

2024 SMAHT SNV/indel Detection Challenge

A Summary of Results

(prepared by DAC; last update on 8/12/24)

Table of contents

Data Release - Version 1.0.....	2
COLO829 truth set.....	4
Overview of the truth set generation.....	4
Distribution of the truth set variants across VAF ranges.....	4
Defining the negative control for the COLO829-T truth set.....	5
SNV/Indel Detection Challenge: Pipelines Used by GCCs/TTDs.....	6
Comparisons of mosaic SNV detection across <i>in silico</i> COLO829-BLT50 samples.....	8
Comparisons of mosaic SNV detection across cell-admixture COLO829-BLT50 samples.....	9
Genome stratification for SNV benchmarking.....	9
Expanding the truth set to include “difficult” genomic regions.....	9
Overview of the truth set across the stratified genomic regions.....	10
Comparisons of mosaic SNV detection across genomic regions.....	11
Broad-GCC: 180X (COLO829-BLT50).....	12
BCM-GCC: 510X (COLO829-BLT50).....	13
Supplementary Information.....	14
Genome Region Stratifications for SMAHT benchmarking truth call sets of SNV/Indels.....	14
Pipeline evaluation for the SNV/Indel Detection Challenge - Additional Plots.....	15

Data Release - Version 1.0

Version 1.0 (June 26, 2024) is an **unofficial internal release**.

*** The V1.0 data are to be shared within the **SMaHT Network only**. ***

The indel truth set will be made available in the near future.

The official public release of the final truth sets may be made after the consortium-wide approval.

What is released in V1.0:

1. Truth set for SNVs in COLO829-BLT50.
2. Negative controls for SNVs: (i) germline variant sites based on COLO829-BL and (ii) homozygous reference / non-variant sites.
3. Genome stratification: (i) easy, (ii) difficult, and (iii) extreme regions.
4. [For the final call sets OK to be published] Submitted VCFs from GCCs and accuracy metrics (precision, sensitivity, F1 score) from the SNV/Indel Challenge.

The original calls submitted by centers are normalized (i.e., split multiallelic variants and left align indels), and, where applicable, files have been re-formatted in the standard VCF format for proper comparisons. The README files describing the pipelines and additional notes provided by submitters are included in the same directory as the VCFs.

All data are immediately available via Globus, a secure data transfer tool that can move large amounts of data between the source and destination storages directly. The data are organized into subfolders. The top folder, denoted as the **SMaHT-DAC SNV/Indel Challenge Release V1** collection on Globus, is the endpoint to which you will be added to access all subfolders. To access the data via Globus, please contact DAC [smhelp@hms-dbmi.atlassian.net] with your Globus ID.

Directory structure at DAC's Globus endpoint

```
└── snvIndelChallenge_Release_V1
    ├── callsets_from_centers
    │   ├── GCC_Broad
    │   │   └── accuracy_scores
    │   │       └── submitted_vcf_normalized
    │   └── GCC_NYGC
    │       └── accuracy_scores
    │           └── submitted_vcf_normalized
    ├── genomeTiers
    ├── negControl_snv_COLO829
    │   ├── germline
    │   └── homozygous_ref
    ├── truthset_snv_COLO829
    └── SMaHT_Benchmark_SNVIndel_Challenge-DataRelease_V1.pdf (this document)
```

Under the accuracy_scores directory for each center are accuracy matrices that contain precision, sensitivity, and F1 scores across VAF bins for *in silico* mixture and cell-admixture BLT50 samples.

The file name prefix of an accuracy matrix is the following:

[SUBMITTER]_[SAMPLE DESCRIPTION]_[DATA GENERATOR]

E.g. File names of the accuracy matrices of the Broad's pipelines applied on 100X *in silico* mixture BLT50 and the cell-admixture BLT50 from BCM:

BROAD_insilico-BLT50-100X.accuracy.tsv
 BROAD_BLT50_BCM.accuracy.tsv

Other Files in Version 1 Release:

Table 1. The SNV truth set used for the COLO829 SNV/Indel Detection Challenge.

Data description	File name	File size
COLO829 SNV truth set - Version 1.0	SMaHT_COLO829_SNV_truth_set_v1.0.vcf	3.3 MB

Table 2. The negative control sets for SNVs in COLO829-BLT50.

Data description	File name	Total file size
Germline variant sites	colo829bl_germline.vcf.gz	280 MB
Homozygous reference sites	(Grouped by chromosomes) /byChr/chr*.merged.bed (Grouped by genomic tiers and chromosomes) /byGenomeTiers/[easy difficult extreme]/chr*.merged.bed	2.2 GB

Table 3. Genome stratification used for the COLO829 SNV/Indel Detection Challenge.

Category	Tier(s)	File name	File size (MB)
Easy	Tier 0	SMaHT_easy_regions_GRCh38_v1.0.bed	111
Difficult	Tiers 1-4 merged	SMaHT_difficult_regions_all_GRCh38_v1.0.bed	116
	Tier 1	SMaHT_difficult_regions_tier1_GRCh38_v1.0.bed	141
	Tier 2	SMaHT_difficult_regions_tier2_GRCh38_v1.0.bed	76
	Tier 3	SMaHT_difficult_regions_tier3_GRCh38_v1.0.bed	43
	Tier 4	SMaHT_difficult_regions_tier4_GRCh38_v1.0.bed	23
Extreme	Tier 5	SMaHT_extreme_regions_GRCh38_v1.0.bed	5.7

COLO829 truth set

The truth set in the Version 1 release includes SNVs with VAF $\geq 25\%$ in the pure COLO829-T cancer cell line (corresponding to VAF $\geq 0.5\%$, expected in the BLT50 mixture samples) to analyze variants considered for the SNV/Indel Detection Challenge.

Overview of the truth set generation

To achieve high sensitivity, the truth set of the COLO829-T cancer cell line was derived by taking the union of somatic variants identified by four callers (Mutect2, Strelka2, VarNet, and RUFUS) using the Illumina WGS data. To achieve high precision, the truth set was validated by PacBio HiFi reads using `bcftools mpileup` with BQ and MQ ≥ 30 (**Figure 1**). We considered SNVs in three different genomic regions where Illumina short-reads are mapped with different levels of confidence.

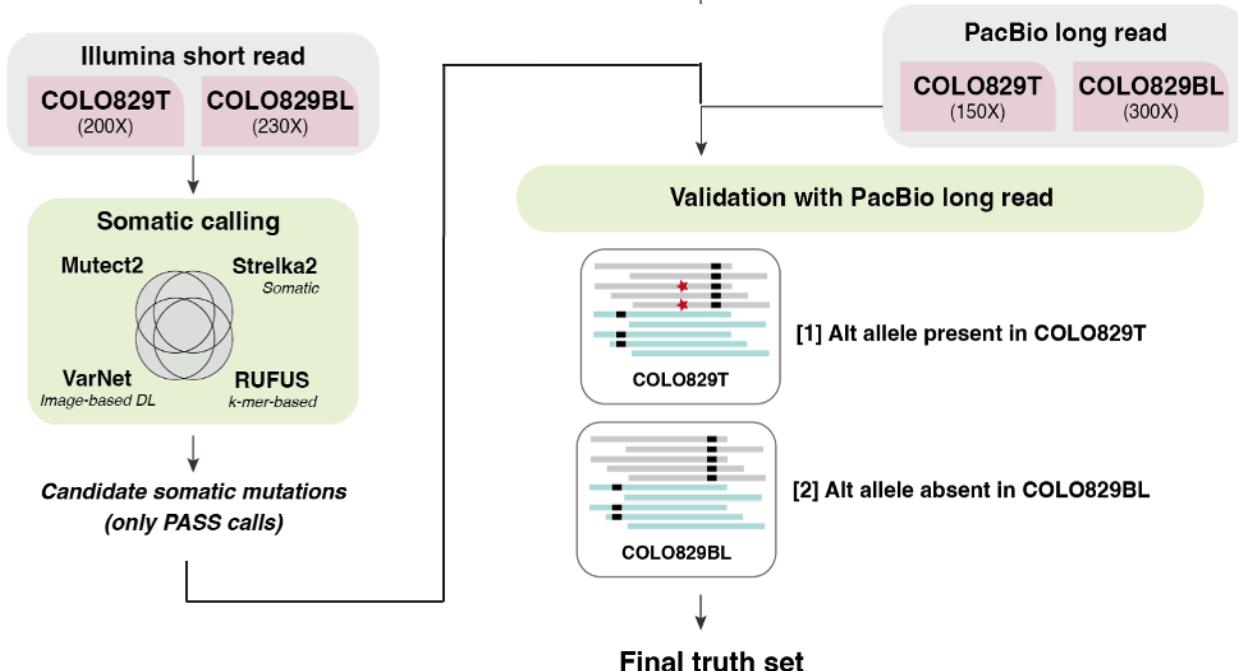


Figure 1. Overview of the COLO829-T truth set generation method.

Distribution of the truth set variants across VAF ranges

The true variants in COLO829-BLT50 are expected to have VAF under 2% based on the mixing ratio of 49-parts COLO829-BL to 1-part COLO829-T. As expected, the distribution of VAF of the truth set variants shows the majority under 2%, as shown in **Figure 2**.

Distribution of truth set across VAF bins

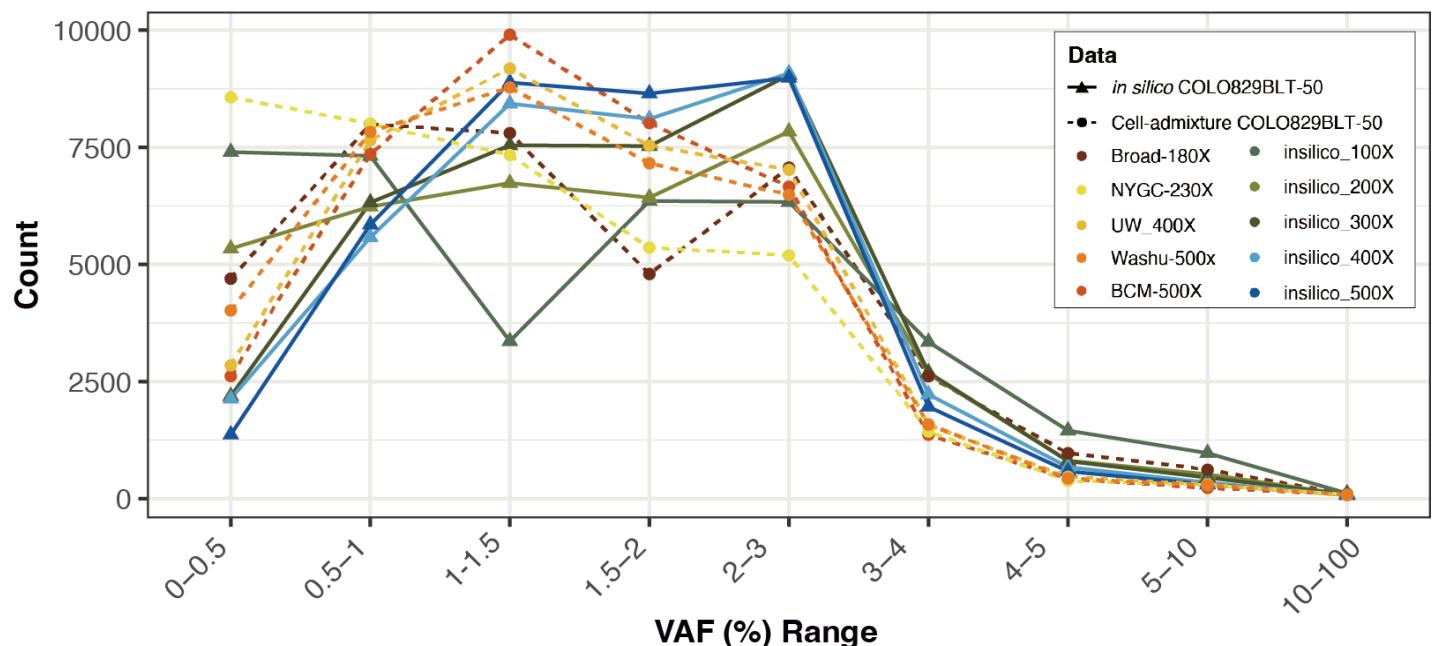


Figure 2. VAF distribution of SNVs in the truth set across different COLO829-BLT50 samples.

Defining the negative controls for the COLO829-T truth set

Sometimes an algorithm makes a call using Illumina data, which is not cross-validated by an orthogonal dataset (e.g., PacBio). We cannot be sure that such a call is false-positive, e.g., due to PacBio data having insufficient coverage at this position. Thus, we derived a “negative control” set consisting of positions that are (almost certainly) not true mosaic variant sites. If the negative control set is not defined, true mosaic variants can be miscategorized as false positive calls (e.g. true variants among the non-PASS Mutect2 calls), potentially leading to the underestimation of precision.

We determined the negative controls as the following:

- Non-variant, homozygous reference positions
 - Two types of candidate sites were identified across the genome, based on raw pileup counts compared between the number of supporting reads for the alt allele in COLO829-T and -BL:
 - (1) Homozygous reference sites
 - Positions that were annotated as 0/0 in GT (genotype), not 1/0 nor 1/1, by Sentieon Haplotype (which implements GATK HaplotypeCaller) using the GVCF mode.
 - (2) Positions that failed during the COLO829-T truth set generation, consisting of:
 - a. Variants that were filtered out during the short-read variant calling process (e.g., those annotated as ‘weak_evidence’ in the FILTER field instead of ‘PASS’)
 - b. Variants that were not validated by PacBio reads because there was no supporting alt allele in COLO829-T.

- Germline variant positions
 - COLO829-BL germline variants were identified by Sentieon Haplotypeper using PacBio data.

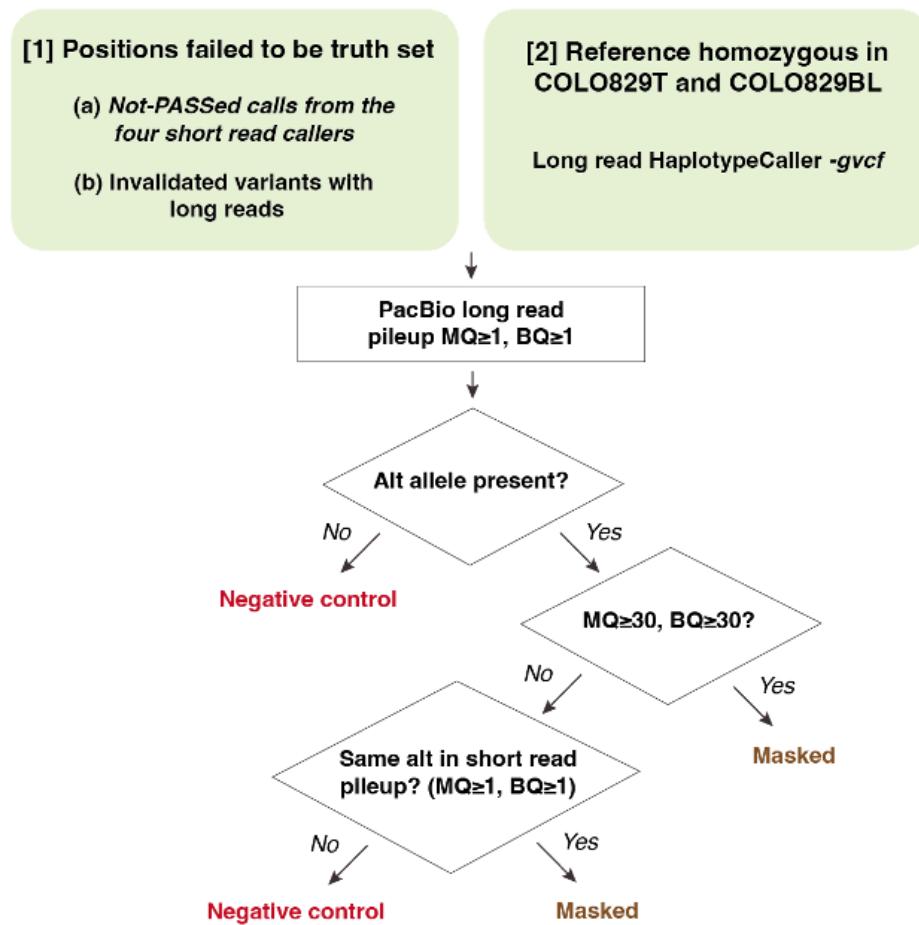


Figure 3. Overview of the negative control design and generation.

Pipelines Used by GCCs/TTDs

Table 4. Pipelines and filtering methods used by GCCs, TTD, and DAC. Column names in the accuracy matrix contain abbreviations of the pipelines and filtering methods, as shown in blue.

Center	Main algorithm	Filtering method & Note
BCM-GCC	DRAGEN (v4.3)	VAF 10%; variant calling with mosaic detection on
Broad-GCC	BROAD_MUTECT2: Mutect2 using the tumor-only mode	Filtered against the publicly available panel of normal and gnomAD; Ran with the GATK best practice. Post-calling filter: Exact binomial test to remove germline calls; Removed calls with Depth <50X or >1,000X; Removed calls in pseudoautosomal regions of chr X and Y; FilterMutectCalls
	BROAD_DRAGEN: DRAGEN (v4.3 - beta)	Used parameters to avoid the RG issue ("--use-single-read-group-for-bam-list) and VAF < 20%

		(“--vc-mosaic-af-filter-threshold = 0.2”) Same post-calling filter as above; Variants with the “PASS” flag
	BROAD_DRAGEN_Direct: DRAGEN (v4.3 - beta)	The same DRAGEN pipeline described above was run directly on their original CRAM file.
NYGC-GCC	NYGC_Lancet: Lancet2 - single sample mode	Filtered against germline calls using HaplotypeCaller + GenotypeGVCFs workflow; removed calls in gnomAD (>0.00001% AF) and 1KGenome (high-coverage panel), and in UCSC simpleRepeats and segDup regions; Depth = 150 - 250, whole coverage = 200 - 300, allele depth > 4, VAF < 20%, strand, allele quality and mapping quality considered; Removed indels of 1 bp.
WashU-GCC	WASHU_Vaf: Mutect2 (GATK 4.4.0.0) - VAF	FilterMutectCalls; VAF < 2 %
	WASHU_Mutect2dbSNP: Mutect2 (GATK 4.4.0.0) - dbSNP	FilterMutectCalls; Remove calls in dbSNP v151
Utah-TTD	Utah_Lenient: RUFUS - Lenient	3 filter levels to reduce the noise levels in the call set: (1) Lenient; (2) moderate, and (3) strict. (* RUFUS only works with a normal control sample - 230X COLO829-BL was used)
	Utah_Moderate: RUFUS - Moderate	
	Utah_Strict: RUFUS - Strict	
UW-GCC	UW_MUTECT2: Mutect2 using the tumor-only mode	FilterMutectCalls; VAF > 5% with Depth > 20
	UW_DRAGEN: DRAGEN Somatic (v3.9.5)	Variants with the “PASS” flag: VAF > 5%
DAC	DAC_MF: MosaicForecast	Removed calls in UCSC SimpleRepeats, SegmentalDuplication, MosaicForecast clustered regions.

Comparisons of mosaic SNV methods across *in silico* COLO829-BLT50 samples

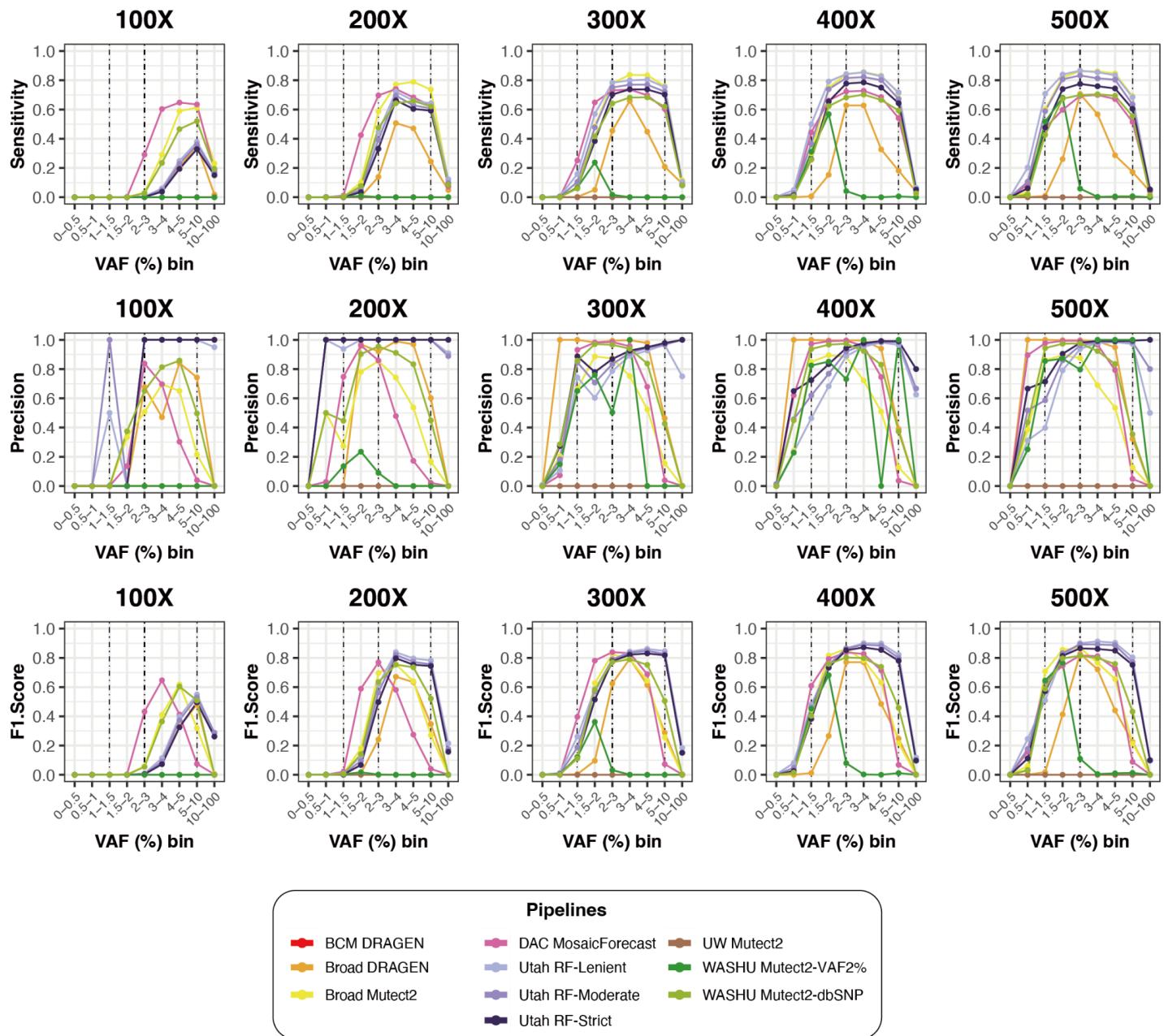


Figure 4. Pipeline performance comparisons across five *in silico* COLO829-BLT50 samples.

Comparisons of mosaic SNV methods across cell-admixture COLO829-BLT50 samples

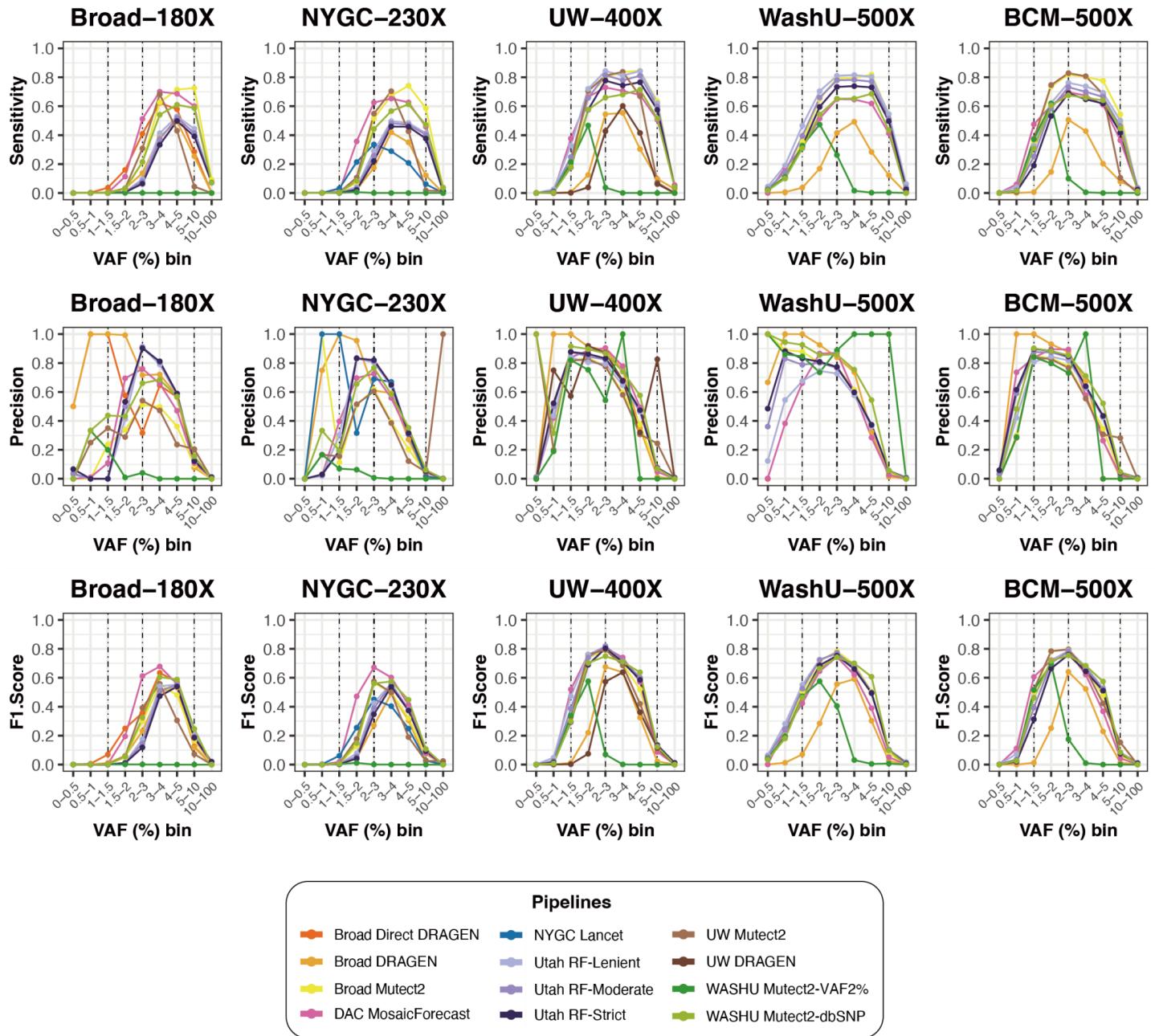


Figure 5. Pipeline performance comparisons across five cell-admixture COLO829-BLT50 samples.

Genome stratification for SNV benchmarking

Expanding the truth set to include “difficult” genomic regions

Mosaic variant detection in genomic regions with low mappability can be challenging. If the variants in the truth set were considered only in the regions with high mappability (i.e. “high-confidence” regions), it becomes an “easy” benchmark study with inflated performance measures.

Thus, we took the GIAB (“Genome in a Bottle”) easy regions as the baseline (in the “Easy” category) and expanded the genome categorization based on k-mer-based read mappability scores using the UMAP software (Karimzadeh M *et al.* (2018) NAR).

Table 5. Genome stratification was applied to categorize the variants in the truth set. The “Easy” regions were determined based on the GIAB regional stratification, which excludes the GIAB difficult region defined as (1) Low Complexity; (2) Functional Technically Difficult; (3) Genome Specific (4) Functional Regions; (5) GC content; (6) mappability; (7) Other Difficult; (8) Segmental Duplications; (9) Union; (10) Ancestry, or (11) XY.

Category	Tier	Description	Bp	% of genome	
Easy	Tier 0	GIAB easy	2,310,050,737	74.80	
Difficult	Tier 1	Additional UMAP24 regions	371,562,536	16.68	12.03
	Tier 2	Additional UMAP36 regions	53,704,601		1.74
	Tier 3	Additional UMAP50 regions	35,085,249		1.14
	Tier 4	Additional UMAP100 regions	54,908,647		1.78
Extreme	Tier 5	Rest of the hg38 genome	262,958,062	8.51	
Total		Entire hg38 genome	3,088,269,832	100	

Overview of the truth set across the stratified genomic regions

The truth set contains 44K high-quality SNVs (plus 644 putative indels; to be distributed in the next release after more refined validation) across the genomic regions described above.

We also categorized the negative control sets into the genomic regions as follows:

- Negative controls for non-variant, homozygous reference positions
 - Total number of sites = 2,772,580,858 (90% of the entire genome)
 - Easy region: 2,229,413,903 (72% of the genome)
 - Difficult region: 475,890,667 (15% of the genome)
 - Extreme region: 67,276,288 (2% of the genome)
- Negative controls for germline variants (COLO829-BL)
 - Total number of sites = 6,208,223
 - Easy: 3,205,125 (52% of the germline variants)
 - Difficult: 2,803,519 (45% of the germline variants)
 - Extreme: 203,189 (3% of the germline variants)

Table 6. True SNV and indel calls across the easy, difficult, and extreme categories of the genomic region. The “All” variant set includes all somatic calls; The “Challenge” set includes only the variants with VAF \geq 25% in COLO829-T, to consider the variants with VAF \geq 0.5% expected in the BLT50 mixture samples for the Challenge.

Category	Tier	SNVs (All)		SNVs (Challenge)		Indels (All)		Indels (Challenge)	
Easy	Tier 0	35,128		29,155		252		166	
Difficult	Tier 1	8,989	6465	7149	5,107	389	278	230	176
	Tier 2		999		784		78		39
	Tier 3		620		505		22		8
	Tier 4		905		753		11		7
Extreme	Tier 5	433		345		3		2	
Total		44,550		36,649		644		398	

Comparisons of mosaic SNV detection across genomic regions

We evaluated the performance of the variant-calling pipelines across different genomic regions. As representative datasets, below are the accuracy of variant calls made from the cell-admixture BLT50 samples with the lowest (180X from Broad-GCC) and the highest (510X from BCM-GCC) sequencing depths, shown in **Figures 6** and **7**, respectively.

Broad-GCC: 180X (COLO829-BLT50)

Broad COLO829BLT-50 (180X)

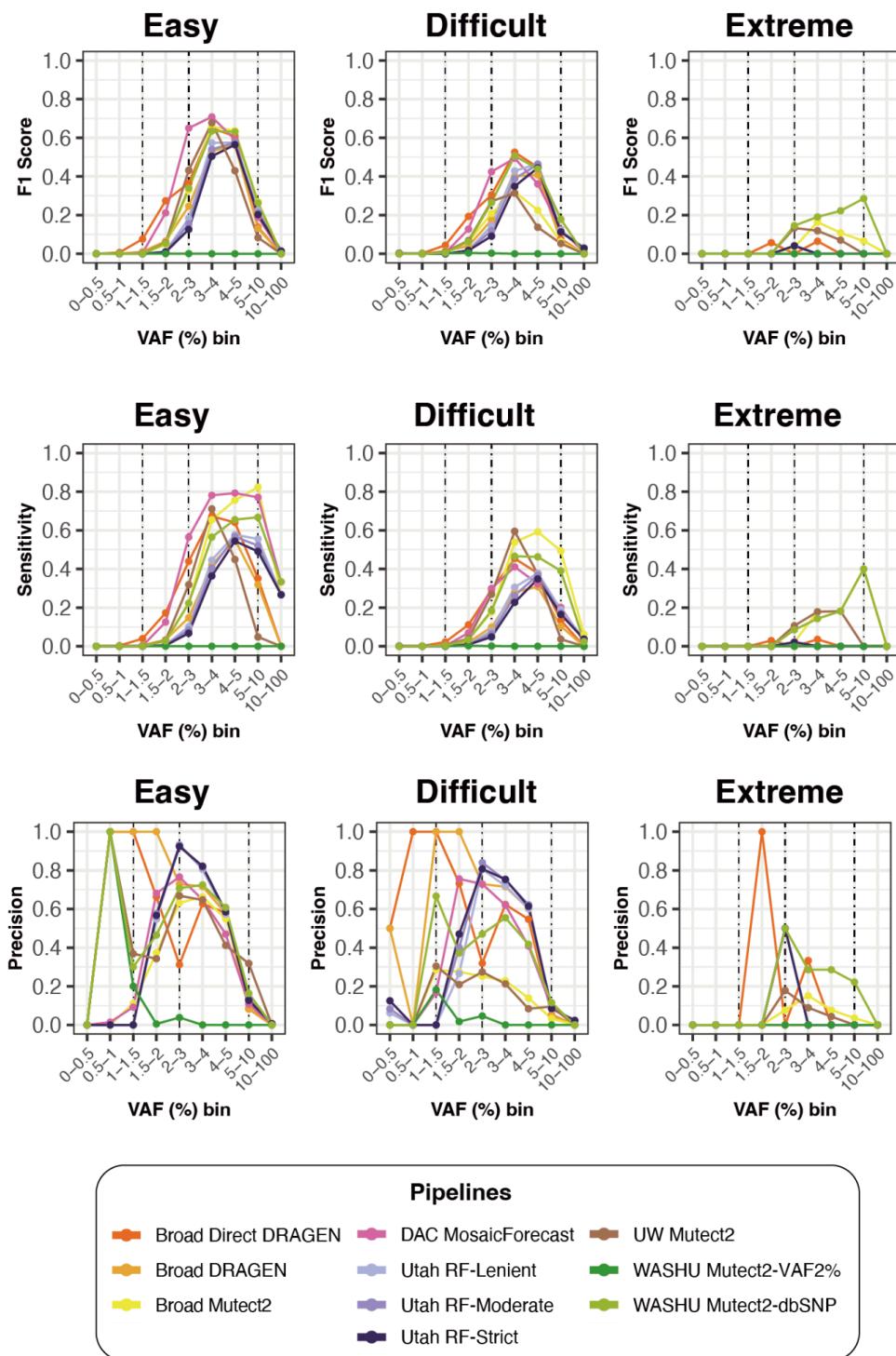


Figure 6. Performance comparisons across different genomic regions in 180X cell-admixture COLO829-BLT50.

BCM-GCC: 510X (COLO829-BLT50)

BCM COLO829BLT-50 (500X)

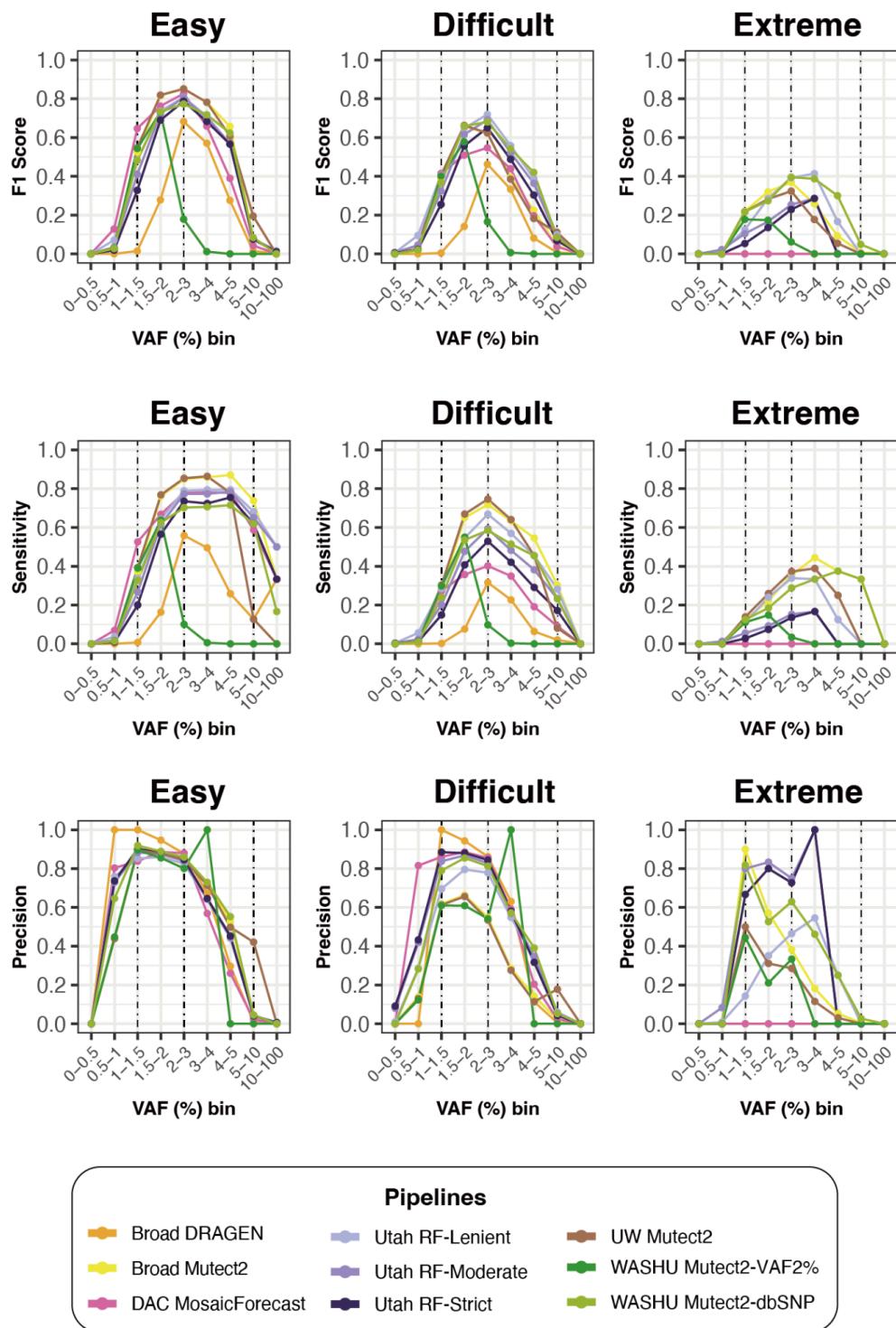


Figure 7. Performance comparisons across different genomic regions in 500X cell-admixture COLO829-BLT50.

Supplementary Information

Genome Region Stratifications for SMaHT benchmarking truth call sets of SNV/Indels

Version: 1.0

Last updated: July 1, 2024

The Data Analysis Center (DAC) of the NIH Somatic Mosaicism across Human Tissues (SMaHT) consortium generates BED files containing genomic regions with different levels of confidence for mosaic variant detection, stratified into three groups, i.e. “Easy”, “Difficult”, and “Extreme”. The “Difficult” region is subclassified into four tiers based on the Umap mappability score. The regions are based on the GRCh38 reference human genome.

Please refer to **Table 3** for the full list of BED files for all genomic regions released in Version 1.

Easy regions

“Easy” regions are high-confidence regions that exclude the difficult regions defined by Genome in a Bottle (GIAB v3.3) in autosomes and sex chromosomes, i.e. Chr1 - 22, X, Y.

We used the BED file for the GIAB v3.3 difficult regions, GRCh38_alldifficultregions.bed.gz under

<https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/genome-stratifications/v3.3/GRCh38@all/Union/>

Difficult regions

Difficult regions are tiered based on the read mappability scores determined by Umap (Karimzadeh M et al. (2018) NAR; <https://doi.org/10.1093/nar/gky677>), which identifies genomic regions with uniquely mappable k-mers. A subset of the GIAB difficult regions is included and tiered based on the mappability of reads of different lengths:

- Tier 1 [*most stringent, good mappability*]: Regions with uniquely mappable reads of 24 bp in length
- Tier 2: Regions with uniquely mappable reads of 36 bp in length
- Tier 3: Regions with uniquely mappable reads of 50 bp in length
- Tier 4 [*most lenient*]: Regions with uniquely mappable reads of 100 bp in length

*Note: The tiers of the “Difficult” regions do not overlap.

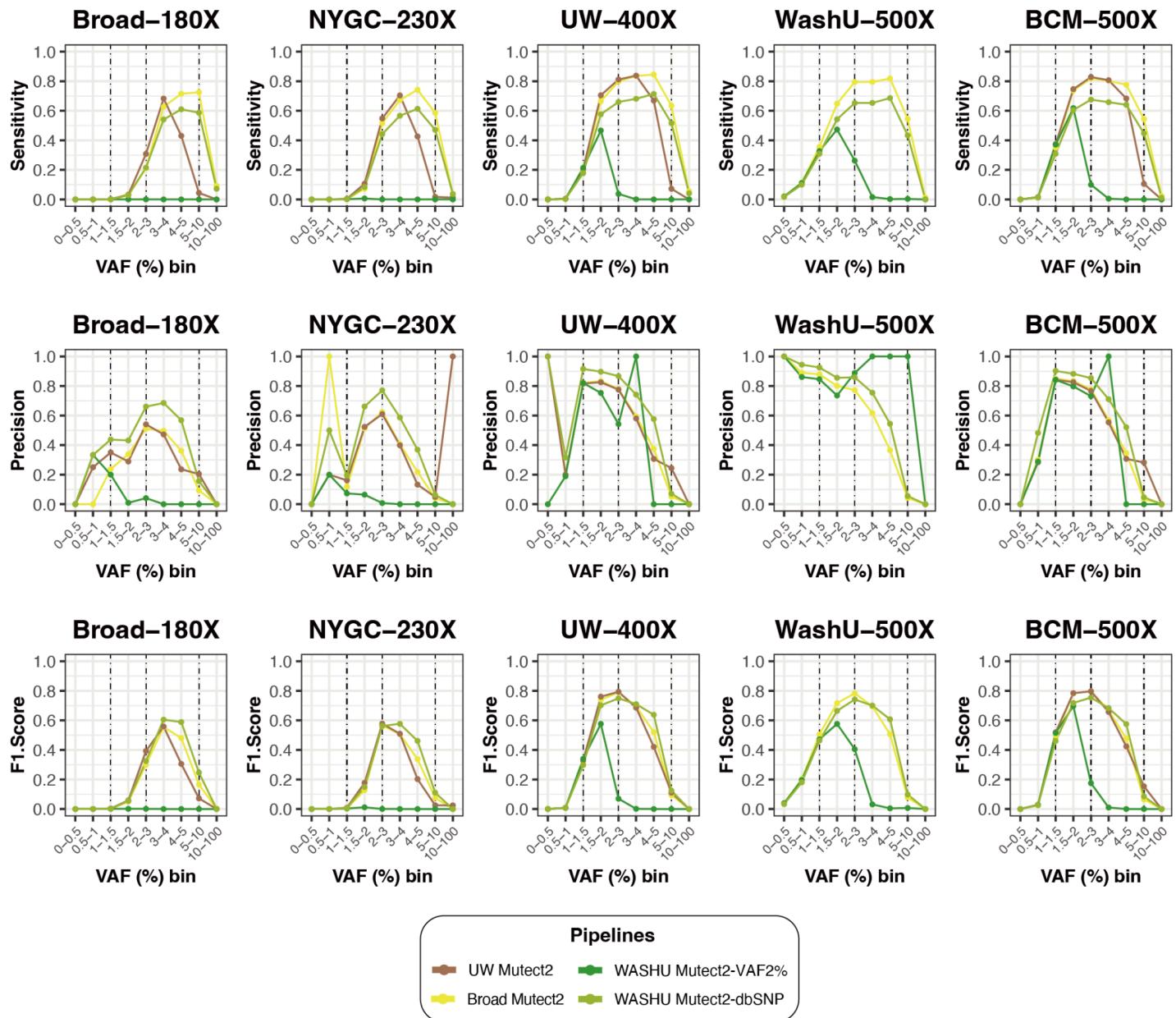
Extreme regions

Genomic regions that are not in the “Easy” and “Difficult” regions are defined as “Extreme” where mosaic variant detection is most challenging due to poor mappability.

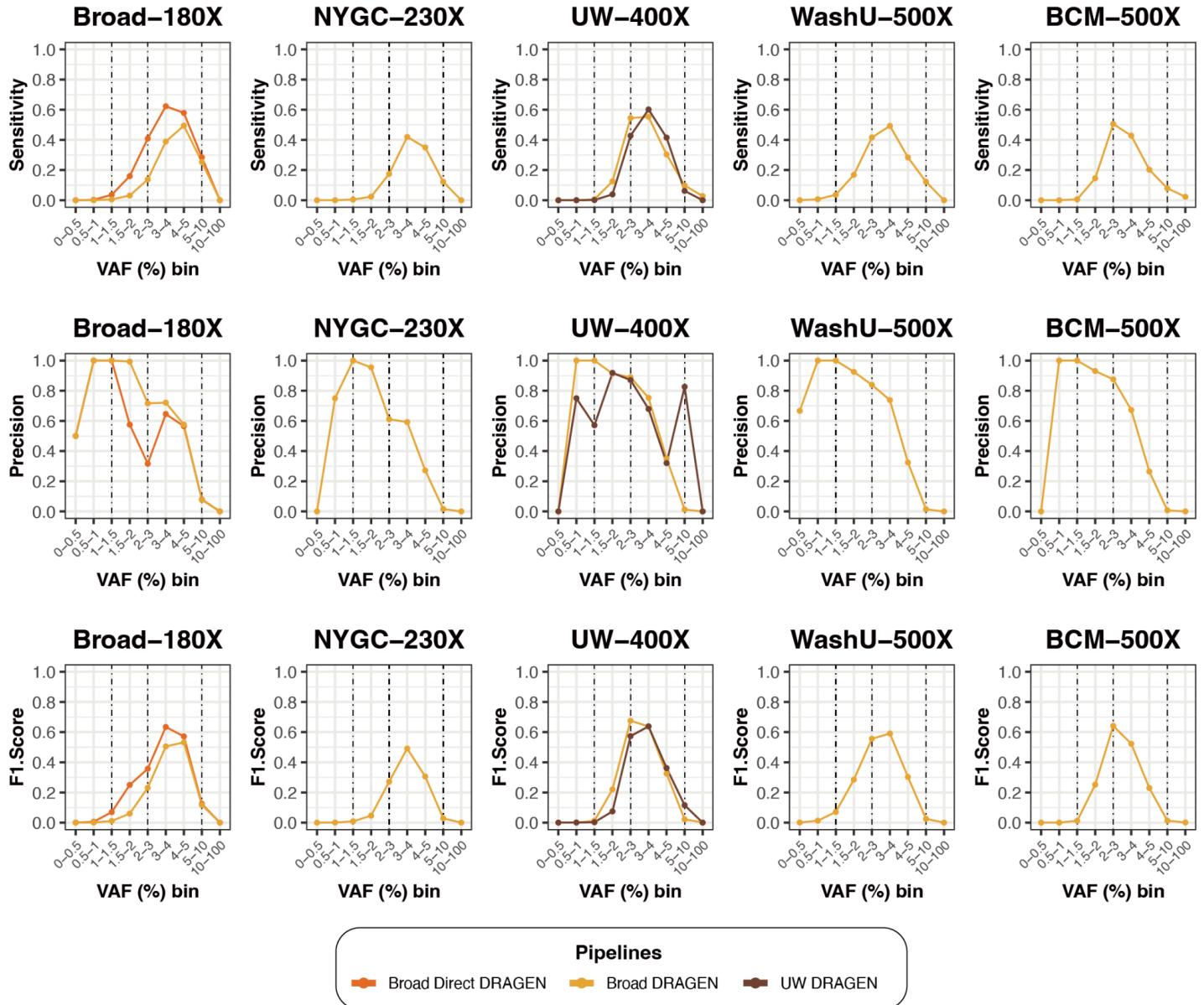
Pipeline evaluation for the SNV/Indel Detection Challenge - Additional Plots

The following plots show the performance of the pipelines grouped by common variant callers used by the centers.

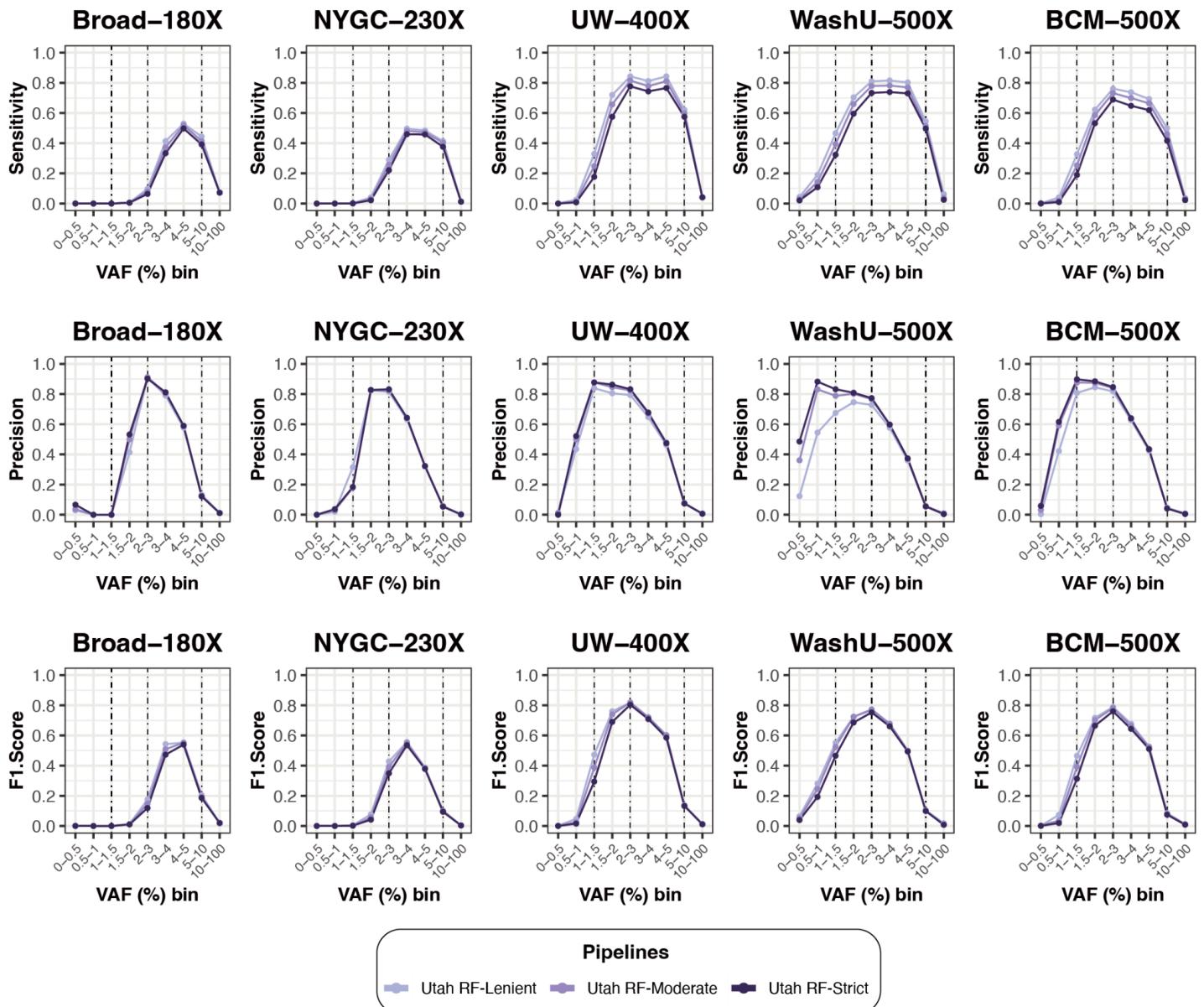
Mutect2



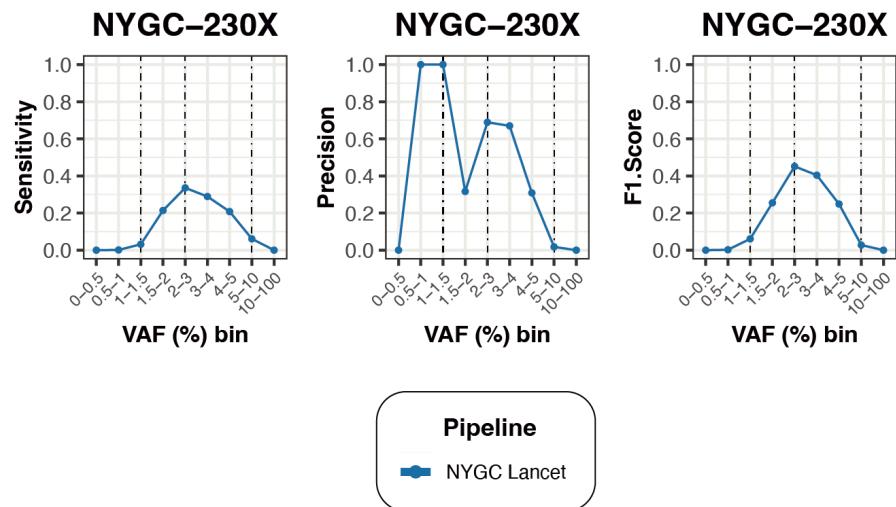
DRAGEN



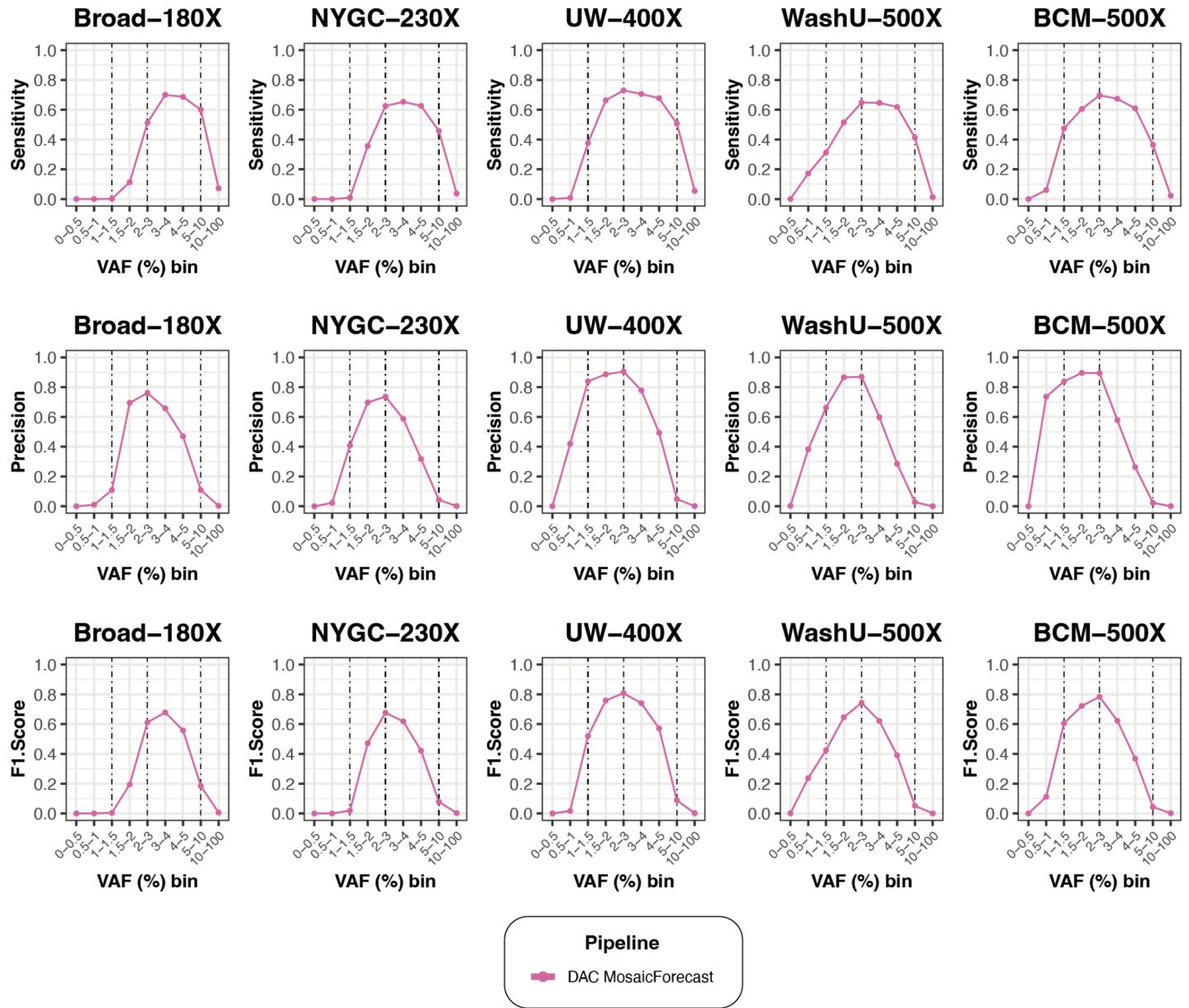
RUFUS



Lancet



MosaicForecast



Supplementary Figure 1. Pipeline performance comparisons across five cell-admixture COLO829-BLT50 samples. The plots are grouped by the pipelines that use the same variant callers by the centers.