ML Method for identifying microsatellite instability DNA samples without matched normal

Dilmi Perera ,PhD; Sahand Khakabi, PhD; Ka Mun Nip, MSc; Sonal Brahmbatt, MSc; Adrian Kense, MSc ; Kevin Tam, PhD; David Mulder, PhD; Melissa K. McConechy, PhD ; David Huntsman, MD PhD; Ruth Miller, PhD; Rosalía Aguirre-Hernández

Published Abstract Number: 4279

Abstract

Tissue samples with high microsatellite instability (MSI-H) can be indicative of tumors that are sensitive to certain types of cancer treatments (e.g., immune modulation-checkpoint inhibitor treatment). MSI-H regions can be identified with Polymerase Chain Reaction (PCR) based assays and next-generation sequencing (NGS). However, these MS regions are susceptible to PCR and sequencing errors. We developed a machine learning method (ML) for detecting microsatellite instability high (MSI-H) tumors using NGS data to accurately identify true MSI-H samples from MS-Stable samples based on an analysis of these MS regions. Furthermore, this method does not require a match normal or pool of normals.

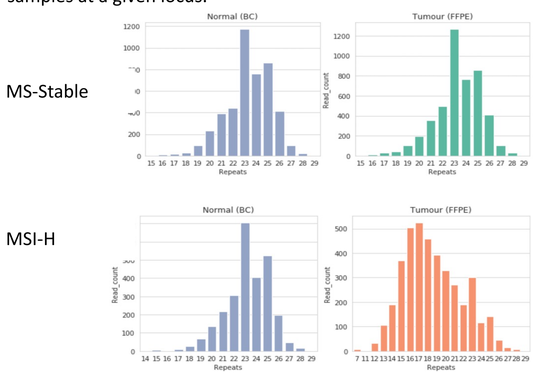
Introduction

A microsatellite (MS) is a tract of repetitive DNA in which certain DNA motifs are repeated:
GGACTCTCTCTCTCTCTCTCTCTATA

These regions are susceptible to DNA replication errors which are corrected by the DNA mismatch repair machinery. In DNA mismatch repair deficient tumours (MMRd/MSI-H) these regions have higher rates of DNA replication errors resulting in alleles with different numbers of repeats, which can be identified using NGS: GGACTCTCTCTCTCTCATA.

Several MSI callers use the distribution of repeats to classify MSS vs MSI-H, see Figure 1.

Figure 1. Length distribution of repeats in MS-Stable and MSI-H samples at a given locus.



Methods and Materials

We developed a proprietary method for classifying a tissue sample as being microsatellite instability high (MSI-H) without using normal tissue from the same person which doubles the sequencing cost, or a pool of normals which is sometimes challenging to build. Furthermore, the algorithm was designed for amplicon targeted assays where it is not always feasible to choose the most predictive MS sites. The MSI pipeline workflow is shown in Figure 3.

The machine learning classifier (ML) algorithm is a random forest algorithm with a training set of known MSI-H and MS-Stable samples to learn the relationship between the MSI status and the distribution of repeats in microsatellite regions of genomes using 21 MS loci. A negative control (NC) was used to normalize the ML features and therefore reduce the effects of PCR and sequencing errors in noisy MS sites, see Figure 2. Formally, the features are defined as the stepwise difference distance metric expressed by

$$d = \sum_{\gamma \epsilon (R_T \cup R_{NC})} |T_\gamma - NC_\gamma|$$

where
T_γ is the fraction of reads with repeats of length γ in the tumor sample at a given MSI locus

NC_γ is the fraction of reads with repeats of length γ in the NC sample at the same MSI locus.

Figure 2. Features normalized by negative control (NC).

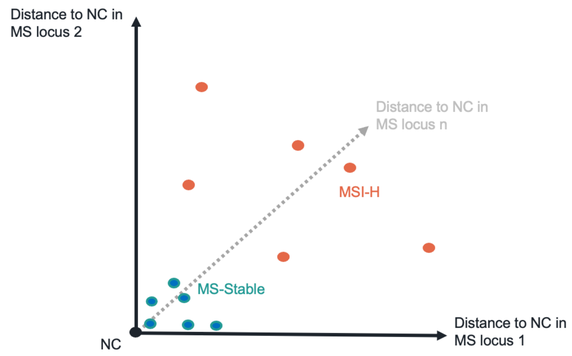
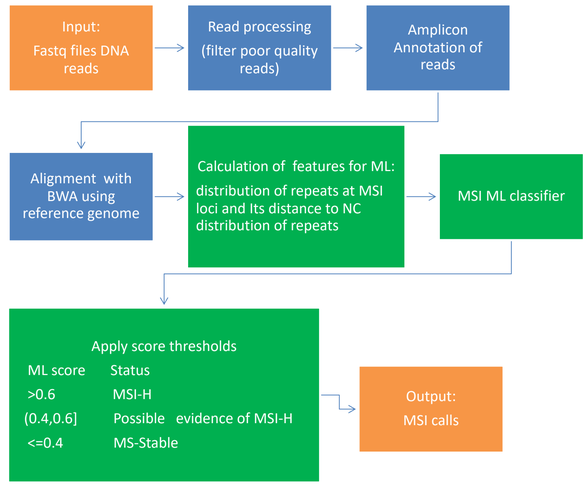


Figure 3. ICH MSI pipeline workflow.



Results

The MSI ML algorithm performance was compared with MSIsensor and MANTIS, see Figure 4.

The MSI detection algorithm was also validated in analytical and clinical samples (Figure 5) achieving an accuracy greater than 98%.

Analytical samples consist of commercial reference standard samples and well characterized FFPE treated cell-lines. Clinical samples consist of clinical FFPE tumor samples from cancer patients that were orthogonally validated using immunohistochemistry (expression of mismatch repair genes, i.e., MMRnormal vs MMRd) on tumor tissue and/or the Promega MSI PCR using matched tumor/normal.

All experiments were performed in the Imagia Canexia Health CAP, CLIA, DAP certified laboratory using the Find It® assay standard operating procedures for detecting genomic mutations in solid tumor tissue.

Figure 4. Performance comparison: MSIsensor, MANTIS and Imagia Canexia Health (ICH) MSI classifier.

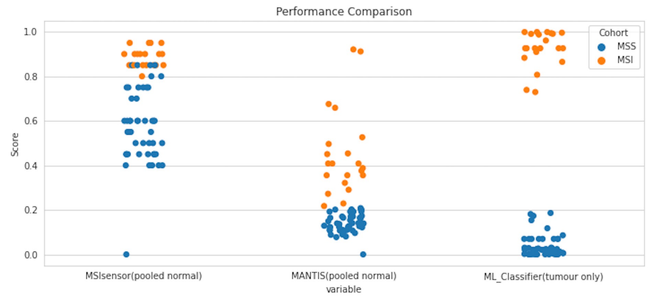
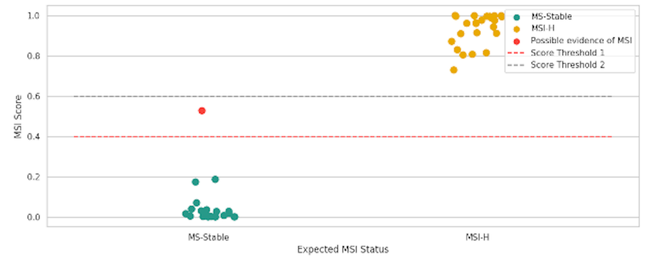


Figure 5. Performance of MSI ML classifier on clinical samples.



Conclusions

The ICH MSI detection algorithm can accurately identify samples with MSI-H tumors on amplicon based assays with limited number of predictive MSI sites. Furthermore, this algorithm does not require a pool of normal or a matched normal tissue which is not always available and doubles the sequencing costs.

When used in a clinical setting, patients with MSI-H samples can then be directed to treatments such as immune-checkpoint inhibitors.

Contact

Rosalía Aguirre-Hernández
Imagia Canexia Health
Email: rosalia.aguirre@imagiacanexia.com
Website: https://imagiacanexiahealth.com/about-canexia/

References

1. 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) 30:1015–6.10.1093/bioinformatics/btt755
2. 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) 8:7452–63. doi: 10.18632/oncotarget.13918
3. Salipante SJ, Scroggins SM, Hampel HL, Turner EH, Pritchard CC. Microsatellite instability detection by next generation sequencing. Clin Chem. (2014) 60:1192–9. doi:10.1373/clinchem.2014.223677
4. 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) 20:225–31. doi: 10.1016/j.jmoldx.2017.11.007