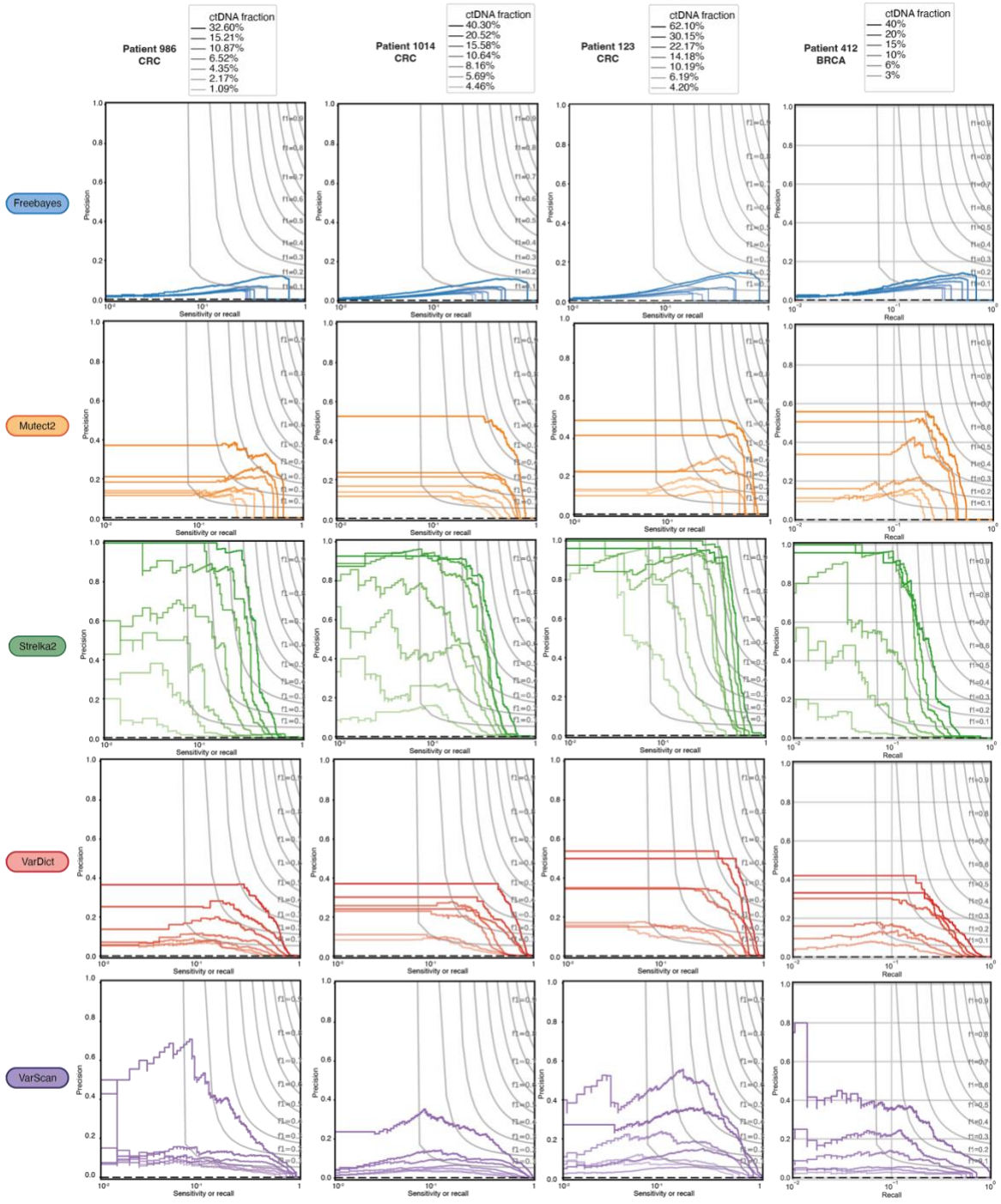
Supplementary Figure 4: Pseudo ground truth repartition from consensus calling on the high circulating tumor DNA sample

Bar plots showing for each studied patient the number of Single Nucleotide Variants passing default filters and thresholds called in the exomic regions of the ~100x WGS high ctDNA sample. The figure is log-scaled on the y-axis due to the wide range of calls that can be predicted by each method. Calls called by more than 5 callers among 8 callers are considered ground truths. Note that SMuRF is excluded for ground truth generation as it is an ensemble caller. Calls falling within the pseudo ground truths list are colored (True Positives) while others are shown in grey (False Positives). The red line indicates the number of consensus pseudo ground truths calls. ctDNA: circulating tumor DNA.



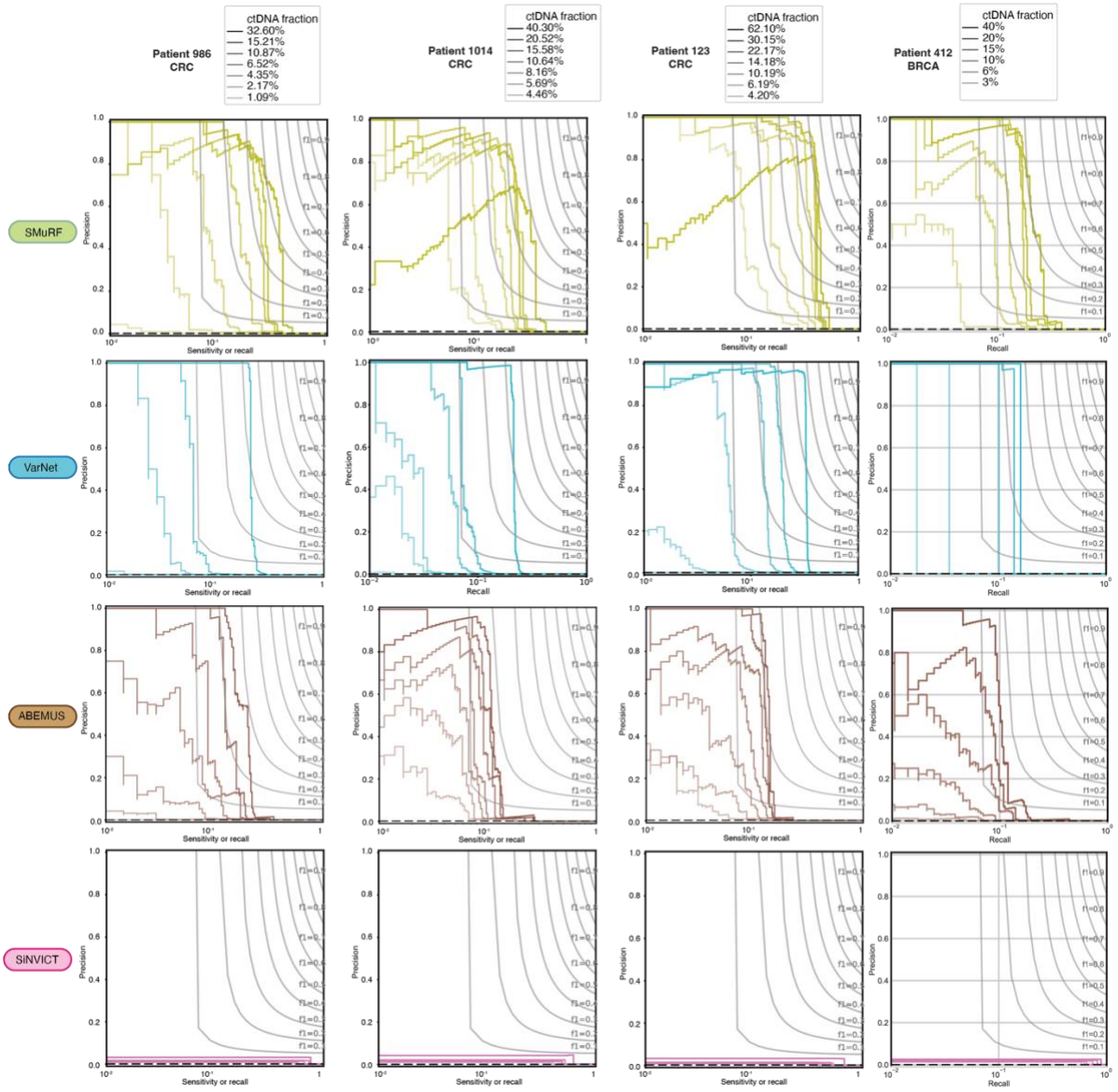
Supplementary Figure 5: Performance results in INDEL calling on created 150x dilution series over the four studied cancer patients

Callers performance on INDELs in exonic regions averaged on the n=4 patients in both benchmark series: the decreasing signal series (A) where TF decreases at a fixed 150x depth of coverage (left, 6 samples) and the increasing noise series (B) where coverage artificially increases (from 70x to 250x) at a fixed number of informative reads (70x) (right, 3 samples). Callers are evaluated in two application settings: **(a)** discovery analysis with the Area Under Precision Recall Curve metric and **(b)** clinical analysis with the sensitivity at a precision fixed to 3%. ABEMUS is not included as it does not handle INDEL calling. The vertical errors bars show the standard error of the mean (s.e.m.) of the averaged metric curves and the horizontal error bars show the s.e.m. of the ctDNA fraction estimates among patients (n=4). INDEL ground-truth counts (n) per patient are shown in the small box: CRC_986 n= 96, CRC_1014 n=277, CRC_123 n=357, BRA_412 n= 79 (pooled N= 809). Here, n denotes the number of true somatic INDELs in the benchmark for that patient. INDEL: Insertion/Deletion. ctDNA: circulating tumor DNA.



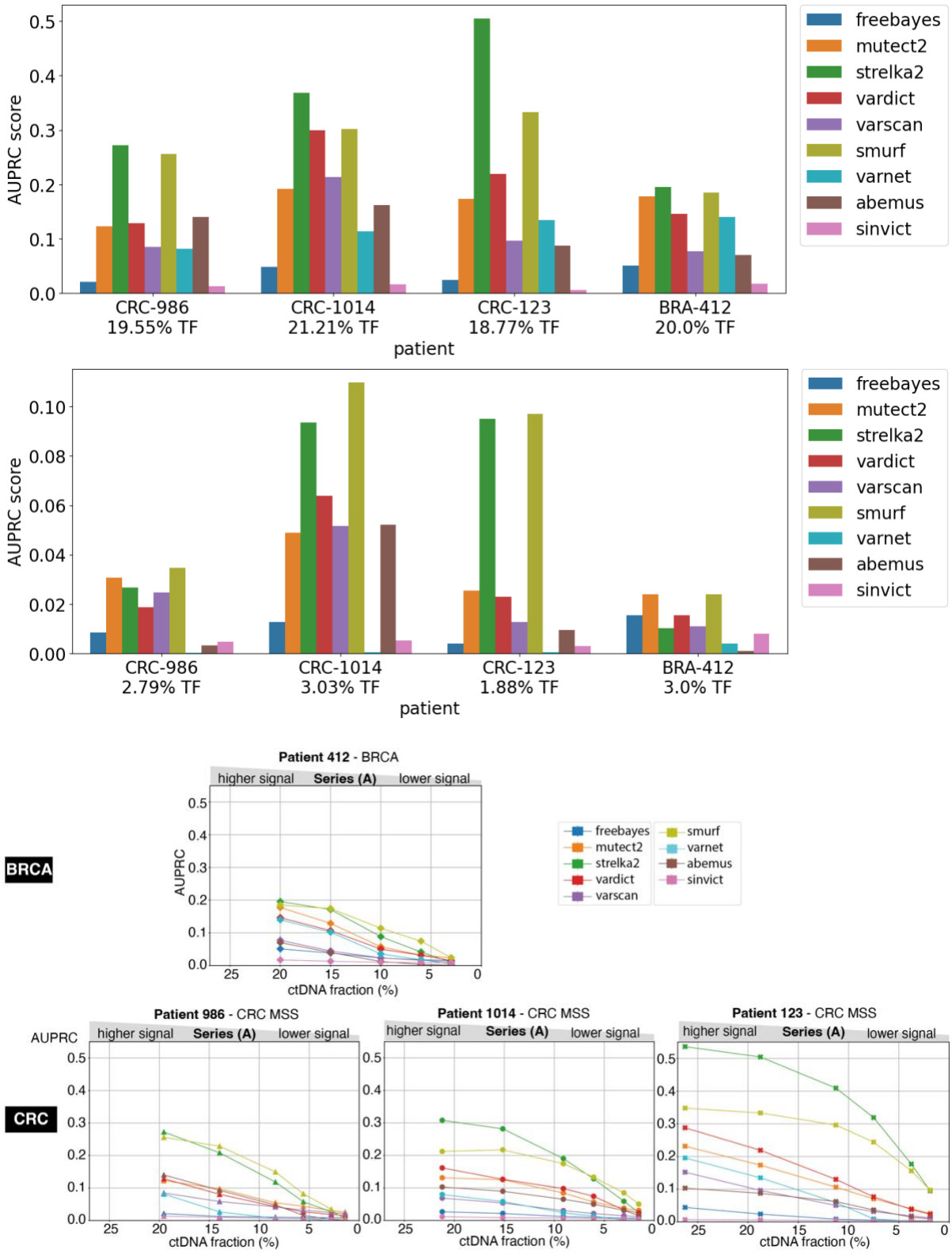
Supplementary Figure 6a: Precision-Recall curves for each somatic mutation caller for Single Nucleotide Variants on each individual studied cancer patient

Namely, the patients are: CRC patient 986, CRC patient 1014, CRC patient 123, and BRCA patient 412. Callers: FreeBayes, Mutect2, Strelka2, VarDict, VarScan. CRC: Colorectal Cancer, BRCA: Breast Cancer.



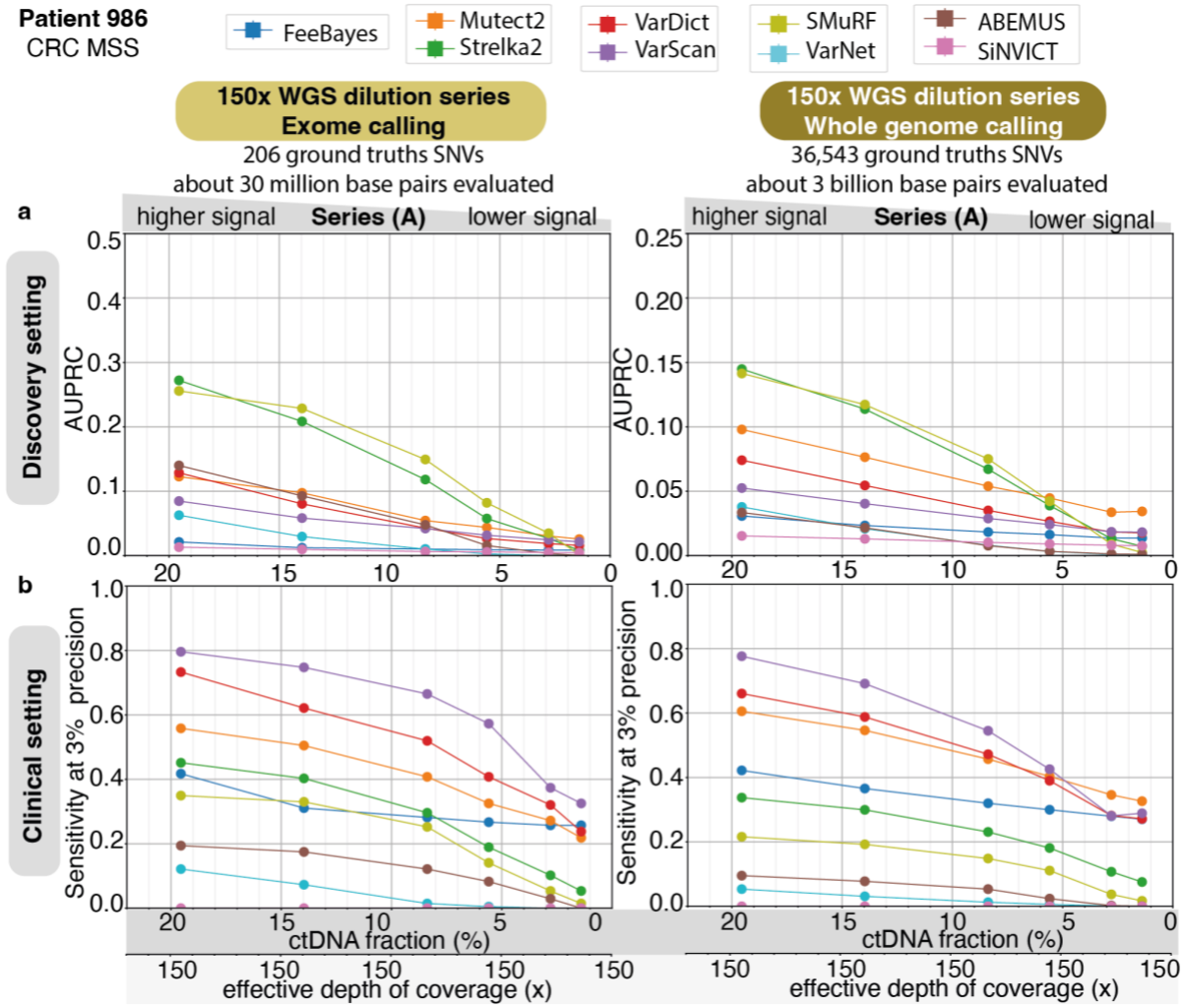
Supplementary Figure 6b: Precision-Recall curves for each somatic mutation caller for Single Nucleotide Variants on each individual studied cancer patient

Namely, the patients are: CRC patient 986, CRC patient 1014, CRC patient 123, and BRCA patient 412. Callers: SMuRF, VarNet, ABEMUS, SiNVICT. CRC: Colorectal Cancer, BRCA: Breast Cancer.



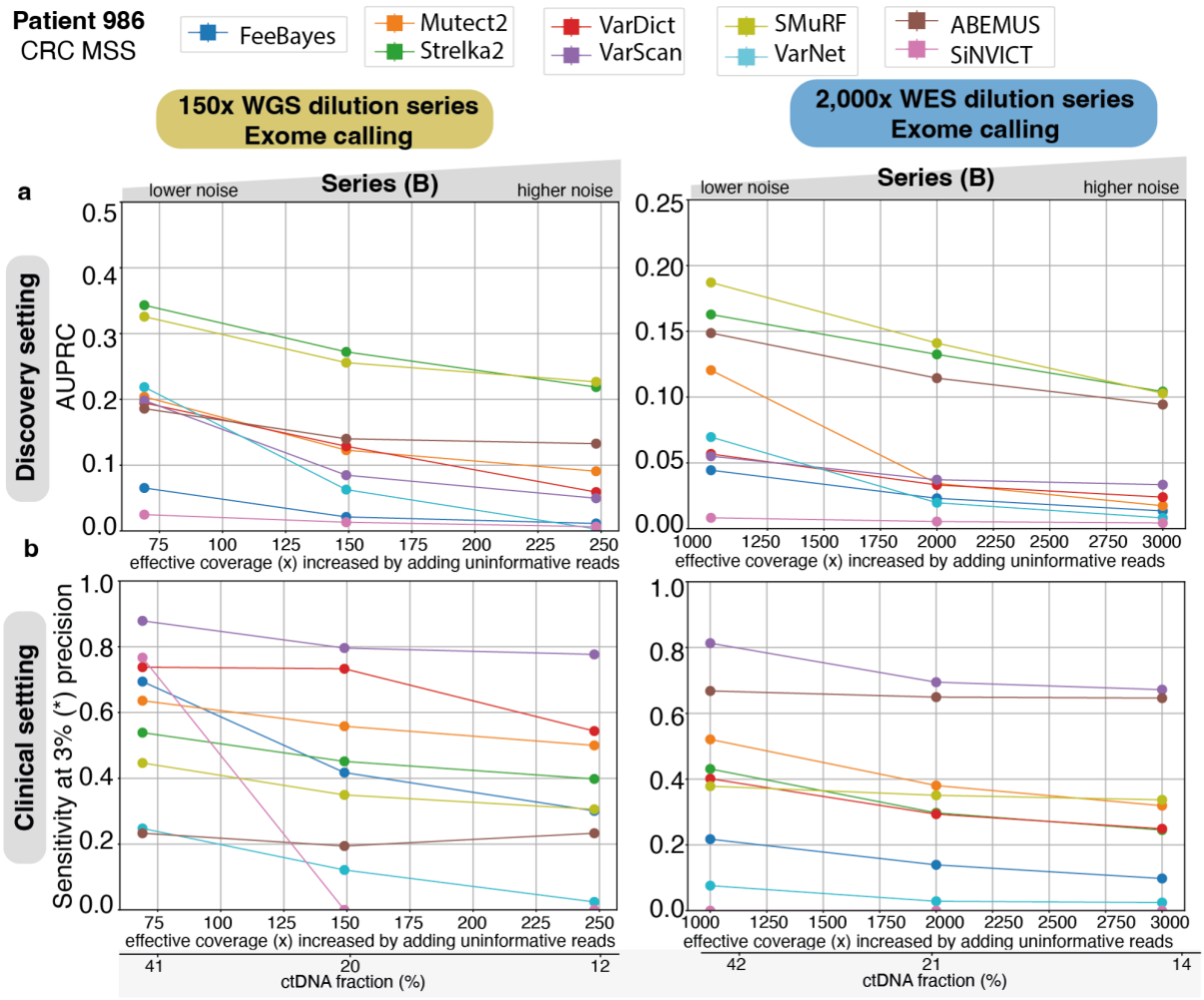
Supplementary Figure 7: Results concordance between the two widespread Colorectal and Breast Cancer types

Callers' performance on Single Nucleotide Variants in exonic regions for each individual patient is displayed in the decreasing signal series (A). The top barplots show the side-by-side AUPRC for each patient in one high TF (TF~20%) and one low TF (TF~3%) benchmark dilution. The bottom plots show the AUPRC across all benchmark dilutions for each patient. AUPRC: Area Under Precision-Recall Curve, TF: Tumor Fraction.



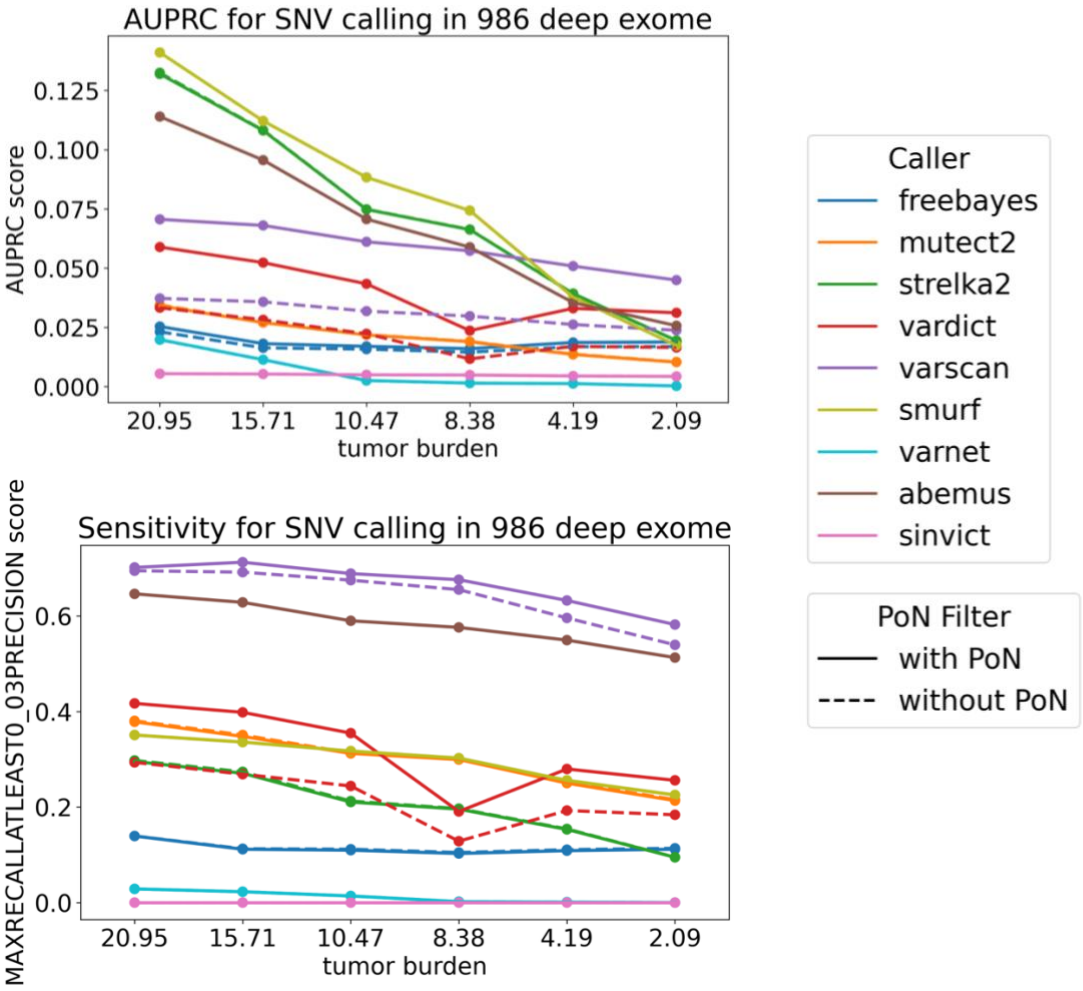
Supplementary Figure 8: Results concordance between exome and whole genome calling

Callers' performance on Single Nucleotide Variants in Colorectal Cancer patient 986 in the decreasing signal dilution series (A) on 150x Whole Genome Sequencing samples when applying callers either on exome (left) or on whole genome (right) in both application settings: for **(a)** discovery analysis with the Area Under Precision- Recall Curve metric and for **(b)** clinical analysis with the sensitivity at precision to 3%.



Supplementary Figure 9: Impact of sequencing depth (150x vs 2,000x) on callers' performance in increasing noise series (B)

Callers' performance on Single Nucleotide Variants in exonic regions of Colorectal Cancer patient 986 in dilution series (B) at two different fixed coverage levels: 150x (left, derived from Whole Genome Sequencing samples) and 2,000x (right, derived from Whole Exome Sequencing samples). Callers are evaluated in two application settings: **(a)** discovery analysis with the Area Under Precision-Recall Curve metric with a zoomed version of the ctDNA fraction range [1-5%] and **(b)** clinical analysis with the sensitivity at a precision fixed to 3%. ctDNA: circulating tumor DNA.



Caller	Max Δ AUPRC	Max Δ Sensitivity at 3% Precision	Interpretation of PoN effect
VarScan	+ 0.0334	+ 0.0425	significant performance improvement
VarDict	+ 0.0256	+ 0.1294	significant performance improvement
FreeBayes	+ 0.0022	0	very slight performance improvement
Mutect2	0	0	No effect
Strelka2	0	0	No effect
SMuRF	0	0	No effect
VarNet	0	0	No effect
SiNVICT	0	0	No effect
ABEMUS	NA	NA	only with PoN result available

Supplementary Figure 10: Effect of a unified Panel-of-Normals (PoN) post-filter on SNV calling (patient 986 deep-WES)

(a) Area under the precision–recall curve (AUPRC) versus tumor burden (% TF). (b) Sensitivity at a fixed 3% precision versus tumor burden. Solid lines show results with the unified PoN; dashed lines show results without PoN. Colors indicate callers (FreeBayes, Mutect2, Strelka2, VarDict, VarScan, SMuRF, VarNet, ABEMUS, SiNVICT). The PoN was trained on unmatched normals from the same capture/pipeline and applied post hoc and identically to all VCFs. (c) Summary of maximum changes (post-PoN − pre-PoN) across burdens: VarScan ΔAUPRC +0.0334, ΔSensitivity +0.0425; VarDict ΔAUPRC +0.0256, ΔSensitivity +0.1294; FreeBayes ΔAUPRC +0.0022, ΔSensitivity 0; Mutect2, Strelka2, SMuRF, VarNet, SiNVICT: no measurable effect. ABEMUS: only PoN-enabled outputs available (N/A deltas).

Patient ID *Sex* *Age at Diagnosis* *Cancer Type and Subtype* *Site of Primary Tumor* *Stage*	*Sample characteristics* *Timepoint* *Sequencing type* *(depth of coverage)*	*Tumor Fraction (TF) estimate*	*Intermediate values to build TF estimate and additional values for orthogonal verification*
Patient ID: 986 Sex: Female Age at Diagnosis: 66 Cancer Type and Subtype: CRC ADC Site of Primary Tumor: Ascending Colon Stage: IVB Microsatellite Status: MSS Tumor markers: NA	**high ctDNA sample** timepoint: 40 days deep WGS (71.63x)	32.6%	ichorCNA: **23%** median of 4 purity estimates: **41.9%** max VAF: **40.7%** 2 x median VAF: 2 x 21.1% = **42.2%**
	ultra-low ctDNA sample timepoint: 454 days deep WGS (~100x)	0%	ichorCNA: **8%** ichorCNA (matched lpWGS): **0%** max VAF: **0%** 2 x median VAF: 2 x 0% = **0%**
Patient ID: 1014 Sex: Male Age at Diagnosis: 46 Cancer Type and Subtype: CRC ADC Site of Primary Tumor: Transverse Colon Stage: IVA Microsatellite Status: MSS Tumor markers: NA	**high ctDNA sample** timepoint: 901 days deep WGS (80.84x)	40.3%	ichorCNA: **40%** median of 4 purity estimates: **45.5%** max VAF: **44.8%** 2 x median VAF: 2 x 20.3% = **40.6%**
	ultra-low ctDNA sample timepoint: 800 days deep WGS (~100x)	2.56%	ichorCNA: **5.1%** ichorCNA (matched lpWGS): **0%** max VAF: **1.9%** 2 x median VAF: 2 x 1.61% = **3.22%**
Patient ID: 123 Sex: Male Age at Diagnosis: 49 Cancer Type and Subtype: CRC ADC Site of Primary Tumor: Rectum Stage: IIIA Microsatellite Status: MSS Tumor markers: NA	**high ctDNA sample** timepoint: 1641 days deep WGS (71.63x)	62.1%	ichorCNA: **65.4%** median of 4 purity estimates: **47%** max VAF: **56.5%** 2 x median VAF: 2 x 29.4% = **58.8%**
	ultra-low ctDNA sample timepoint: 1745 days deep WGS (~100x)	3.15%	ichorCNA: **6.1%** ichorCNA (matched lpWGS): **0%** max VAF: **4.1%** 2 x median VAF: 2 x 1.10% = **2.20%**
Patient ID: 412 Sex: Female Age at Diagnosis: 33 Cancer Type and Subtype: BRCA IDC Site of Primary Tumor: Right Breast Stage: IIIA Microsatellite Status: NA Tumor markers: ER+/PR+/HER2-	**high ctDNA sample** timepoint: 1414 days deep WGS (71.63x)	67.6%	ichorCNA: **66.4%** median of 4 purity estimates: **NA** max VAF: **68.9%** 2 x median VAF: **NA**
	ultra-low ctDNA sample timepoint: 1218 days deep WGS (~100x)	0%	ichorCNA: **0%** ichorCNA (matched lpWGS): **0%** max VAF: **0%** 2 x median VAF: 2 x 0% = **0%**

Supplementary Table 1: Patient Clinical Details and Tumor Fraction estimates of initial samples before dilutions

Selected cancer plasma samples detailed information to build a reliable TF estimate for each initial deep WGS sample. For high ctDNA samples, the TF is calculated as the mean of ichorCNA [44] TF and twice the median VAF of known cancer driver mutations which were first screened in the matched deep targeted sample. For ultra-low ctDNA samples, the TF is taken as the mean of the max VAF and of twice the median VAF of those known cancer driver mutations. The TF estimate relied on VAF in this case as the ichorCNA [44] TF estimate found was not consistent between the matched shallow and deep WGS samples. CRC: Colorectal Cancer; ADC: Adenocarcinoma; MSS: Macrosatellite Stable; BRCA: Breast Cancer; IDC: Invasive Ductal Carcinoma.

Dilution ID	from low ctDNA sample	from high ctDNA sample	Total depth of coverage	Approximative TF estimate	Type of dilution series
0	70x	0x	70x	40%	Reference
1	70x	80x	150x	20%	Series (A) (6 mixtures excluding reference)
2	50x	100x	150x	15%	
3	30x	120x	150x	10%	
4	20x	130x	150x	6%	
5	10x	140x	150x	3%	
6	5x	145x	150x	1.5%	
0	70x	0x	70x	40%	Series (B) (3 mixtures including reference)
1	70x	80x	150x	20%	
8	70x	180x	250x	12%	

Supplementary Table 2: Dilution ratio choice taken in experimental design for dilution series creation at 150x whole genome sequencing

Sample ID = 0 will be the reference undiluted high ctDNA sample. From a pair of high/ultra-low ctDNA samples of one cancer patient, 8 samples are created for Series (A), Series (B), and reference sample. ctDNA: circulating tumor DNA, TF: Tumor Fraction.

Dilution ID	from low ctDNA sample	from high ctDNA sample	Total depth of coverage	Approximative TF estimate	Type of dilution series
0	1,000x	0x	1,000x	40%	Reference
1	1,000x	1,000x	2,000x	20%	Series (A) (6 mixtures excluding reference)
2	750x	1,250x	2,000x	15%	
3	500x	1,500x	2,000x	10%	
4	400x	1,600x	2,000x	6%	
5	200x	1,800x	2,000x	3%	
6	100x	1,900x	2,000x	1.5%	
0	1,000x	0x	1,000x	40%	Series (B) (3 mixtures including reference)
1	1,000x	1,000x	2,000x	20%	
8	1,000x	1,900x	2,000x	12%	

Supplementary Table 3: Dilution ratio choice taken in experimental design for dilution series creation at 2,000x whole exome sequencing

Sample ID = 0 will be the reference undiluted high ctDNA sample. From a pair of high/ultra-low ctDNA samples of one cancer patient, 8 samples are created for Series (A), Series (B), and reference sample. ctDNA: circulating tumor DNA, TF: Tumor Fraction.

Caller	Type	Model	AUPRC	Sensitivity (at 3% precision)
FreeBayes	naive	ODDS	0.001 ± 0.000	0.001 ± 0.001
	cfDNA-tuned	Decision Tree	0.030 ± 0.010	0.127 ± 0.024
		Random Forest	**0.114 ± 0.026**	**0.317 ± 0.022**
		Gradient Boosting	0.019 ± 0.007	0.263 ± 0.022
		Logistic Regression Lasso	0.053 ± 0.010	0.305 ± 0.028
		Logistic Regression Ridge	0.054 ± 0.010	0.306 ± 0.029
		Logistic Regression ElasticNet	0.046 ± 0.011	0.297 ± 0.020
		Multi-Layer Perceptron	0.110 ± 0.028	0.287 ± 0.017
Mutect2	naive	TLOD	0.081 ± 0.012	**0.492 ± 0.071**
	cfDNA-tuned	Decision Tree	0.070 ± 0.030	0.263 ± 0.063
		Random Forest	**0.278 ± 0.051**	0.490 ± 0.070
		Gradient Boosting	0.042 ± 0.011	0.353 ± 0.116
		Logistic Regression Lasso	0.229 ± 0.052	0.472 ± 0.054
		Logistic Regression Ridge	0.229 ± 0.052	0.471 ± 0.054
		Logistic Regression ElasticNet	0.220 ± 0.046	0.469 ± 0.051
		Multi-Layer Perceptron	0.273 ± 0.034	0.470 ± 0.054
Strelka2	naive	EVSomatic	0.212 ± 0.067	0.398 ± 0.083
	cfDNA-tuned	Decision Tree	0.080 ± 0.021	0.237 ± 0.060
		Random Forest	**0.262 ± 0.063**	0.472 ± 0.048
		Gradient Boosting	0.103 ± 0.038	0.315 ± 0.026
		Logistic Regression Lasso	0.177 ± 0.030	0.426 ± 0.070
		Logistic Regression Ridge	0.175 ± 0.028	0.415 ± 0.063
		Logistic Regression ElasticNet	0.190 ± 0.035	0.452 ± 0.071
		Multi-Layer Perceptron	0.239 ± 0.035	**0.476 ± 0.023**
VarDict	naive	SSF	0.055 ± 0.016	0.515 ± 0.091
	cfDNA-tuned	Decision Tree	0.013 ± 0.003	0.149 ± 0.028
		Random Forest	0.053 ± 0.008	0.473 ± 0.060
		Gradient Boosting	0.016 ± 0.010	0.155 ± 0.125
		Logistic Regression Lasso	0.010 ± 0.003	0.093 ± 0.047
		Logistic Regression Ridge	0.010 ± 0.003	0.093 ± 0.047
		Logistic Regression ElasticNet	0.010 ± 0.004	0.106 ± 0.055
		Multi-Layer Perceptron	**0.060 ± 0.012**	**0.521 ± 0.056**
VarScan	naive	SPV	0.043 ± 0.011	0.545 ± 0.057
	cfDNA-tuned	Decision Tree	0.014 ± 0.005	0.150 ± 0.034
		Random Forest	0.029 ± 0.009	0.282 ± 0.065
		Gradient Boosting	0.027 ± 0.004	0.512 ± 0.066
		Logistic Regression Lasso	0.039 ± 0.011	0.483 ± 0.055
		Logistic Regression Ridge	0.039 ± 0.011	0.486 ± 0.054
		Logistic Regression ElasticNet	0.039 ± 0.011	0.484 ± 0.055
		Multi-Layer Perceptron	**0.101 ± 0.031**	**0.554 ± 0.043**
SMuRF	naive	FALSE.	0.190 ± 0.033	0.290 ± 0.043
	cfDNA-tuned	Decision Tree	0.113 ± 0.029	0.219 ± 0.039
		Random Forest	0.251 ± 0.047	0.324 ± 0.047
		Gradient Boosting	0.056 ± 0.044	0.129 ± 0.070
		Logistic Regression Lasso	0.257 ± 0.041	0.357 ± 0.045
		Logistic Regression Ridge	0.257 ± 0.041	**0.357 ± 0.044**
		Logistic Regression ElasticNet	**0.262 ± 0.045**	0.356 ± 0.040
		Multi-Layer Perceptron	0.229 ± 0.038	0.323 ± 0.040

Supplementary Table 4: Alternative machine learning models for cell-free DNA-specific versions of tissue callers

The table displays the mean metric over three test CRC patients ± the standard error of the mean metric (s.e.m.). Permutations of the three CRC patients were used to create three different models. Models were trained on a 10% subset of two out of three CRC patients' benchmark dilution series at 150x and their matched pseudo ground truths. Subsequently, the models were tested on the full set of dilution series of the third unseen patient. The models' parameters are described in Methods Section 5.4.2. The best score for each category is displayed in bold. CRC: Colorectal cancer, cfDNA: cell-free DNA, AUPRC: Area Under Precision-Recall Curve.

Caller	Feature	Group/Comment	Full Description
FreeBayes	MQM	mapping quality	Mean mapping quality of observed reference alleles
	DPRA	VAF	Alternate allele depth ratio. Ratio between depth in samples with each called alternate allele and those without
	SRR	strand bias	Number of reference observations on the reverse strand
	ABP	VAF	Allele balance probability at heterozygous sites: Phred-scaled upper-bounds estimate of the probability of observing the deviation between ABR and ABA given E(ABR/ABA) ~0.5, derived using Hoeffding's inequality
	MQMR	mapping quality	Mean mapping quality of observed reference alleles
	SRF	strand bias	Number of reference observations on the forward strand
	ODDS	score	The log odds ratio of the best genotype combination to the second-best
	SAP	strand bias	Strand balance probability for the alternate allele: Phred-scaled upper-bounds estimate of the probability of observing the deviation between SAF and SAR given E(SAF/SAR) ~0.5, derived using Hoeffding's inequality
	SAR	strand bias	Number of alternate observations on the reverse strand
	SAF	strand bias	Number of alternate observations on the forward strand
	DPB	read count	Total read depth per bp at the locus; bases in reads overlapping / bases in haplotype
	RPP	mapping quality and base quality	Read Placement Probability: Phred-scaled upper-bounds estimate of the probability of observing the deviation between RPL and RPR given E(RPL/RPR) ~0.5, derived using Hoeffding's inequality
	AB	VAF	Allele balance at heterozygous sites: a number between 0 and 1 representing the ratio of reads showing the reference allele to all reads, considering only reads from individuals called as heterozygous
	PAIRED	paired-end	Proportion of observed alternate alleles which are supported by properly paired read fragments
	RPR	mapping quality and base quality	Reads Placed Right: number of reads supporting the alternate balanced to the right (3') of the alternate allele
	RPL	mapping quality and base quality	Reads Placed Left: number of reads supporting the alternate balanced to the left (5') of the alternate allele
	PAIREDR	paired-end	Proportion of observed reference alleles which are supported by properly paired read fragments
	EPPR	mapping quality and base quality	End Placement Probability for reference observations: Phred-scaled upper-bounds estimate of the probability of observing the deviation between EL and ER given E(EL/ER) ~0.5, derived using Hoeffding's inequality
	AC	read count	Total number of alternate alleles in called genotypes
	SRP	strand bias	Strand balance probability for the reference allele: Phred-scaled upper-bounds estimate of the probability of observing the deviation between SRF and SRR given E(SRF/SRR) ~0.5, derived using Hoeffding's inequality
	RPPR	mapping quality and base quality	Read Placement Probability for reference observations: Phred-scaled upper-bounds estimate of the probability of observing the deviation between RPL and RPR given E(RPL/RPR) ~0.5, derived using Hoeffding's inequality
	EPP	mapping quality and base quality	End Placement Probability for reference observations: Phred-scaled upper-bounds estimate of the probability of observing the deviation between EL and ER given E(EL/ER) ~0.5, derived using Hoeffding's inequality
Mutect2	TLOD	score	Variant existing versus not existing (log likelihood ratio)
	NALOD	matched normal: potential CHIP	Artifact in normal with same VAF as tumor (negative log odds)
	NLOD	matched normal: germline variant	Normal diploid het or hom alt genotypes (log likelihood ratio)
	FS	strand bias	Strand bias (Fisher exact p-value)
	MQRankSum	mapping quality	Alt vs Ref read mapping qualities (Wilcoxon rank sum z-score)
	RPRS = ReadPosRankSum	read position	Alt vs. Ref read position bias (Wilcoxon rank sum z-score)
	ECNT	read count	Number of events in this haplotype
Strelka2	SomaticEVS	score	Somatic Empirical Variant Score (EVS) expressing the phred-scaled probability of the call being a false positive observation
	RPRS = ReadPosRankSum	read position	Z-score from Wilcoxon rank sum test of Alt Vs. Ref read-position in the tumor
	QSS_NT	quality score	Quality score reflecting the joint probability of a somatic variant and NT
	QSS	quality score	Quality score for any somatic SNV, i.e., for the ALT allele to be present at a significantly different frequency in the tumor and normal
	MQ	mapping quality	RMS Mapping Quality
	MQ0	mapping quality	Total Mapping Quality Zero Reads
	SNVSB	strand bias	Somatic SNV site strand bias
VarDict	SSF	score	P-value
	SOR	strand bias	Strand odds ratio
	MSI	microsatellite instability	MicroSatellite. >1 indicates MSI
	SHIFT3	INDEL property	No. of bases to be shifted to 3 prime for deletions due to alternative alignment
	MSILEN	microsatellite instability	MSI unit repeat length in bp
VarScan	SPV	score	Fisher's Exact Test P-value of tumor versus normal for Somatic/LOH calls
	SS	mainly germline or somatic	Somatic status of variant (0=Reference,1=Germline,2=Somatic,3=LOH, or 5=Unknown)
	SSC	quality score	Somatic score in Phred scale (0-255) derived from somatic p-value
	GPV	germline	Fisher's Exact Test P-value of tumor+normal versus no variant for Germline calls
SMuRF	s2_QSS	quality score	Strelka2 QSS
	m2_TLOD	Mutect2 score	Mutect2 TLOD
	s2_SomaticEVS	Strelka2 score	Strelka2 SomaticEVS
	vd_SSF	VarDict score	VarDict SSF
	FALSE.	score	score
	m2_MQ	mapping quality	Mutect2 MQ
	m2_NLOD	matched normal	Mutect2 NLOD
	s2_RPRS	read position	Strelka2 ReadPosRankSum
	v2_SSC	quality score	VarScan SSC
	m2_RPRS	read position	Mutect2 ReadPosRankSum
	v2_SPV	VarScan score	VarScan SPV
	f_MQMR	mapping quality	FreeBayes MQMR
	FILTER_Mutect2	Mutect2 binary prediction with default threshold	Mutect2 FILTER
	s2_MQ	mapping quality	Strelka2 MQ
	FILTER_Strelka2	Strelka2 binary prediction with default threshold	Strelka2 FILTER
	vd_MSI	microsatellite instability	VarDict MSI
	f_ODDS	score	FreeBayes ODDS
	f_MQM	mapping quality	FreeBayes MQM
	vd_SOR	strand bias	VarDict SOR
	m2_MQRankSum	mapping quality	Mutect2 MQRankSum

Supplementary Table 5: Description of most important features in cell-free DNA-tuned versions of tissue-specific callers