B/tMSI-CAST: a systematic approach for microsatellite instability detection in both tumor tissue and plasma specimens from next-generation sequencing data

Fengchang Huang[1,#], Lili Zhao[2,#], Hongyu Xie[2,#], Jian Huang[1], Xiaoqing Wang[2], Jun Yang[1], Yuanyuan Hong[2], Jingchao Shu[2], Jianing Yu[2], Qingyun Li[2], Hongbin Zhang[1], Weizhi Chen[2], Ji He[2], Wenliang Li[1,*]

[1] Department of Oncology, the First Affiliated Hospital of Kunming Medical University, 650032 Kunming, Yunnan, China

[2] GeneCast Biotechnology Research Institute, 100088 Beijing, China

E-mail: liwenliang@kmmu.edu.cn (Li W)

Running title: *Huang F et al / b/tMSI-CAST*

Abstract

Microsatellite instability (MSI) is a phenotypic manifestation of mismatch repair (MMR) deficiency and has been used as a biomarker through NGS-based assays to aid clinical cancer diagnosis and prognosis, as being advantageous over traditional electrophoresis-based tests and immunohistochemistry. Critical to detection accuracy but seldomly systematically discussed and poorly resolved by currently available methods are topics such as influences of changing experimental conditions, MSI-specific duplication removal and tradeoffs associated with loci selection strategies. In seeking solution to these questions, we have developed b/tMSI-CAST (blood/tissue MSI Caller Adjusted with Sequence duplicaTes), an integrated method consisting of a set of loci selection principles, an MSI-specific duplication removal strategy and a two-mode calling algorithm targeting both blood and tissue samples, equipped with a baseline construction method based on duplication distributions. We benchmarked b/tMSI-CAST against two established tools, mSINGS and MSIsensor-pro and found b/tMSI-CAST showed matched performance as MSIsensor-pro and outperformed mSINGS. We further applied it to a retrospective cohort of 18084 clinical samples to present a comprehensive landscape view on MSI prevalence across 26 cancer types, the largest of its kind in Chinese population so far, and demonstrated it on presurgical cfDNA samples of patients with early stage cancers.

KEYWORDS: MSI; cfDNA; deduplication removal; loci selection; MSI baseline

Introduction

Microsatellite instability (MSI) is a phenotype of mismatch repair (MMR) deficiency and an emerging guideline recommended biomarker in the diagnosis of Lynch syndrome (1) and better prognosis of patients with MSI high metastatic colorectal cancer (CRC) (2) in addition to other solid tumors (3). Compared to electrophoresis-based tests that target several microsatellites (4) or immunohistochemistry (IHC) , NGS-based methods are gaining popularity in clinical settings owing to more flexible choices of panel, comparable performance without matched normal, improved accuracy and compatibility with panel designed for other purposes.

In the context of NGS-based MSI algorithms that are based on comparing the repeat lengths of alleles between a test sample and a reference (5), the majority of MSI tools were intended for tissue samples (6), with few for detection of MSI events from circulating tumor DNA (ctDNA) (7, 8). The later poses greater challenges in discriminating significantly lower level of signals from background noises, similar to somatic single nucleotide variant (SNV) calling (9-11). Workarounds include deeper sequencing, use of unique molecular identifiers and a statistically described confidence in how likely observed alleles are authentic somatic events.

Including more loci in a panel without selection would not necessarily increase the sample-level sensitivity and specificity as non-specific loci decrease the sensitive loci ratio. Microsatellite locus selection started since the development of MSI-PCR methods (4, 12) and are widely applied in NGS-based assays using either MSI-targeting panels (13, 14) or panels designed with other primary purposes (15, 16). Mononucleotide repeats are generally preferred over dinucleotides or higher repeat unit lengths due to higher probability of somatic mutation given MMR deficiency and the number of repeats fewer than 7 are poor indicators (17-20). Regarding the type of repeat unit nucleotides, there have been reported that A/T repeats are less mutable than C/G (21). Given possible dependence of polymorphism on ethnic lines (22), screening on microsatellites by monomorphism and controlling for ethnicity shall help improve sample-level specificity. Capture efficiency and ranked importance based on classification accuracy were also used (8).

Patient-matched normal samples are ideal means to identify population monomorphism, and yet tumor-only methods could get through by using mononucleotide microsatellites that are mostly monomorphic (4, 23, 24) and mitigating the impact of locus-level false positives caused by sporadic heterozygous loci through sample-level joint calling using a combination of

microsatellite loci. However, existing tumor-only methods do not screen microsatellite loci by monomorphism (6, 25). Under paired mode, matched normal samples serve as the background noise against which tumor signals are compared (26, 27), whereas a baseline built using a group of MSI-negative samples can be alternatively used for the tumor-only mode.

With regard to the choice of duplication removing tool, published NGS-based MSI detection methods either explicitly stated choosing Picard (7, 25) , assumed post-deduped BAMs as input (6, 27) or did not mention their choices (**Table 2**). Picard MarkDuplicates chooses the family member with the greatest sum of sequenced base qualities and has been widely used in non-MSI applications. To our best knowledge, it has not been previously assessed whether such algorithm is suitable on microsatellite-residing families consisting of alleles with different repeat lengths.

In general, calling sample-level MSI status is performed via an initial locus-level features selection/abstraction, followed by sample-level model training (optional) and testing upon selected loci (**Table 2**). Features are either transformed into binary stability results through a statistical test first before being fed to a sample-level model, or directly fed into such a model. Kolmogorov-Smirnov test was a typical choice for the measurement of such differences and applied in a couple of studies with relatively low coverages (18, 28), while having high specificity and being too conservative (20). Another branch of likelihood maximization-based algorithms (6, 29) originally applied to genotyping of STR associated applications (30, 31) treated microsatellite-overlapping reads as data, and as parameters the intrinsic probability of extension/contraction with in a single reproduction cycle. Both methods modeled the observed read as obeying multinomial distributions regardless of the exact PCR cycles reads are produced. Other methods include tests based on Gaussian, a binary test of whether gain in allele count is greater than zero, or simply comparing the sum of stepwise differences to a fixed value. Sample-level models include those that are based on classification methods such as random forest and k-means clustering (32), and simply putting thresholds on the fraction of positive loci.

Available MSI methods are more or less ambiguous about the scope of application regarding coverages, library preparation PCR cycles, sequencing data or input mass. Relevant works include setting a minimal quality control coverage (18), setting thresholds to exclude alleles with less than 5% reads compared to the majority allele (25), setting an upper limit of 1.75% on the percentage of singletons (family of one read) that implicitly controls the family size distributions of families with more than one read (32). Direct and quantitative measurement of, or efforts to

systematically assess the impact of variations associated these parameters are lacking. In the following sections, we show that MSI algorithms based on measuring differences between test and reference signals could be more accurate if controlled for aforementioned experimental variables.

Here we present b/tMSI-CAST, an NGS-based composite MSI detection method that has a two-mode caller compatible with both tissue and plasma samples and does not require matched normal samples. Compared to existing methods and tools, b/tMSI-CAST improves on critical steps in the analysis chain including loci selection, duplication removal strategy, calling algorithms and reference comparison. Besides being an independent method that can be followed to perform MSI analysis on any given panels, the loci selection principles, duplication removal strategy and the locus-level callers can be applied independently of other parts and integrated into one's own pipeline.

Results

A set of 118 MSS PWBC samples from Chinese patients with non-blood cancers were used to filter out loci with greater than 5% population polymorphism (**Fig. 1**). Considering repeats longer than 15 would make it hard to determine population monomorphism and keeping only loci showing significantly differential DeletionRatios or KLD scores between 20 MSI-H and 20 MSS samples determined by MSI-PCR ($P < 0.05$, Rank Sum test), a total of 37 10-15 bp long A/T mononucleotide microsatellite loci with high specificity were eventually chosen. Built-in loci provided by mSINGS and MSIsensor-pro overlapped with our panel on 187 and 1344 loci and used, overlapping on zero and 37 loci with b/tMSI-CAST, respectively, and overlapping on 180 loci mutually.

Duplication removal

b/tMSI-CAST adopts a majority voting duplication removal strategy, assuming an allele of an original molecule remains the majority throughout the library preparation as well as sequencing PCRs. We noticed that longer mononucleotide repeats were frequently accompanied by lower base qualities on bases flanking the 3' direction of the repeats and the base qualities within the repeat or upstream regions were not affected, plausibly caused by phasing/unphasing issues associated with sequencers. We further hypothesized that, among family members fully spanning a microsatellite locus and having equal number of bases sequenced, the sum of base

qualities of a member with longer mononucleotide repeats might be lower than a member with shorter repeats. Under such "longer repeats with worse quality" tendency, deduplicating strategies that choose family members with the greatest sum of sequenced bases qualities, such as one applied by Picard MarkDuplicates, would biasedly choose alleles with shorter repeats around mononucleotide microsatellite loci and result in left-tilted stutters with increased DeletionRatios, decreased InsertionRatios, and a slightly decreased germline allele ratio as the net effect of the "germline allele left-shifting" and "extended alleles left-shifting" (**Table.4**). In contrast, algorithms that authentically conduct duplication removal shall end with shrinkage of the extended and contracted alleles towards the central germline allele.

To test this hypothesis, GermlineRatios, DeletionRatios and InsertionRatios were calculated with three deduplication strategies on six loci (five with reference mononucleotide repeat sizes ranging from 10-15 bp plus NR-24 (**Fig. 2**). Three types of ratios calculated without deduplication removal were flat regardless of dup-ratio as expected. A majority voting strategy led to decreased deletion-ratio, increased germline allele-ratio and decreased insertion-ratio with ascending dup-ratio. This means the post-deduplication stutter shrank towards the central allele and is intuitively correct as increased duplication level of families should indeed restore the authentic allele with a better chance. On the contrary, a sum-of-base-quality-based strategy used by Picard even had negative effect, worse than doing nothing, as indicated by decreased germline allele ratios and increased DeletionRatios. Such patterns became more obvious with longer mononucleotide repeats. Deduplication is in nature a process to discriminate the true signal from noises and the accuracy is function of dup-ratio, repeat itself, and deduplication strategy. Results here further stressed the importance of comparing results under the same dup-ratio and baseline.

Saturation curve

At a given spanning coverage, repeated random sampling of a sample resulted in a distribution of calculated deletion ratios or KLD scores. The expectation and variation both decreased with increased spanning coverage until saturation (**Fig. 3A**). As a result, the saturation points for different loci using deletion-ratio and KLD methods were determined to be in the rage of ~100-200X and ~200-350X respectively. Both locus-level scores (**Fig. 3B**) and overall sample-level score (**Fig. 3C**) monotonically correlated to dup-ratios, again highlighting the benefits of matching test samples to baselines by dup-ratios.

Analytical and empirical performance evaluation

Analytical sensitivity was evaluated on cell line samples diluted to eight levels from 0.9% to 20% (**Fig. 4A**). b/tMSI-CAST had better sensitivity at lower CCF than mSINGS and MSIsensor-pro. Specifically, KLD mode showed 75% or higher sensitivity as low as 2.0% CCF, compared to deletion-ratio mode (6.7%), mSINGS (10.0%) and MSIsensor-pro (20.0%). All samples reached saturation and dup-ratios were in the range of ~70%-80%. (**Fig. 4B**).

In an empirical analysis of 163 clinical tumor tissue samples tested by MSI-PCR (39 positive and 124 negative), cutoff MSI scores of mSINGs and MSIsensor-pro were determined as 0.078 and 0.085, corresponding to the greatest AUC through ROC analyses. Thresholds for the two modes of b/tMSI-CAST were determined by negative predictive values (NPV). By MSI-PCR results, b/tMSI-CAST deletion-ratio mode and MSIsensor-pro both showed 100% sensitivity (39/39) and 98.3% specificity (122/124). mSINGS had 97.4% sensitivity (38/39) and 98.3% specificity (122/124) (**Fig. 5A**). Two samples classified as MSI negative by MSI-PCR were positive by deletion-ratio mode, out of which one sample was also labeled positive by MSIsensor-pro and mSINGS. These samples have purity estimated in a relative low range of 20%~40% as determined by pathologists, and both were calculated to be TMB extra-high. Both patients had one stop-gain mutations found on MMR genes (GCST85: MLH1, chr3: g.37053343C>A, VAF=12.4%, p.S193X; GCST124: MLH1, chr3: g.37048493C>G, VAF=14.6%, p.S131X). We suspect the true status of these samples are MSI-high. MSI scores of these two samples calculated by mSINGS and MSIsensor-pro laid close to their thresholds (**Table. 5**). Each of the two samples (GCST122 and GCST133) classified as negative by MSI-PCR was classified positive by MSIsensor-pro and mSINGS, and again MSI scores were close to their thresholds. Neither patients had TMB high nor mutations on MMR genes, and both samples had high dup-ratios (0.94 and 0.95, respectively). We hypothesized these two cases were indeed MSS that both MSIsensor-pro and mSINGS had incorrectly classified, possibly due to the negative effects caused by the sum-of-base-quality strategy worsened by high dup-ratios as discussed above. To test the hypothesis, both samples were down-sampled to dup-ratios of 0.71 and 0.68, respectively, and classified as negative by MSIsensor-pro and mSINGS. This unintuitive observation (less data but more accurate) can be explained by the mitigated negative deduplication effects.

For a last sample (GCST35) that mSINGS classified as MSS, MSIsensor-pro and b/tMSI-CAST had resulted as MSI and agreed with MSI-PCR. To further compare the separation effects of all three methods, Cohen'd values were used to measure the effect sizes obtained using the three tools, with b/tMSI-CAST showing the greatest effect size (6.74) compared to mSINGS and MSIsensor-pro (2.82 and 5.40, respectively).

MSI-CAST with KLD mode resulted in 100.0% negative percent agreement (NPA, 95% CI 92.0%-100.0%) on 44 cfDNA samples of which 28 were from healthy donors and 16 from cancer patients with patient-matched tissues negative for MSI-PCR. Then on 29 plasma cfDNA samples with patient-matched tissue determined as positive by MSI-PCR, b/tMSI-CAST KLD mode showed an overall positive percent agreement (PPA) of 31% (9/29, 95% CI 17.3%~49.2%) for patients across all clinical stages, with a clear upward trend as the stage advances. In particular for metastatic and locally advanced stage patients (stage III and IV) the PPA of plasma MSI status called by b/tMSI-CAST KLD mode was 54.5% (6/11, 95% CI 28.0%~78.7%). (**Fig. 5B**).

MSI incidence rates by cancer types on more clinical samples

We next applied the b/tMSI-CAST algorithm on tissue and plasma samples which had been received and tested in the CAP accredited GeneCast clinical laboratory, with the DeletionRatio mode used for tumor tissue samples and KLD mode for plasma samples. 9068 tumor tissue and 9011 plasma samples from patients spreading 26 cancers were initially considered and showed an overall percentage of MSI-H rate of 1.8% in tissue and 0.7% in plasma samples regardless of cancer stages or corresponding ctDNA fraction. To investigate incidence rates by cancer types, we restricted to types with at least 20 total cases and excluded clinically undetermined types, it resulted in 7667 tumor tissue and 7600 plasma samples and similar overall MSI-H rates (1.9% for tissue and 0.7% for plasma). Considering the wide range of sources of the plasma samples in terms of clinical stages and timing of sampling which could be either pre-treatment or post-treatment, we further restrained the analysis on plasma samples with CCF values over 0.02. The prevalence of blood MSI increased to 1.4% in this population (**Fig. 6**). As expected, UCEC (18.0% and 13.6% for tissue and plasma, respectively) and PRAD (9.5% and 3.7% for tissue and plasma, respectively) had the highest MSI-H incidence rates, in agreement with previous landscape studies (8, 19, 20, 33). Among NSCLC cases, squamous cell carcinoma had slightly higher incident rates than adenocarcinoma in tissue (1.3% and 0.5%) and

plasma specimens (1.4% and 0.45%). Among CRC cases, COC had higher incident rates than REC in both tissue (8.8% and 3.0%) and plasma (4.0% vs. 2.0%).

Discussion

On a test set of 169 clinical tumor tissue samples, b/tMSI-CAST under deletion-ratio mode, mSINGS and MSIsensor-pro all classified most cases (164/169) the same as MSI-PCR results. With respect to the five disputable samples, we further considered TMB, CCF, mutations on MMR genes, dup-ratios and how close their MSI scores were to their thresholds. At low CCFs, somatic alleles may not form obvious second peaks and MSI-PCR are particularly prone to false negatives. Out of the two MSI-PCR negative samples, all three tools unanimously classified one sample (GCST85) as MSI positive, where mSINGS and MSIsensor-pro had MSI scores barely greater than thresholds. b/tMSI-CAST classified the other one (GCST124) as MSI-positive while mSINGS and MSIsensor-pro gave negative results with MSI scores again laid closely to their thresholds. TMB high, mutations on MMR genes and relatively low tumor fractions make it plausible to argue b/tMSI-CAST correctly classified the MSI status against MSI-PCR results. mSINGS and MSIsensor-pro individually classified two samples (GCST35, GCST122) differently than the other methods, the MSI score of which once again was close to the threshold. Therefore, b/tMSI-CAST showed valuable capability to better determine the MSI status in hard-to-tell samples, as also measured by Cohen'd.

In addition to tissue samples which are generally well above a CCF of 20%, b/tMSI-CAST under KLD mode was able to detect MSI positive samples at CCFs as low as 2% and exhibited 100% specificity on clinical plasma samples. Observed increasing PPAs with increasing cancer stages can be explained by positive dependence of CCFs on cancer stages (34). To apply b/tMSI-CAST under KLD mode below 2%, high dup-ratio is necessary, optionally with unique molecular identifiers and error suppressing techniques which b/tMSI-CAST is well compatible with.

Mass of input DNA, numbers of PCR cycles, sequencing coverages, and cancer cell fractions (CCF) are critical experimental variables involved with NGS-based library preparation and sequencing pipelines and variations are mostly inevitable in practical settings. Differences in these variables between a test sample and an MSS reference, if evaluated and calibrated properly, would help better identify potential somatic signals. Therefore, we incorporated b/tMSI-CAST

with a baseline method to minimize the impact of unequal experimental conditions of test and reference samples. We started by seeking a metric that is able to abstract and fully characterize relevant covariates. A sample with greater library complexity would require a higher share on a sequencing run to maintain similar dup-ratio compared to a sample with lower complexity. For fixed input DNA, increasing sequencing data would ideally shrink and center the stutter continuously while the shrinkage rate diminishes and plateaus. Saturation implies spanning coverages no longer influence the calling algorithms after certain point, after which dup-ratios fully describe the combined effect of input mass and sequencing depth. It agrees with a previous finding that there was substantial drop in called MSI events at low coverage of 20-30X (18) and no meaningful correlation between coverage and calling results in the range of 30X-100X (18) or even higher (20). Baselines shall be built for a fixed and stable experimental pipeline and inspected for suspicious outliers or abnormally high variances before being applied to test samples. The intervals between discrete dup-ratios can preferably be calculated as piecewise-linearly dependent on dup-ratio, e.g. looser intervals in high dup-ratios and denser in low ones.

Having stressed the representativity of dup-ratio, we note that the one factor that dup-ratio is not able to represent is PCR cycles. Varying PCR cycles reflects on microsatellite alleles complexity of the resulted library. Two samples with similar dup-ratios while having experienced distinctive PCR cycles might result in completely different MSI status. Therefore, to use b/tMSI-CAST or any other MSI detection tools, one would have to evaluate the influences brought by discrepant PCR cycles. In practice, the number of PCR cycles and pooling volumes are often dynamically adjusted given collectable input DNA, undermining this assumption. Researches and modeling of PCR in general (35, 36) and polymerase slippage around microsatellite loci have been ongoing for more than a decade (37, 38). The dependence of observed percentages of extension/contraction alleles after a number of cycles on varying repeat unit length, repeat unit nucleotides, number of repeats and polymerase types have recently been studied (39), providing a basis for normalization of PCR cycles of compared samples in MSI detection algorithms. Throughout this study, a total of 22-26 PCR cycles (8-12 prePCR + 14 cycles post-capture-PCR) were used which was well below 47 cycles for 10-15 A/T mononucleotide repeats and therefore we deemed it safe to neglect PCR normalization in the present pipeline. Approximately 20-30 library preparation PCR in a typical capture-based

experimental flow and 35 cycles of bridge-PCR associated with sequencers, choosing A/T mononucleotide repeats fewer than 15 should be safe.

We have only provided necessary evidences for the superiority of a majority voting deduplication strategy employed by b/tMSI-CAST over a sum-of-base-qualities strategy used by Picard exclusively on mononucleotide microsatellite loci. A sufficient prove on the matter demands in the future a thorough study and understanding of the base quality scoring mechanism immediately after a mononucleotide repeat region given a specific sequencer.

Polymerase slippage induced MSI has been observed to show preference for contraction than extension (6, 31, 38, 39), although the underlying mechanism has not entirely been clear (40). We have demonstrated in this study that deletion-ratio is a good abstraction of allele count information. Compared to DeletionRatios and features used by other MSI tools that abstract the counts of the true alleles and either one or the combination of extended and contracted alleles, KLD scores additionally capture the inherent order information of the alleles in a stutter, thus conferring more comprehensive characterization of a stutter pattern and providing higher resolution in distinguishing stutters. As the cost, KLD scores generally require higher spanning coverages to reach a low variation level than deletion ratios (**Fig. 3**). In practice, assays with panel size around or below 2-3 Mb would mostly have saturated spanning reads and the validation samples involved in the present report all satisfied this criterium. We have not validated b/tMSI-CAST on exome or whole genome panels where sub-saturation is expected and yet we speculate that b/tMSI-CAST would function as well. Theoretically, positive MSI events would reflect on variability in the number of alleles, counts of alleles and the order or the pattern of the counts of alleles. KLD method and a similar algorithm (8) that utilized more of these variables should be able to provide increased accuracy in calling challenging MSI events, e.g. low cancer cell fractions, than those only consider the number of alleles and the counts of alleles. On cfDNA samples with sufficient spanning coverage, the KLD method indeed showed higher sensitivity than the deletion ratio method at various tumor fractions, and clinical samples stratified by disease stages. Both methods achieved high and comparable specificity as mSINGS and MSISensor-pro.

Several previous studies have described the landscape of MSI status across various cancer types based on TCGA data (19, 20) or NGS data by in-house NGS panels (8, 33) in both tumor tissue (8, 19, 20, 33), and in plasma samples (8). In the current study we investigated the MSI

prevalence in multiple cancer types in 11,266 tumor tissue and 7,625 plasma cfDNA samples from Chinese patients diagnosed with solid tumors. The prevalence of plasma MSI across 13 major cancer types included for analysis has a very similar distribution pattern to that in tumor tissues, providing further validation on the reliability of the MSI assay and the b/tMSI-CAST algorithm. The MSI positive rates across cancer types are generally lower in plasma than in tissue, which can be explained by the fact that this current cohort spans wide clinical stages, and the tumor-derived ctDNA can be very low in quantity or even nonexistent in early stage cancer blood samples. On the other hand, the overall percentages of both tissue and plasma MSI-H in this multiple cancer type cohort (1.9% for tissue and 1.4% for plasma) are at comparable level to what has been reported previously (8, 19), with cervical/endometrial/uterine cancers, prostate cancer, colorectal and gastric cancers leading on the top of the list. Interestingly we have observed a slight but significant difference of MSI prevalence in adenocarcinoma and squamous cell carcinoma of the lung both in tissue and in plasma, implying an expanded population of lung cancer patients who may potentially benefit from immune checkpoint inhibitor regimen, for either metastatic treatment setting or in neoadjuvant settings in management of earlier stages of cancer.

A couple of other groups have published methods and results on blood MSI analysis on retrospective samples in metastatic diseases settings and in particular, preliminary results on clinical validity of such methods in prediction of patient response to immune checkpoint inhibitors (7, 8). The current study is limited by the fact that the clinical plasma samples used for landscape profiling compose a heterogenous mixture of cancer types, clinical stages, treatments, and timing of sampling. Clinical validity and utility of b/tMSI-CAST algorithm on prediction of sensitivity to immunotherapy remains to be investigated in prospective interventional studies. b/tMSI-CAST is more than a stand-alone tool. The loci selection principles applied on the current panel resulted in 37 loci retrospectively. Ideally, loci selection is recommended on a whole genome scale from the panel design stage. Majority voting deduplication strategy could be applied in universal panels with mononucleotide repeats. Use of dup-ratio matched baseline could almost certainly improve the calling accuracy if incorporated in pipelines of existing methods including the two benchmarked here. We have evaluated the performance of b/tMSI-CAST on samples prepared and sequenced in only a limited range of experimental conditions. However, b/tMSI-CAST can be potentially suitable for any NGS capture-based workflows. The

task of MSI detection using tumor tissue and plasma are typically treated dichotomously owing to distinct ranges of CCF with representative experimental conditions. In practice however, targeted panels can vary greatly in size and even samples of the same type could experience completely different experimental conditions, therefore posing stratified requirements on the limit of detection and difficulties. Our method features a first-of-its-kind two-mode MSI caller. Rather than a mechanical combination of two tools that handles two sample types, it was intended to cover two mutually exclusive and collectively exhaustive regions on a continuous space which is fully described by a set of experimental conditions and where sample types are just a confounding factor. Despite differed choice of reported performance metrics, a finer evaluation of all MSI detection methods calls for further comparisons on more hard-to-call samples together with other dimensions of evidences as we have illustrated.

Materials and methods

Study design and sample information

Samples involved in this study were captured using a solid tumor pan-cancer panel (543 genes, 1.6Mb) originally designed for somatic SNV and insertion/deletion mutations. A training set of 118 peripheral white blood cell (PWBC) samples and 40 tumor tissue samples (20 positive/20 negative by MSI-PCR) were used exclusively for microsatellite loci selection. 143 non-small cell lung carcinoma (NSCLC) FFPE samples with a low MSI incidence rate of ~0.6% were assumed negative and used for baseline construction of b/tMSI-CAST deletion-ratio mode, MSIsensor-pro and mSINGS. Another 173 NSCLC plasma samples were used for baseline construction of b/tMSI-CAST KLD mode. For analytical performance assessment, 50ng initial input DNA of four MSI-H cell line gDNA were sonicated and titrated with NA12878 (COBIOER Inc., China), resulting in a series of eight tumor fractions (0.9%, 2.0%, 3.0%, 4.4%, 6.7%, 10.0%, 20.0% and 100.0%). All dilutions were sequenced to an average coverage of 1100X, with an averaged 73% dup-ratio.

Empirical assessment was conducted on 133 tumor tissue samples (39 positive/94 negative by MSI-PCR), and 30 PWBC samples from healthy controls. Evaluation of b/tMSI-CAST KLD mode was run on a total of 73 plasma samples (16 negative /29 positive by MSI-PCR, 28 were healthy donors, out of which 24 have both plasma and tissue samples). A retrospective study on

18084 clinical cancer samples examined in the past two years was conducted, including 9068 cases with FFPE samples, 9016 cases with plasma samples.

Experimental procedures

gDNA and cfDNA were extracted from FFPE (Tiagen, China) and plasma samples (Thermo Fisher, USA) per instructions. Targeted region selection was performed with HyperCap Target Enrichment Kit (Roche, Swiss) following manufacturer's protocol. A fixed mass of products from 8-12 cycles of pre-capture PCR depending on input were fed into 14 cycles post-capture PCR. The captured products were sequenced using 2 x150 paired-end reads (NovaSeq 6000, Illumina).

MSI PCR experiment

PCR-based MSI detection was conducted using MSI Analysis System (Promega, Madison, WI) that includes five mononucleotide markers (BAT-25, BAT-26, NR-21, NR-24 and MONO-27) and two pentanucleotide markers (Penta C and Penta D). The latter were used to detect potential sample mismatch. Two or more positive out of five indicates sample-level MSI positive.

NGS data bioinformatic pipeline

Adaptors of raw read pairs were trimmed using Trimmomatic (v0.36) (41). Clean reads were mapped against the human reference genome (build hg19, UCSC) and aligned using BWA bwa-mem (v0.7.12) (42) and sorted using SAMtools (v1.7) (43). A detailed comparison of majority voting-based deduplication strategy applied by b/tMSI-CAST and Picard MarkDuplicates (v2.1.0) is described in the duplication removal section. MarkDuplicates was performed under the default mode, followed by local indel realignment using GATK version v3.7. For each microsatellite locus, all spanning read pairs were extracted from realigned BAM file. Spanning reads are defined as fully covering the coordinates of the microsatellite in reference plus 2 bp in both 5' and 3' directions, while a pair of spanning reads only counts as one. Following deduplication, the length of the mononucleotide repeat in each deduped alignment was counted and tallied by lengths.

Main quality control criteria are as follows: 1. lower limit on input DNA was 20 ng for tissue and 5 ng for cfDNA. 2. For deletion ratio and KLD methods, lower spanning coverage thresholds were separately determined for each locus through saturation curves as described below. 3. A minimum of 15 loci passing the above thresholds are required.

Microsatellite loci selection and filtering.

b/tMSI-CAST adopts a set of heuristics to screen for high-performance microsatellites given a panel. First, discard microsatellites suffering from low capture efficiency as evaluated by the ratio of spanning reads to the average coverage in training samples. Second, screen for microsatellites that shows significant similarities to 10-bp sequences in the 5' or 3' direction as calculated by

$$s = \sum_{i=1,\ n_i=n_{ms}}^{10} \frac{11-i}{10}$$

Where n_i is the i-th nucleotide next to the microsatellite, n_{ms} represents the repeat unit nucleotide of the microsatellite and only count those n_i same as n_{ms}. If s in either direction exceeds a threshold of 2, leave the locus out for future considerations. Third, choose loci with A and T mononucleotide repeats with the number of repeats bounded between 10 and 15. Dinucleotide or higher repeats are generally associated with lower sensitivity and higher population polymorphism albeit being used occasionally. Fourth, evaluate the monomorphism of candidate loci with a group of normal samples. A locus is considered monomorphic if more than 95% samples have germline allele frequency greater than a defined threshold, with longer repeats corresponding to lower thresholds. Repeat length longer than 15 is not recommended as the monomorphism becomes harder to determine. Fifth, perform a significance test (i.e. Rank Sum test, $p < 0.05$) on candidate loci with MSI status labeled samples.

Spanning coverage

Spanning coverage is a more accurate metric as non-spanning coverage provides invalid information about a locus. We statistically determined a saturation spanning coverage beyond which calculated scores become invariant. Specifically, deletion ratio and KLD scores at a given spanning coverage were assumed Gaussian and Gamma distributed, respectively, and the expectation corresponding to each spanning coverage was calculated. Iterative linear regressions on a dynamic window of three consecutive spanning coverages from low to high were conducted to calculate the curvature of a saturation curve. A spanning coverage was determined as the saturation point as soon as the curvature reached below a defined threshold.

Duplication removal

b/tMSI-CAST is equipped with a self-contained deduplication strategy that proves more accurate on mononucleotide microsatellites than base quality-based strategies used by tools such as Picard, which tends to choose shorter repeats (see result and discussion sections for details).

The duplication removal is performed as follows. First, determine spanning reads and group families by the start position and insert size. If unique molecular identifiers are used, group families by identical UMI, start position and insert size. Second, discard families with fewer than 'min-reads' contributing reads. This parameter ought to be kept consistent among processes for test samples and baselines. Third, discard family members with a mapping quality lower than a specified threshold. Fourth, count observed microsatellite repeat length of all members by CIGAR values despite of SNV mutations. Fifth, choose the mode of the family. If there are more than one most frequent families, discard the family.

Calculation of dup-ratio

Dup-ratios stand an averaged abstraction of family size distributions across families spanning the same microsatellite locus. Dup-ratio is calculated per microsatellite locus per sample as

$$dup\_ratio = \frac{\sum_{k=1}^{N_{family}} (n_{total,k} - 1)}{\sum_{k=1}^{N_{family}} n_{total,k}}$$

where k is the index of total N_{family} families covering the locus, n_{total} is the number of total alignments.

MSI calling

We first define or clarify the terminology that will be used throughout this report. An original single strand molecule and its PCR duplications constitute a family, duplication removal of which result in a called allele associated with a specific microsatellite locus. Counts of a mixture of alleles either truly or falsely called from families covering the same microsatellite loci are often tallied in a histogram, constituting a so-called stutter. Theoretically, larger families should generally result in more accurate deduplication results regardless of the mutation type. In the task of MSI detection, this means more shrank and centered post-deduplication stutters. For methods free of duplication removal, i.e. MSI-PCR and amplicon-based MSI methods (29), stutter is a mixture of alleles from all families. Therefore, deconvolution of a family is a problem associated with duplication removal while locus-level MSI calling is essentially a task to tell apart *in vivo* MSI-positive-cell-originated somatic signal stutters from germline stutters. The ratio of the count of deleted/contraction alleles, inserted/extension alleles and the sum of these two to the total number of alleles are termed deletion-ratio, insertion-ratio and noise-ratio, respectively.

b/tMSI-CAST locus-level MSI status calling can operate under two alternative modes that give best performance in different combinations of experimental conditions. The binary locus-

level results are further fed into a second module that calls sample-level MSI status or any tools that call sample-level MSI status from binary locus-level input. The first locus-level caller is based on comparing DeletionRatios calculated for test samples against baseline values at each locus and labeled unstable if exceeding the mean plus 4 times the standard deviation. Baselines corresponding to a grid of discrete dup-ratios should be built in advance with a group of MSS samples down-sampled to each dup-ratio level. The second locus-level caller calculates Kullback-Leibler divergence (KLD) (44) as a score as calculated by

$$D_{KL}(p|q) = \sum_{x} p(x) \cdot log \frac{p(x)}{q(x)}$$

where $p(x)$ and $q(x)$ are test sample and baseline allele frequency distributions over a set of discrete allele lengths, respectively. To prepare $q(x)$ at a specific dup-ratio, a group of MSS samples are down-sampled to that dup-ratio, and then the allele frequencies of each present allele are averaged to represent the corresponding probability of that allele in $q(x)$. During a test, the null hypothesis is that $p(x)$ at a locus is indistinguishable from the baseline distribution $q(x)$, determined by the calculated KLD score. The threshold is determined through an independent set of MSS samples, each of which is used to calculate KLD with respect to $q(x)$. This represents a background level and any observed sample with KLD score falling outside the 99.7% confidence interval is considered unstable at that specific locus.

b/tMSI-CAST classifies sample dichotomously MSI/MSS using binary results from either the deletion ratio or the KLD method (45). The fraction of unstable loci out of the total number of loci passing QC check was calculated and compared to a trained threshold.

Identification of somatic mutations and assessment of tumor mutational burden (TMB)

TMB has been shown in correlation with MSI status clinical responses (46). To gain etiological insight on the MSI status of discrepant samples, the TMB value, tumor purity and mutation in four MMR genes (MLH1, MSH2, PMS2, MSH6) of the two samples were evaluated. VarScan2 was used to call SNV with the following filters: (i) located in intergenic regions or intronic regions; (ii) synonymous SNVs; (iii) allele frequency $\geq$ 0.002 in the database exac03 or gnomad_exome; (iv) the value of ljb2_pp2hdiv = “B” and ljb2_pp2hvar = “B”; (v) allele frequency <0.05 in the tumor sample ; and (vi) allele depth <5. For the determination of TMB, the number of somatic nonsynonymous SNVs (with depth >100X and allele frequency $\geq$ 0.05) detected on NGS (interrogating Mb of the genome) were quantified, and the value was

extrapolated to the whole panel using a validated algorithm. Alterations likely or known to be *bona fide* oncogenic drivers were excluded. TMB was measured in mutations per Mb and were divided into four groups: low (<3.8 mutations/Mb), intermediate (3.8–8.4 mutations/Mb), high (8.4≥13.8 mutations/Mb) and extra-high (≥13.8 mutations/Mb).

Estimation of ctDNA Content Fraction (CCF)

CCF of plasma samples were estimated by a maximum likelihood model based on SNVs and CNVs in the paired blood cell and plasma samples. Somatic and germline SNPs met the following criteria were used to build the model: 1) with a minimum depth of 50x in the paired samples; 2) not on genes with high polymorphism; 3) no InDels in the 50bp upstream or downstream regions; 4) not in a copy number gain region; 5) germline SNPs with significantly different variant allele frequencies (VAFs) in the paired samples, or somatic SNPs with VAFs significantly higher than background noise. These SNPs were defined as informative SNPs and clustered into multiple groups according to their VAFs, local copy numbers and hypothetic genotypes. The hypothetic genotypes in cfDNA and ctDNA were determined by the VAFs in the paired samples and the copy number in the plasma sample. Each cluster represents a unique ctDNA source. We then calculated the likelihood of observing SNPs under given CCFs in each cluster. By maximizing the likelihood, CCF of each cluster could therefore be estimated. Cluster with the highest CCF was considered to be from the main source of ctDNA, and its CCF was output as the final estimation.

Author contributions

WL, WC and JH designed the project. FH, JH, JY and HZ provided the clinical samples. LZ implemented b/tMSI-CAST scripts. LZ, JS, HX, and JNY formulated the algorithms and performed data analysis. XW and YH performed the wet lab experiments. HX and LZ wrote the manuscript. All authors read and approved the final manuscript.

Competing interests

All other authors declared that they have no competing interests.

Acknowledgments

This work is supported in part by National Natural Science Foundation of China No.31660312. The authors thank the patients and staff at the First Affiliated Hospital of Kunming Medical University.

Data

Oral and written consent was obtained from the study participants or their representatives prior to enrollment. The studies were approved by the First Affiliated Hospital of Kunming Medical University Ethics Committee (2017-L3).

b/tMSI-CAST can be applied as a complete tool, or the integral parts other than the calling algorithms can be combined with a chosen calling algorithm in general, all of which are freely accessible at https://github.com/GeneCast/btMSI-CAST.

The raw sequence data reported in this paper have been deposited in the Genome Sequence Archive (47) in National Genomics Data Center (48), Beijing Institute of Genomics (China National Center for Bioinformation), Chinese Academy of Sciences, under accession number subHRA000493 that are publicly accessible at http://bigd.big.ac.cn/gsa-human.

Table 1 A comparative summary of MSI detection methods

	Type[a]	Tool availability	Sample type	Matched required	Min. depth	Deduplication method	Model		
							Feature abstraction	Locus-level	Sample-level
mSINGS (25)	Method	Yes	Tissue	No	30	GATK Picard	Allele count	Tumor vs. baseline mean allele counts assuming Gaussian (MSI if exceeding 3 X SD)	20% positive sites
MOSAIC (20)	Method & landscape	Yes	Tissue	Yes	30X	GATK Picard	Gain in allele count & chr8:7679723–7679741	mSINGS + "Gain in allele count>0"	random forest
MANTIS (26)	Method	Yes	Tissue	Yes	N/A	GATK Picard	Peak-count	sum of stepwise difference	average of locus-level scores scaled from 0 to 2
MSIsensor (27)	method	Yes	Tissue	Yes	20X	Assume input of deduped BAM	Allele length counts	tumor vs. normal through chi-square test with Benjamini correction	Fraction of positive sites
MSIsensor-pro (6)	method	Yes	Tissue	No	20X	Assume input of deduped BAM	The probability p of deletion for each replication step	tumor vs. baseline assuming Gaussian (MSI if p exceeding 3 X SD)	Fraction of positive sites

Kim et al. 2013 & Cortes-Ciriano et al. 2017	Method & Landscape	Upon request	Tissue	Yes	30X	N/A	Allele length counts	compare tumor to matched normal through a Kolmogorov-Smirnov test	random forest + conformal prediction
MSI NGS (32)	Assay	No	Tissue	No	500X (PE300)	N/A	Allele count & avg. length of alleles	k-means: Euclidean distance between a test sample and the centroid of training clusters	
Willis et al. 2019 (8)	Assay & landscape	No	cfDNA	No	N/A	N/A	Allele frequencies	AIC model: tumor vs. baseline	Fraction of positive sites
Georgiadis et al. 2019 (7)	Assay	No	cfDNA	No	N/A	N/A	Local peaks	3 bp shorter than hg19 reference	2 loci classified as MSI
bMSISEA (49)	Method	Yes	cfDNA	No	N/A	Assume input of deduped BAM	Read coverage pattern of MSS/MSI-H samples	P value calculated via a hypothesis test assuming hypergeometric distribution	MS-score = sum of H value of each marker normalized by baseline mean, with cutoff determined by CV using tumor vs. baseline assuming Gaussian (MSI if p exceeding 3 X SD)

a. Assay: panel and method;

b. In this study, dilutions of MSI-H cell line DNA with MSS cell line DNA resulted in a purity LOD of 2.5%, while mixing of tumor tissue with pathologically determined tumor content and adjacent normal tissue resulted in a purity of 10%.

c. The study did not explicitly specify a lower purity limit. 5% is merely the lowest tested and reported.

Table 2 Theoretical dependence of GermlineRatios, DeletionRatios, and InsertionRatios on increasing dup-ratio using three duplication removing strategies

The DeletionRatios, InsertionRatios and GermlineRatio were predicted to be in the order of majority-voting < deduplication-free < sum-of-base-qualities, sum-of-base-qualities < deduplication-free < majority-voting, majority-voting ~= sum-of-base-qualities < deduplication-free, respectively.

Strategies	DeletionRatio	GermlineRatio	InsertionRatio
Deduplication-free	-	-	-
Sum-of-base-qualities	Increasing	Slightly decreasing	Decreasing
Majority-voting	Decreasing	Increasing	Decreasing

Table 3 Information for 5 disputable samples

PID	TMB	CCF	Dup-ratio	MSI-PCR	Deletion-ratio (score)	mSINGS (score)	MSIsensor-pro (score)
GCST85	Extra High	20-40%	0.83	MSS	MSI-H (0.56)	MSI-H (0.08)	MSI-H (0.10)
GCST122	Moderate	20-40%	0.94	MSS	MSS (0.03)	MSS (0.07)	MSI-H (0.13)
GCST124	Extra High	20-40%	0.73	MSS	MSI-H (0.32)	MSS (0.01)	MSS (0.04)
GCST35	Extra High	N/A	0.78	MSI-H	MSI-H (0.38)	MSS (0.06)	MSI-H (0.29)
GCST133	Low	40-60%	0.95	MSS	MSS (0.00)	MSI-H (0.10)	MSS (0.08)

References

1. Hampel H, Frankel W, Panescu J, Lockman J, Sotamaa K, Fix D, et al. Screening for Lynch syndrome (hereditary nonpolyposis colorectal cancer) among endometrial cancer patients. Cancer Res. 2006;66(15):7810-7.
2. Le DT, Uram JN, Wang H, Bartlett BR, Kemberling H, Eyring AD, et al. PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. N Engl J Med. 2015;372(26):2509-20.
3. Le DT, Durham JN, Smith KN, Wang H, Bartlett BR, Aulakh LK, et al. Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade. Science. 2017;357(6349):409-13.
4. Boland CR, Thibodeau SN, Hamilton SR, Sidransky D, Eshleman JR, Burt RW, et al. A National Cancer Institute Workshop on Microsatellite Instability for cancer detection and familial predisposition: development of international criteria for the determination of microsatellite instability in colorectal cancer. Cancer Res. 1998;58(22):5248-57.
5. Baudrin LG, Deleuze JF, How-Kit A. Molecular and Computational Methods for the Detection of Microsatellite Instability in Cancer. Front Oncol. 2018;8:621.
6. Jia P, Yang X, Guo L, Liu B, Lin J, Liang H, et al. MSIsensor-pro: Fast, Accurate, and Matched-normal-sample-free Detection of Microsatellite Instability. Genomics Proteomics Bioinformatics. 2020.
7. Georgiadis A, Durham JN, Keefer LA, Bartlett BR, Zielonka M, Murphy D, et al. Noninvasive Detection of Microsatellite Instability and High Tumor Mutation Burden in Cancer Patients Treated with PD-1 Blockade. Clin Cancer Res. 2019;25(23):7024-34.
8. Willis J, Lefterova MI, Artyomenko A, Kasi PM, Nakamura Y, Mody K, et al. Validation of Microsatellite Instability Detection Using a Comprehensive Plasma-Based Genotyping Panel. Clin Cancer Res. 2019;25(23):7035-45.
9. Gerstung M, Papaemmanuil E, Campbell PJ. Subclonal variant calling with multiple samples and prior knowledge. Bioinformatics. 2014;30(9):1198-204.
10. Kamps-Hughes N, McUsic A, Kurihara L, Harkins TT, Pal P, Ray C, et al. ERASE-Seq: Leveraging replicate measurements to enhance ultralow frequency variant detection in NGS data. PLoS One. 2018;13(4):e0195272.

11. Newman AM, Lovejoy AF, Klass DM, Kurtz DM, Chabon JJ, Scherer F, et al. Integrated digital error suppression for improved detection of circulating tumor DNA. Nat Biotechnol. 2016;34(5):547-55.
12. Goel A, Nagasaka T, Hamelin R, Boland CR. An optimized pentaplex PCR for detecting DNA mismatch repair-deficient colorectal cancers. PLoS One. 2010;5(2):e9393.
13. Zhu L, Huang Y, Fang X, Liu C, Deng W, Zhong C, et al. A Novel and Reliable Method to Detect Microsatellite Instability in Colorectal Cancer by Next-Generation Sequencing. J Mol Diagn. 2018;20(2):225-31.
14. Waalkes A, Smith N, Penewit K, Hempelmann J, Konnick EQ, Hause RJ, et al. Accurate Pan-Cancer Molecular Diagnosis of Microsatellite Instability by Single-Molecule Molecular Inversion Probe Capture and High-Throughput Sequencing. Clin Chem. 2018;64(6):950-8.
15. Middha S, Zhang L, Nafa K, Jayakumaran G, Wong D, Kim HR, et al. Reliable Pan-Cancer Microsatellite Instability Assessment by Using Targeted Next-Generation Sequencing Data. JCO Precis Oncol. 2017;2017.
16. Latham A, Srinivasan P, Kemel Y, Shia J, Bandlamudi C, Mandelker D, et al. Microsatellite Instability Is Associated With the Presence of Lynch Syndrome Pan-Cancer. J Clin Oncol. 2019;37(4):286-95.
17. Bacher JW, Flanagan LA, Smalley RL, Nassif NA, Burgart LJ, Halberg RB, et al. Development of a fluorescent multiplex assay for detection of MSI-High tumors. Dis Markers. 2004;20(4-5):237-50.
18. Cortes-Ciriano I, Lee S, Park WY, Kim TM, Park PJ. A molecular portrait of microsatellite instability across multiple cancers. Nat Commun. 2017;8:15180.
19. Bonneville R, Krook MA, Kautto EA, Miya J, Wing MR, Chen HZ, et al. Landscape of Microsatellite Instability Across 39 Cancer Types. JCO Precis Oncol. 2017;2017.
20. Hause RJ, Pritchard CC, Shendure J, Salipante SJ. Classification and characterization of microsatellite instability across 18 cancer types. Nat Med. 2016;22(11):1342-50.
21. Fujimoto A, Fujita M, Hasegawa T, Wong JH, Maejima K, Oku-Sasaki A, et al. Comprehensive analysis of indels in whole-genome microsatellite regions and microsatellite instability across 21 cancer types. Genome Res. 2020.
22. Buhard O, Cattaneo F, Wong YF, Yim SF, Friedman E, Flejou JF, et al. Multipopulation analysis of polymorphisms in five mononucleotide repeats used to determine the microsatellite instability status of human tumors. J Clin Oncol. 2006;24(2):241-51.
23. Buhard O, Suraweera N, Lectard A, Duval A, Hamelin R. Quasimonomorphic mononucleotide repeats for high-level microsatellite instability analysis. Dis Markers. 2004;20(4-5):251-7.
24. Hatch SB, Lightfoot HM, Jr., Garwacki CP, Moore DT, Calvo BF, Woosley JT, et al. Microsatellite instability testing in colorectal carcinoma: choice of markers affects sensitivity of detection of mismatch repair-deficient tumors. Clin Cancer Res. 2005;11(6):2180-7.
25. Salipante SJ, Scroggins SM, Hampel HL, Turner EH, Pritchard CC. Microsatellite instability detection by next generation sequencing. Clin Chem. 2014;60(9):1192-9.
26. Kautto EA, Bonneville R, Miya J, Yu L, Krook MA, Reeser JW, et al. Performance evaluation for rapid detection of pan-cancer microsatellite instability with MANTIS. Oncotarget. 2017;8(5):7452-63.

27. Niu B, Ye K, Zhang Q, Lu C, Xie M, McLellan MD, et al. MSIsensor: microsatellite instability detection using paired tumor-normal sequence data. Bioinformatics. 2014;30(7):1015-6.
28. Kim TM, Laird PW, Park PJ. The landscape of microsatellite instability in colorectal and endometrial cancer genomes. Cell. 2013;155(4):858-68.
29. Walker CJ, Miranda MA, O'Hern MJ, Blachly JS, Moyer CL, Ivanovich J, et al. MonoSeq Variant Caller Reveals Novel Mononucleotide Run Indel Mutations in Tumors with Defective DNA Mismatch Repair. Hum Mutat. 2016;37(10):1004-12.
30. Gymrek M, Golan D, Rosset S, Erlich Y. lobSTR: A short tandem repeat profiler for personal genomes. Genome Res. 2012;22(6):1154-62.
31. Willems T, Zielinski D, Yuan J, Gordon A, Gymrek M, Erlich Y. Genome-wide profiling of heritable and de novo STR variations. Nat Methods. 2017;14(6):590-2.
32. Pabla S, Andreas J, Lenzo FL, Burgher B, Hagen J, Giamo V, et al. Development and analytical validation of a next-generation sequencing based microsatellite instability (MSI) assay. Oncotarget. 2019;10(50):5181-93.
33. Vanderwalde A, Spetzler D, Xiao N, Gatalica Z, Marshall J. Microsatellite instability status determined by next-generation sequencing and compared with PD-L1 and tumor mutational burden in 11,348 patients. Cancer Med. 2018;7(3):746-56.
34. Abbosh C, Birkbak NJ, Swanton C. Early stage NSCLC - challenges to implementing ctDNA-based screening and MRD detection. Nat Rev Clin Oncol. 2018;15(9):577-86.
35. Sun F. The polymerase chain reaction and branching processes. J Comput Biol. 1995;2(1):63-86.
36. Pritchard L, Corne D, Kell D, Rowland J, Winson M. A general model of error-prone PCR. J Theor Biol. 2005;234(4):497-509.
37. Clarke LA, Rebelo CS, Goncalves J, Boavida MG, Jordan P. PCR amplification introduces errors into mononucleotide and dinucleotide repeat sequences. Mol Pathol. 2001;54(5):351-3.
38. Leclercq S, Rivals E, Jarne P. DNA slippage occurs at microsatellite loci without minimal threshold length in humans: a comparative genomic approach. Genome Biol Evol. 2010;2:325-35.
39. Raz O, Biezuner T, Spiro A, Amir S, Milo L, Titelman A, et al. Short tandem repeat stutter model inferred from direct measurement of in vitro stutter noise. Nucleic Acids Res. 2019;47(5):2436-45.
40. Yoshioka KI, Matsuno Y, Hyodo M, Fujimori H. Genomic-Destabilization-Associated Mutagenesis and Clonal Evolution of Cells with Mutations in Tumor-Suppressor Genes. Cancers (Basel). 2019;11(11).
41. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114-20.
42. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754-60.
43. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9.
44. Kullback S LR. On information and sufficiency. The annals of mathematical statistics. 1951 Mar 1;22(1):79-86.

45. Laghi L, Bianchi P, Malesci A. Differences and evolution of the methods for the assessment of microsatellite instability. Oncogene. 2008;27(49):6313-21.
46. Romero D. TMB is linked with prognosis. Nat Rev Clin Oncol. 2019;16(6):336.
47. Wang Y, Song F, Zhu J, Zhang S, Yang Y, Chen T, et al. GSA: Genome Sequence Archive<sup/>. Genomics Proteomics Bioinformatics. 2017;15(1):14-8.
48. National Genomics Data Center M, Partners. Database Resources of the National Genomics Data Center in 2020. Nucleic Acids Res. 2020;48(D1):D24-D33.
49. Cai Z, Wang Z, Liu C, Shi D, Li D, Zheng M, et al. Detection of Microsatellite Instability from Circulating Tumor DNA by Targeted Deep Sequencing. J Mol Diagn. 2020.

Figure legends

Figure 1 NoiseRatios calculated at loci with varying lengths

Box plots showing NoiseRatios (sum of the DeletionRatio and InsertionRatio) calculated at six representative loci covering a range of numbers of mononucleotide repeats (5 loci of length 10-15 and NR-24), using 118 PWBC samples from patients with solid tumors. Loci with greater than 5% population polymorphism were filtered out already. Sample-level polymorphism was determined as showing a NoiseRatio significantly higher than the population mean. NoiseRatios were positively correlated with repeat lengths. Points beyond the top whisker indicate possible sporadic population polymorphism, evidencing the need of loci selection. With-in group mean and variation also roughly increased with longer repeats. Whiskers correspond to 1.5 interquartile range (IQR).

Figure 2 Characterization and comparison of three deduplication strategies

GermlineRatios (**A**), DeletionRatios (**B**), and InsertionRatios (**C**) calculated with three deduplication strategies, namely a sum-of-base-quality strategy by Picard MarkDuplicates, a majority-voting strategy used by b/tMSI-CAST and counting without deduplication. A range of dup-ratios were created via down-sampling of one plasma sample (GCSP49) for six representative loci. With increasing sequencing data and therefore dup-ratios, GermlineRatios, DeletionRatios and InsertionRatios are expected to increase, decrease and decrease, respectively, after deduplication. A majority-voting strategy succeeded (green) while a sum-of-base-quality strategy (blue) had even negative effects compared to doing nothing (red), obvious at NR-24.

Figure 3 Characterization of the dependence of locus-level and sample-level scores on spanning coverage or dup-ratios

(**A**) At each locus, pre-deduplication data from 195 NSCLC clinical samples were down-sampled to a range of predetermined spanning coverages, the expectations of which for all samples were

plotted as connected dots. Such saturation curves showed that both the averages and variations of calculated DeletionRatios and KLD scores stabilized with increasing spanning coverages until saturation. (**B**) DeletionRatios and KLD scores decreased with increasing dup-ratios calculated for 5 representative mononucleotide microsatellite loci. 143 tissue and 173 plasma samples, respectively, were down-sampled to the same dup-ratios created *in silico* ranged from 0.2 to 0.9 with a step size of 0.01. (**C**) Sample-level MSI scores generally increased with increasing dup-ratios, illustrated by four cell line samples diluted to eight levels.

Figure 4 Limit of detection comparison of b/tMSI-CAST under both modes and two benchmarked tools

Probit regression performed on sensitivities calculated with DeletionRatios and KLD modes compared to MSIsensor-pro and mSINGS on four diluted cell line samples ((22RV1, DLD1, LNCAP, RL952)) at eight CCF levels (0.9%, 2.0%, 3.0%, 4.4%, 6.7%, 10.0%, 20.0% and 100.0%). 95% confidence intervals are shown as shades.

Figure 5 Empirical evaluation of b/tMSI-CAST under two modes on clinical samples

(**A**) Beeswarm plot showing the classification results on 163 tissue samples under deletion-ratio mode, benchmarking mSINGS and MSIsensor-pro. Dots are colored red (MSS) or black (MSI) if agreeing with MSI-PCR, green otherwise. Box plots for both MSI and MSS groups were plotted for each detection method. Cohen'd was calculated for each method accordingly. (**B**) Calculated positive prediction values (PPV) by tumor stages (I to IV) under KLD mode tested on 29 cfDNA samples which are tested positive by MSI-PCR on patient-matched tissue specimen. For each stage, the total number of cases (positive by b/tMSI-CAST) was labeled.

Figure 6 A landscape view of the incident rates of MSI-H cases across cancer types

Histogram of the total counts (labeled) and incident rates (y-axis) of MSI-H cases from tumor tissue samples (upwards) and plasma cfDNA samples (downwards) across 13 cancer types with highest incident rates. Two criteria (all samples or CCF greater than 2%) for including plasma samples were used.

Tables

Table 1 A comparative summary of MSI detection methods

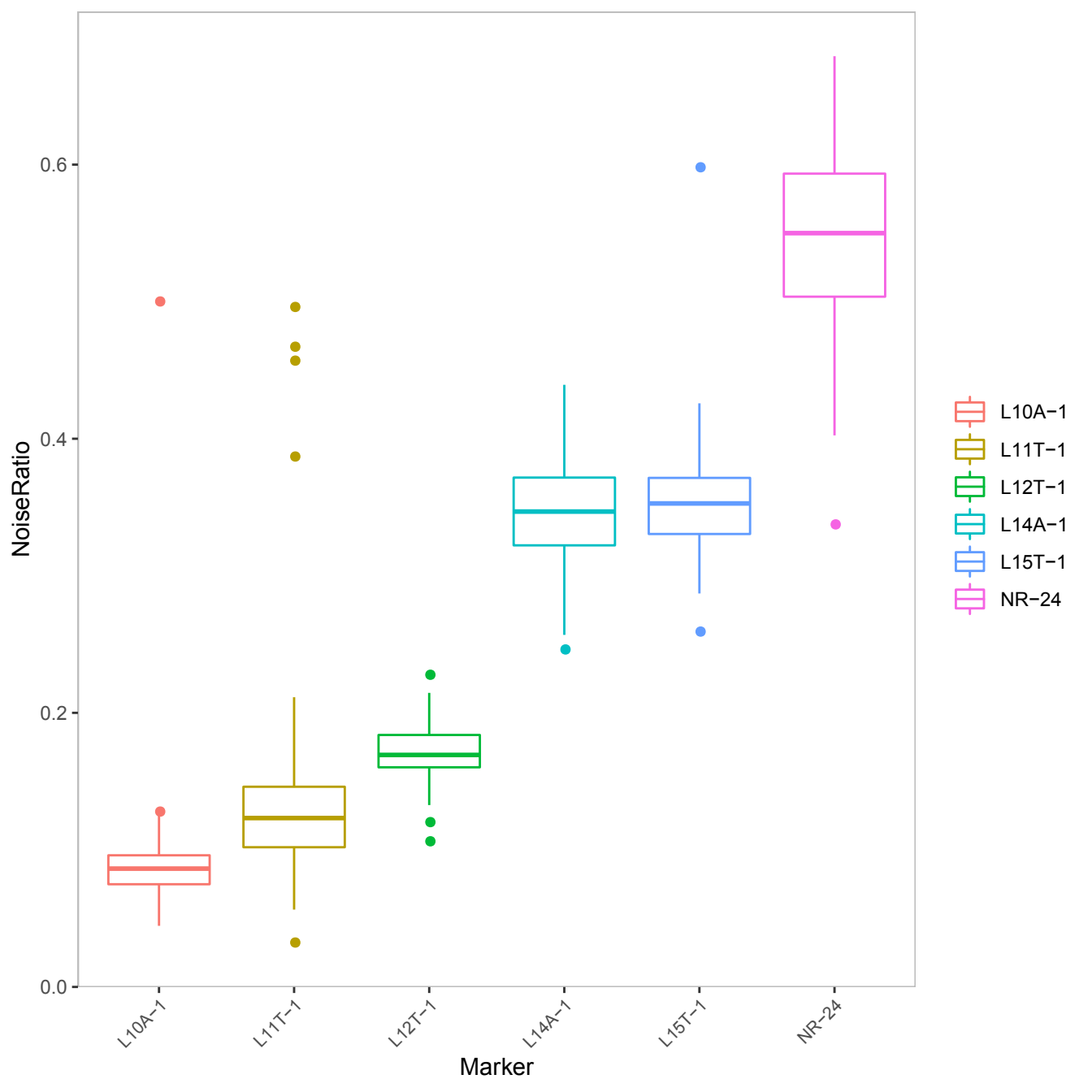
Table 2 Theoretical dependence of GermlineRatios, DeletionRatios, and InsertionRatios on increasing dup-ratio using three duplication removing strategies

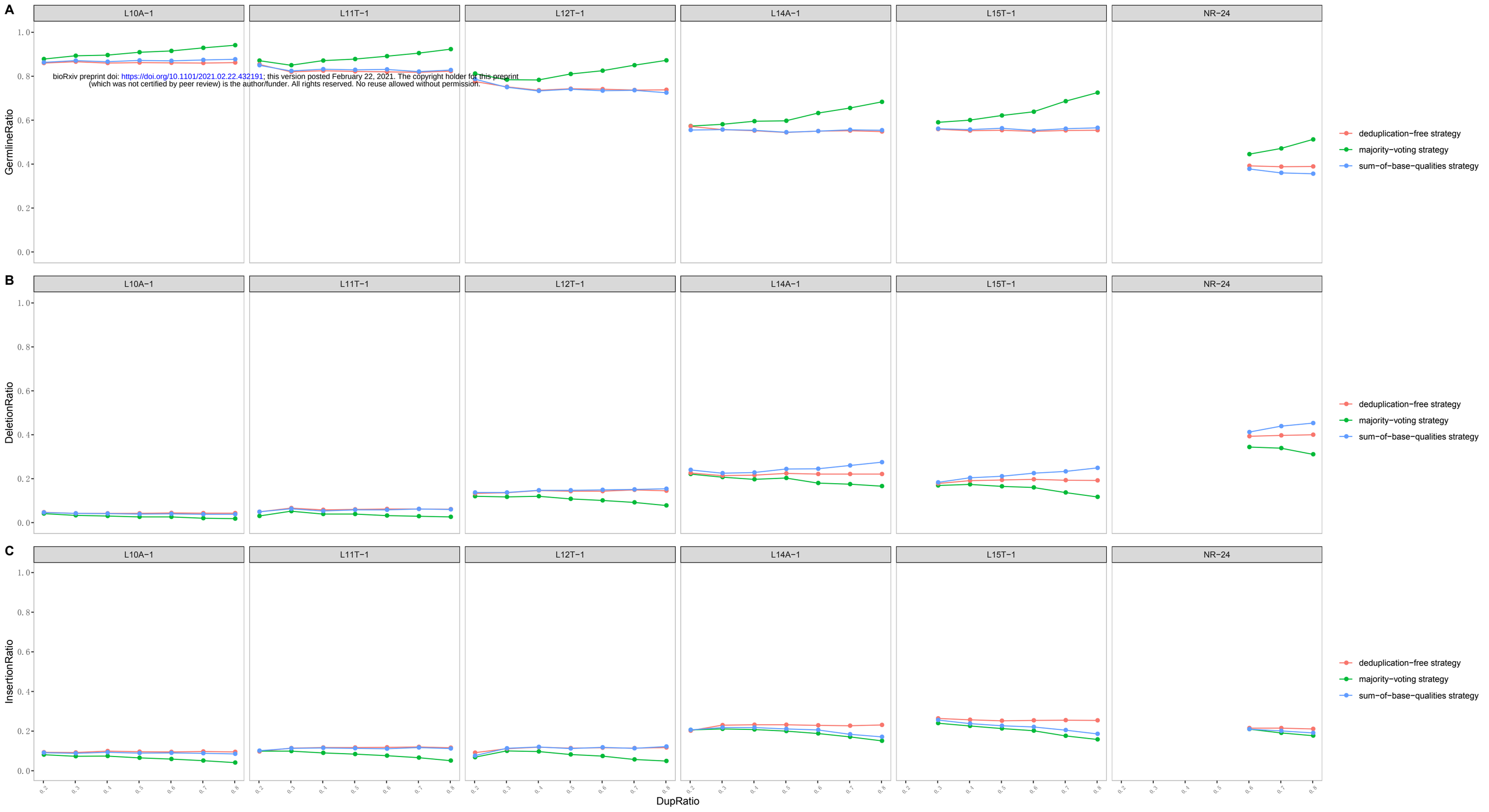
Table 3 Information for 5 disputable samples

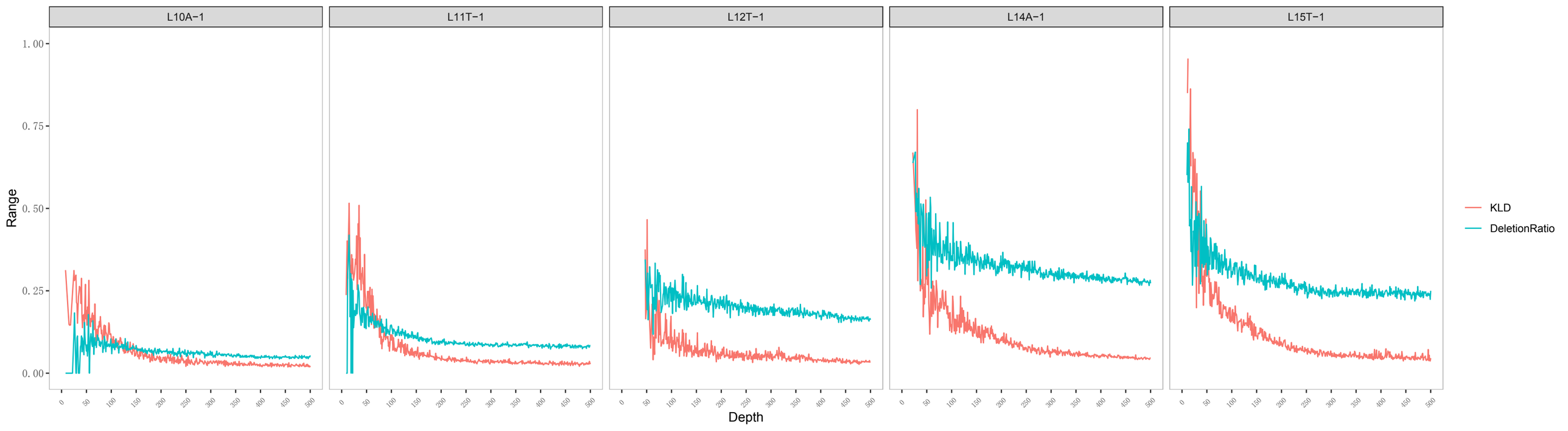
Supplementary material

Supplementary Table 1 Sample information

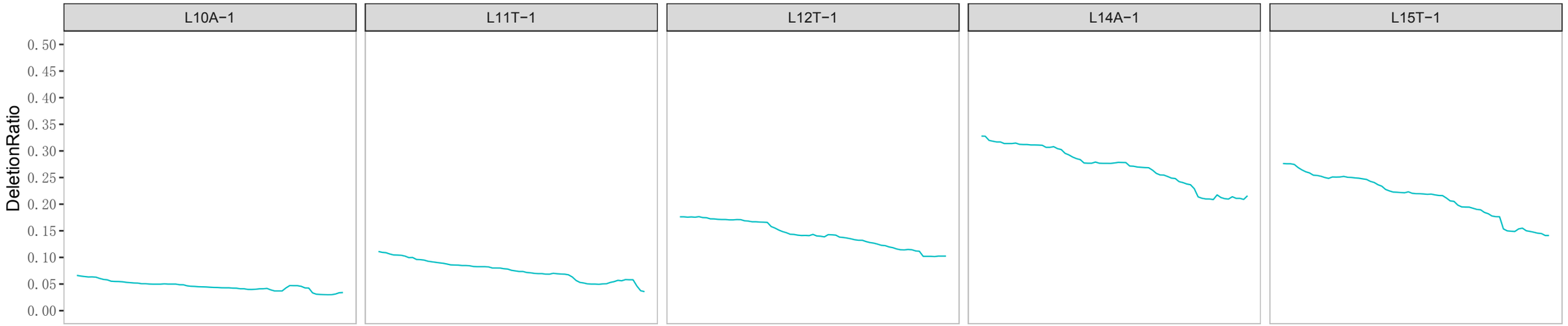
Supplementary Table 2 Detailed information of MSI loci selected and used in this study


NoiseRatio
0.0
0.2
0.4
0.6
L10A−1
L11T−1
L12T−1
L14A−1
L15T−1
NR−24
Marker
L10A−1
L11T−1
L12T−1
L14A−1
L15T−1
NR−24

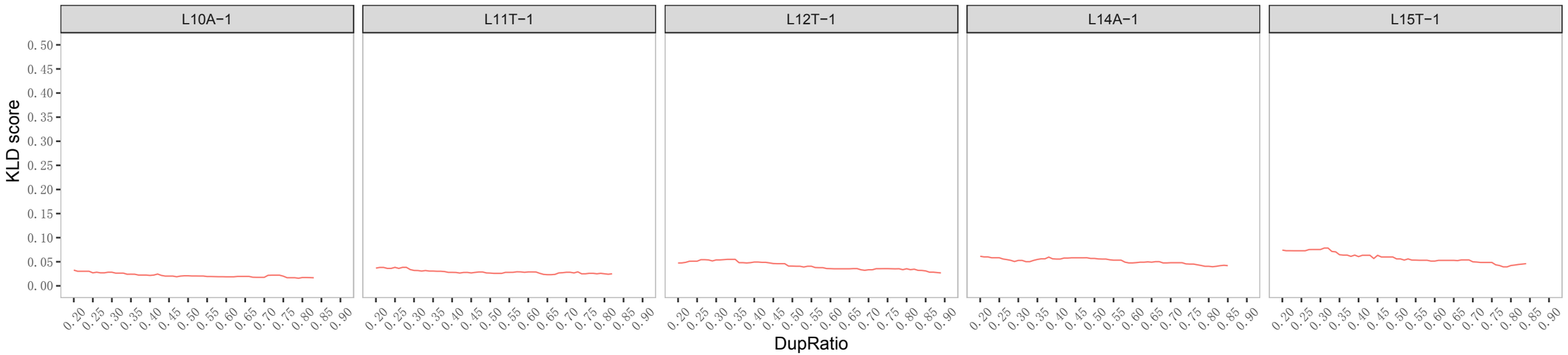


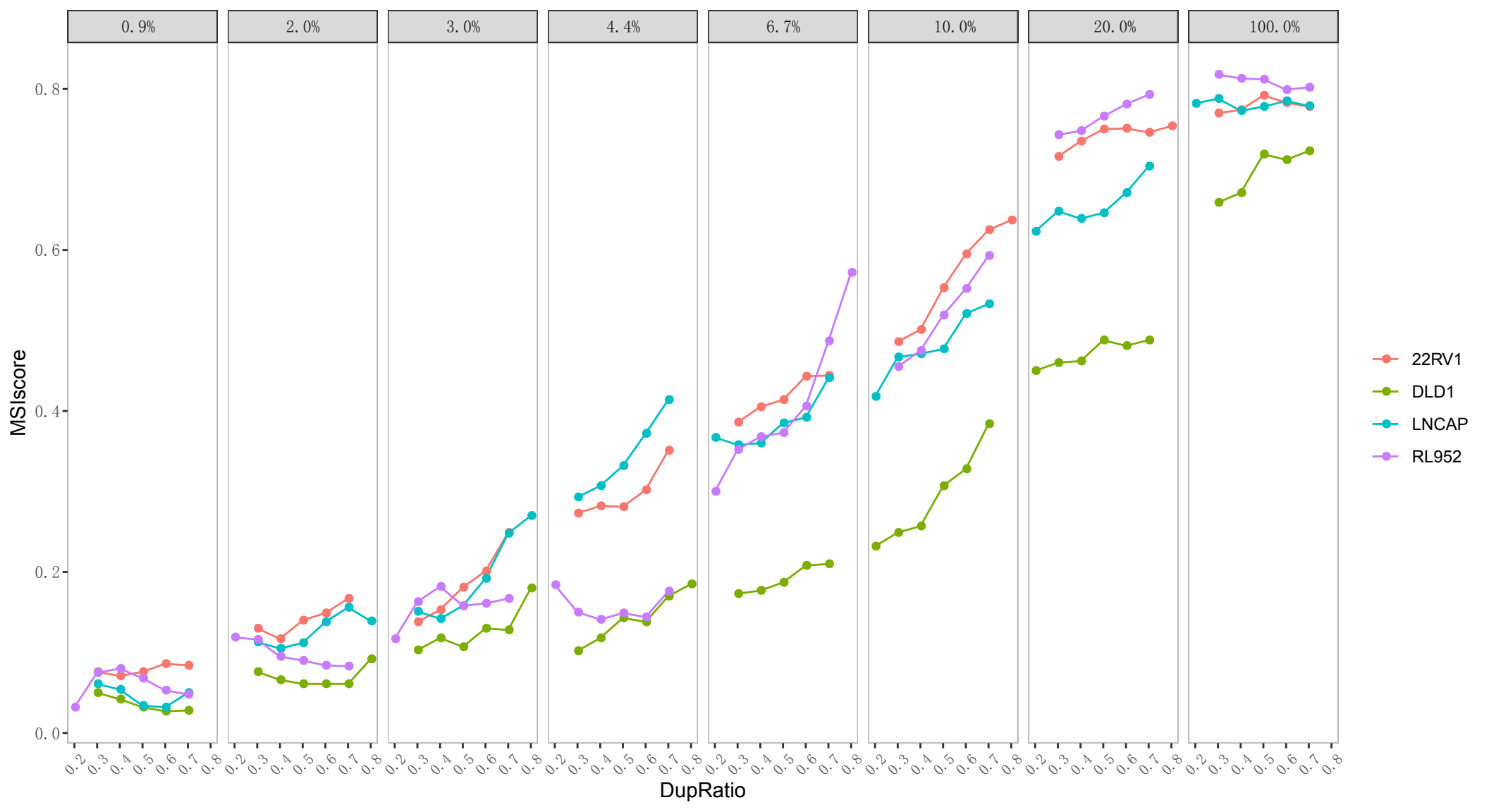

L10A-1
L11T-1
L12T-1
L14A-1
L15T-1
Range
1.00
0.75
0.50
0.25
0.00
0
50
100
150
200
250
300
350
400
450
500
Depth
KLD
DeletionRatio

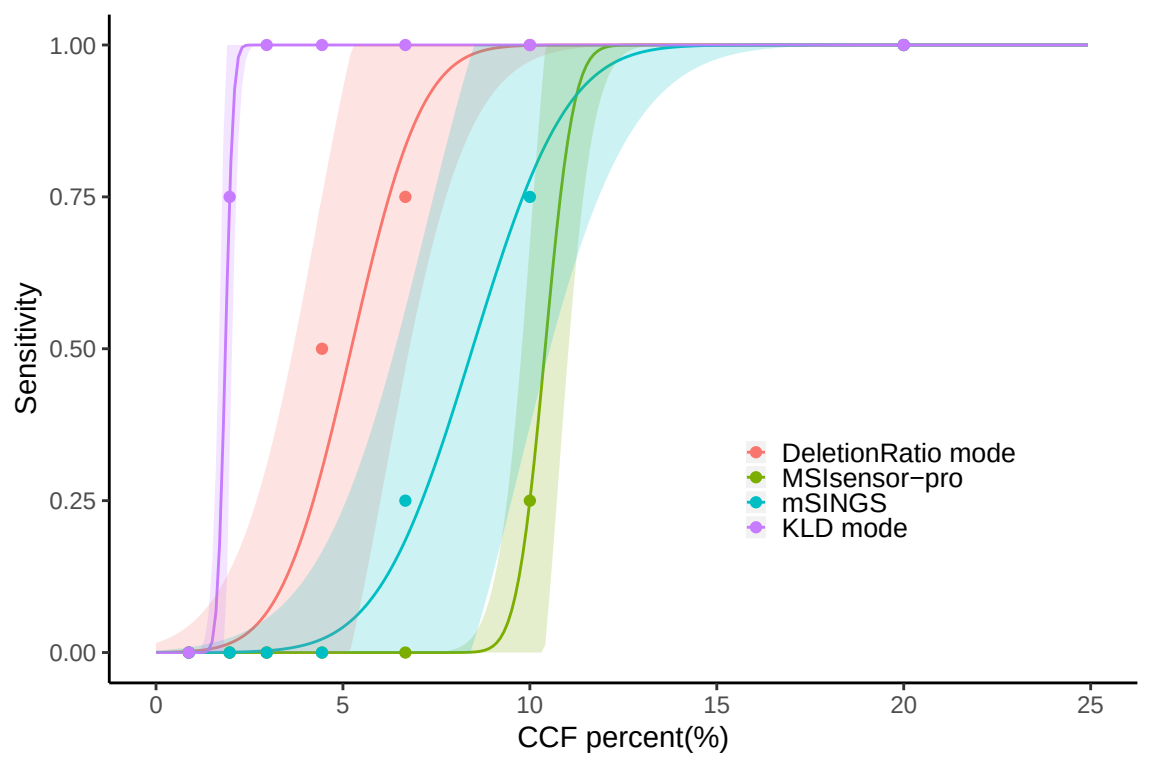
DeletionRatio

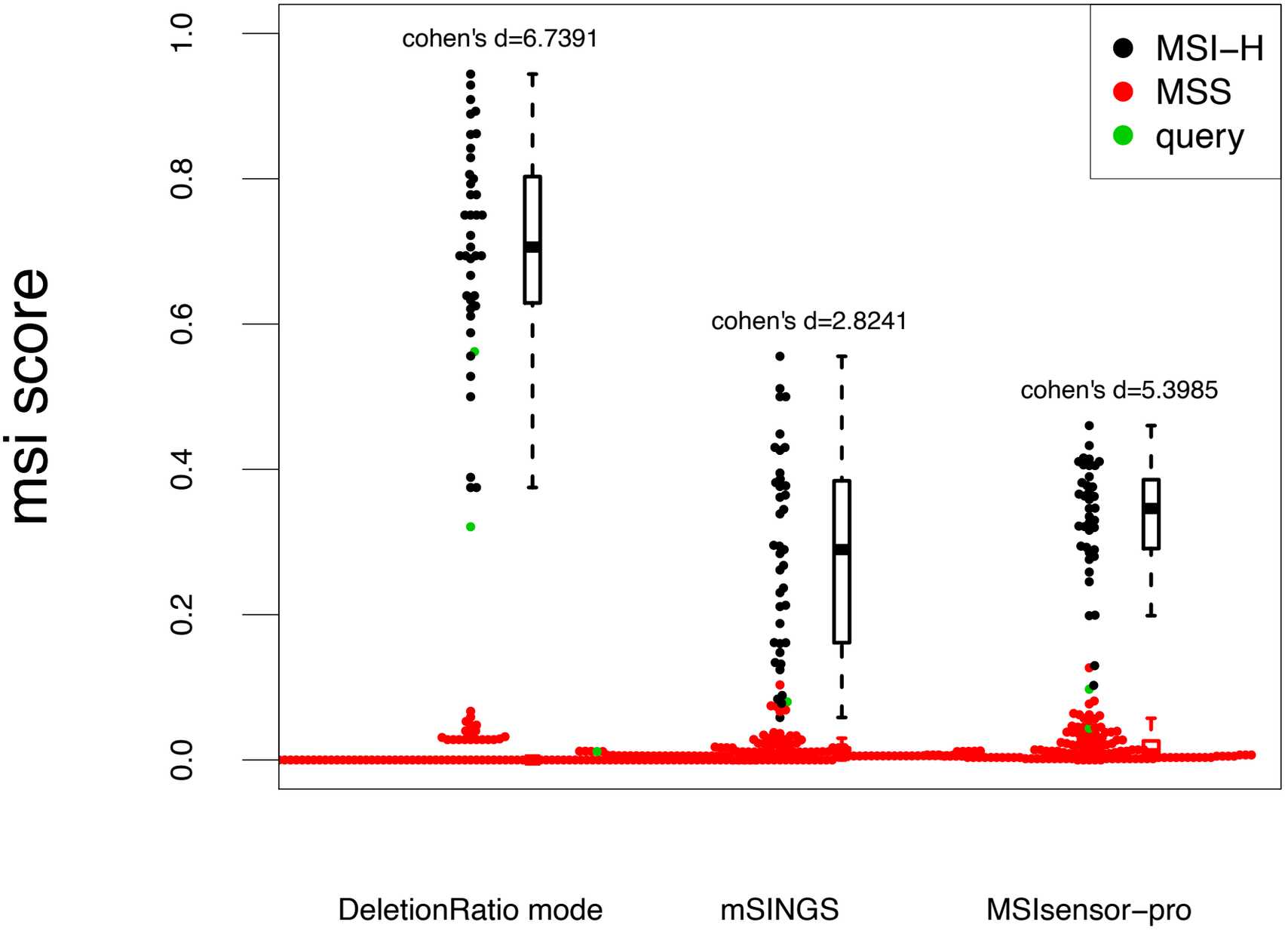


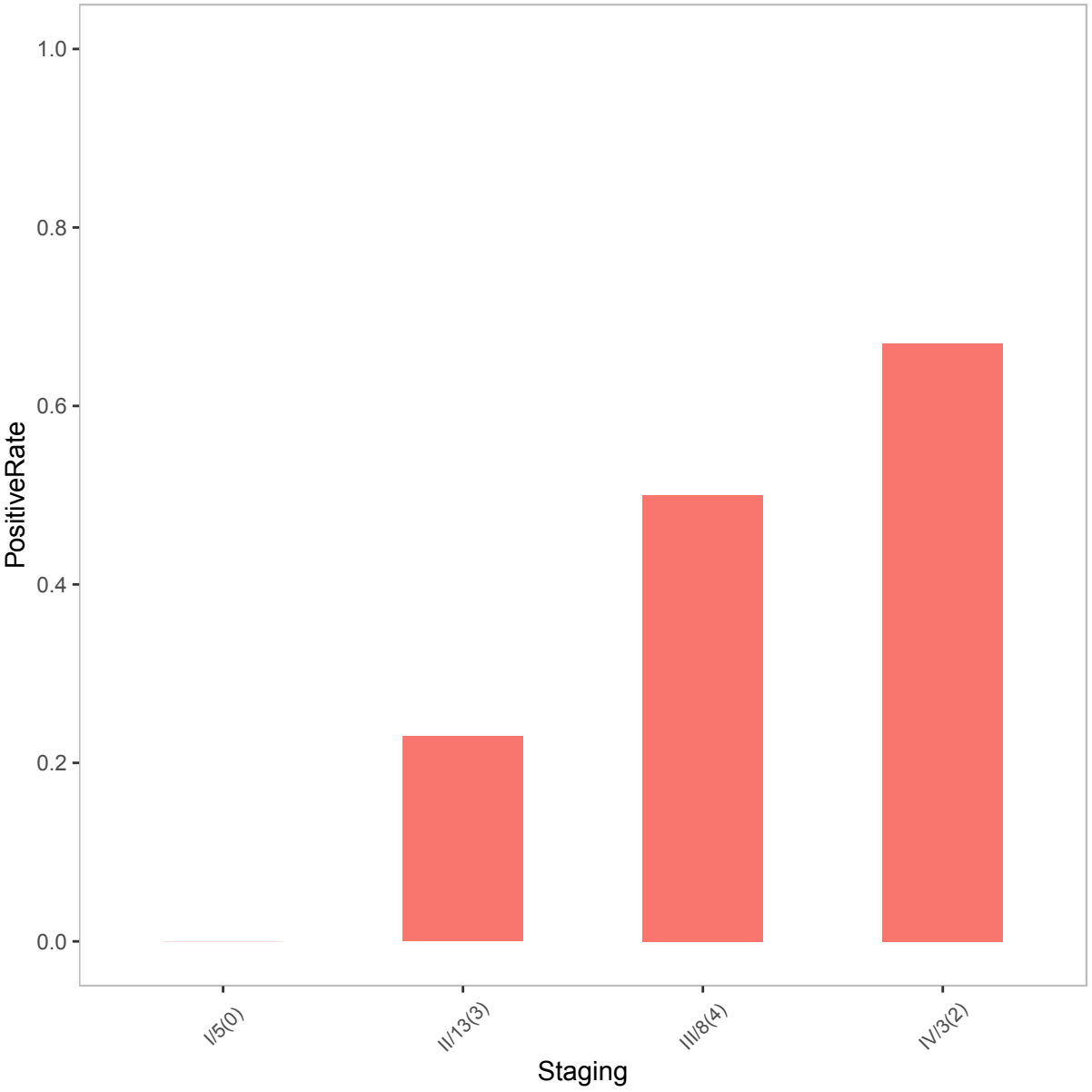
KLD score

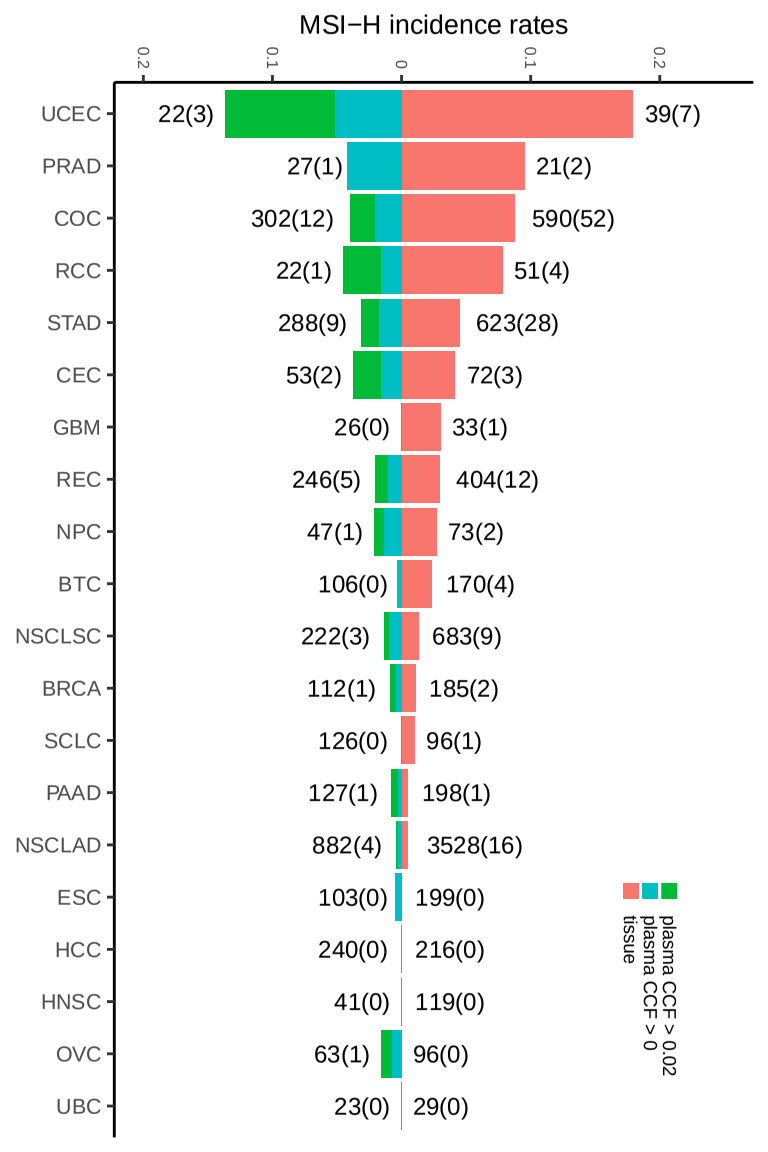



0.9%
2.0%
3.0%
4.4%
6.7%
10.0%
20.0%
100.0%
MSIscore
0.0
0.2
0.4
0.6
0.8
0.2
0.3
0.4
0.5
0.6
0.7
0.8
DupRatio
22RV1
DLD1
LNCAP
RL952


1.00
0.75
0.50
0.25
0.00
Sensitivity
0
5
10
15
20
25
CCF percent(%)
DeletionRatio mode
MSIsensor−pro
mSINGS
KLD mode


msi score
0.0
0.2
0.4
0.6
0.8
1.0
cohen's d=6.7391
cohen's d=2.8241
cohen's d=5.3985
MSI−H
MSS
query
DeletionRatio mode
mSINGS
MSIsensor−pro


1.0
0.8
0.6
0.4
0.2
0.0
PositiveRate
I/5(0)
II/13(3)
III/8(4)
IV/3(2)
Staging


MSI−H incidence rates
0.2
0.1
0
0.1
0.2
UCEC
PRAD
COC
RCC
STAD
CEC
GBM
REC
NPC
BTC
NSCLSC
BRCA
SCLC
PAAD
NSCLAD
ESC
HCC
HNSC
OVC
UBC
22(3)
39(7)
27(1)
21(2)
302(12)
590(52)
22(1)
51(4)
288(9)
623(28)
53(2)
72(3)
26(0)
33(1)
246(5)
404(12)
47(1)
73(2)
106(0)
170(4)
222(3)
683(9)
112(1)
185(2)
126(0)
96(1)
127(1)
198(1)
882(4)
3528(16)
103(0)
199(0)
240(0)
216(0)
41(0)
119(0)
63(1)
96(0)
23(0)
29(0)
plasma CCF > 0.02
plasma CCF > 0
tissue