

# SCIENTIFIC REPORTS

OPEN

## MSIpred: a python package for tumor microsatellite instability classification from tumor mutation annotation data using a support vector machine

Chen Wang<sup>1</sup> & Chun Liang<sup>1,2</sup>

Microsatellite instability (MSI) is characterized by high degree of polymorphism in microsatellite lengths due to deficiency in mismatch repair (MMR) system. MSI is associated with several tumor types and its status can be considered as an important indicator for patient prognosis. Conventional clinical diagnosis of MSI examines PCR products of a panel of microsatellite markers using electrophoresis (MSI-PCR), which is laborious, costly, and time consuming. We developed MSIpred, a python package for automatic MSI classification using a machine learning technology – support vector machine (SVM). MSIpred computes 22 features characterizing tumor somatic mutational load from mutation data in mutation annotation format (MAF) generated from paired tumor-normal exome sequencing data, subsequently using these features to predict tumor MSI status with a SVM classifier trained by MAF data of 1074 tumors belonging to four types. Evaluation of MSIpred on an independent testing set, MAF data of another 358 tumors, achieved overall accuracy of  $\geq 98\%$  and area under receiver operating characteristic (ROC) curve of 0.967. Further analysis on discrepant cases revealed that discrepancies were partially due to misclassification of MSI-PCR. Additional testing of MSIpred on non-TCGA data also validated its good classification performance. These results indicated that MSIpred is a robust pan-tumor MSI classification tool and can serve as a complementary diagnostic to MSI-PCR in MSI diagnosis.

Microsatellites are tandemly repeated sequences with typical repeat unit length varying from 1 to 6 bases<sup>1</sup>. Slippage events during DNA replication can lead to gain or loss of repeat units from microsatellite loci throughout genome. Under normal circumstances, these spontaneous mutations can be sensed and corrected by mismatch repair (MMR) system. When MMR system is inactivated or affected due to mutations or epigenetic silencing in its associated genes, including *MSH2*, *MSH3*, *MSH6*, *MLH1*, *MLH3*, *PMS1*, *PMS2*, consequent high degree of polymorphism in microsatellite length, defined as microsatellite instability (MSI), frequently occurs<sup>3</sup>. Clinical diagnosis of MSI is usually achieved by examining lengths of PCR products of five informative microsatellite loci (BAT25, BAT26, D2S123, D5S346, and D17S250) of National Cancer Institute microsatellite panel by capillary electrophoresis (known as MSI-PCR method). According to Bethesda guideline<sup>4</sup>, tumors are termed as microsatellite stable (MSS) if none of five loci tested is mutated, and tumors are termed as microsatellite instability low (MSI-L) if only one tested locus is mutated and microsatellite instability high (MSI-H) if two or more tested loci are mutated. MSI is best known in colorectal tumor where clear clinical implications have been established. 15% cases of colorectal tumors are observed with MSI. 3% of them are inheritable and are defined as Lynch syndrome where inactivating germline mutations in genes of MMR system are detected. The other 12% are sporadic cases caused by somatic mutations in genes of MMR system<sup>5,6</sup>. MSI-H colorectal tumors tend to have better prognosis and are less prone to metastasis<sup>7</sup>. Though less understood, MSI is also reported in endometrial, gastric, glioblastoma and prostate tumors<sup>8,9</sup>. Classification of MSI is of great importance due to its significant relationship with therapeutic decisions, but clinical diagnosis of MSI by MSI-PCR is laborious, costly, and time consuming. With

<sup>1</sup>Department of Biology, Miami University, Oxford, OH, 45056, USA. <sup>2</sup>Department of Computer Science & Software Engineering, Miami University, Oxford, OH, 45056, USA. Correspondence and requests for materials should be addressed to C.L. (email: [liangc@miamioh.edu](mailto:liangc@miamioh.edu))

Project	Training	Testing
COAD	278	99
READ	92	39
STAD	334	103
UCEC	370	117
Total	1074	358

**Table 1.** Composition of training and testing data sets according to tumor types.

MSI status	Training	Testing
MSS*	705	227
MSI-L*	138	53
MSI-H	231	78
Total	1074	358

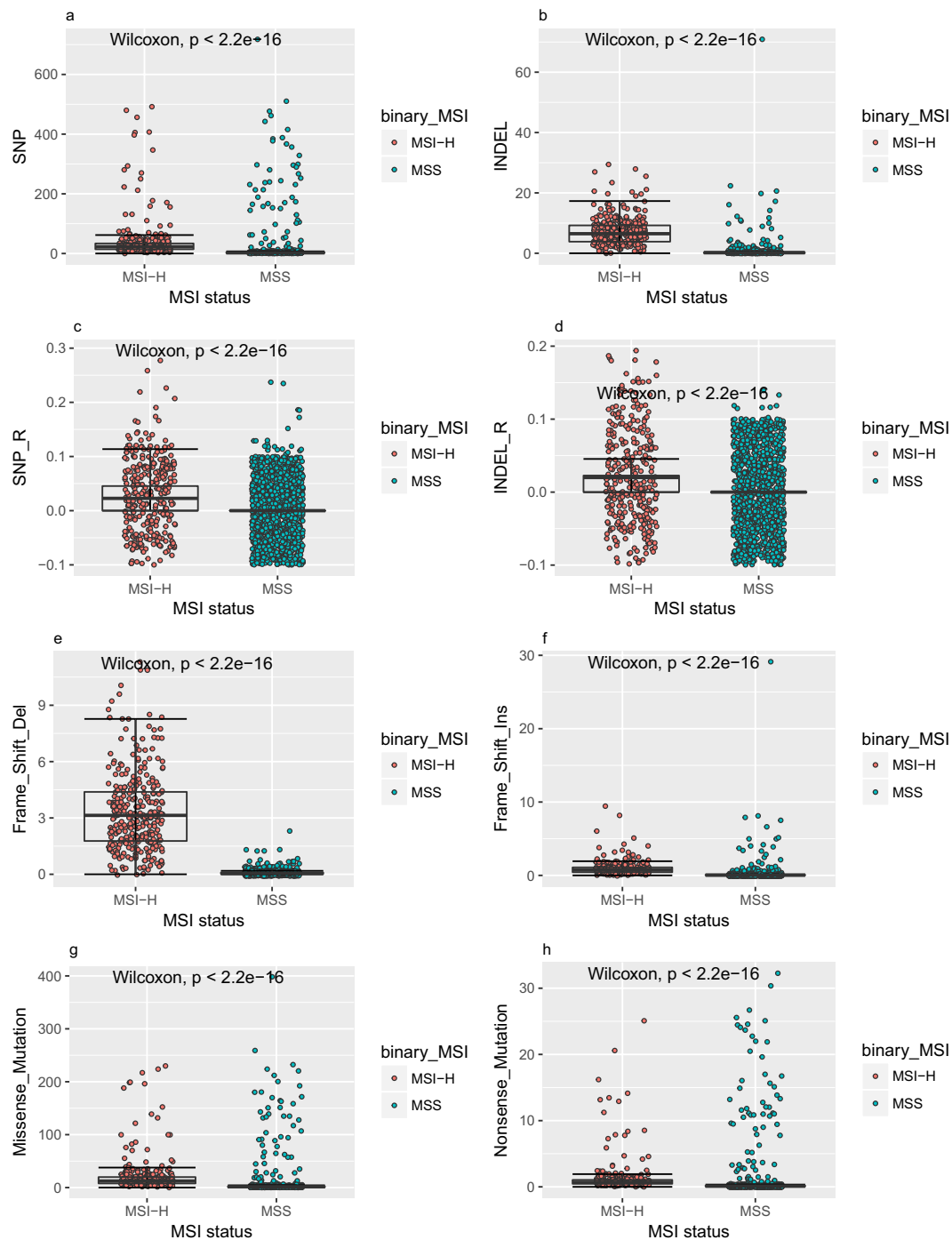
**Table 2.** Composition of training and testing data sets according to MSI-PCR determined MSI status. \*Both MSS and MSI-L are labelled as MSS for binary classification in our study.

the development of next generation sequencing (NGS) technology, several computational tools utilizing NGS data were developed for MSI diagnosis. MSIsensor<sup>10</sup> and mSINGS<sup>11</sup> determine tumor MSI status by measuring prevalence of unstable microsatellite loci in paired tumor-normal sequencing data. These two tools showed good performance but they both require lots of computational resources since they directly examine aligned reads in BAM format. MSIsseq, MOSAIC, and MIRMMR<sup>12–14</sup> utilize machine learning algorithms to predict MSI status. MSIsseq and MOSAIC implemented decision tree classifiers for MSI classification depending on just a single feature derived from tumor somatic mutation information, which is prone to over-fit their training datasets. MIRMMR built a logistic regression classifier requiring both methylation and mutation information of genes belonging to MMR system. A recent study revealed that, in addition to MSI, MMR-deficient tumors also display relatively higher somatic mutational load<sup>15</sup>. Thus, it is a reasonable direction to predict tumor MSI status through tumor mutation information obtained by exome sequencing. Recently, mutation calling has become a routine analysis of paired tumor-normal sequencing data. Several different pipelines for mutation calling have been published<sup>16–21</sup>. These pipelines typically create a file in mutation annotation format (MAF). MAF files provide somatic mutation information of tumors including single nucleotide mutations, micro-indels, and their detailed translational/regulatory effects (for example, missense, nonsense, and silence mutations) on annotated genes. In this study, we explored distributions of somatic mutational load in MSI-PCR termed MSI-H and non MSI-H (MSS and MSI-L) tumors using their corresponding MAF files obtained from The Cancer Genome Atlas (TCGA) Research Network<sup>22</sup>, and subsequently developed a python package, MSIPred, implementing a pan-tumor binary MSI classifier to predict MSI status from tumor MAF files.

## Results

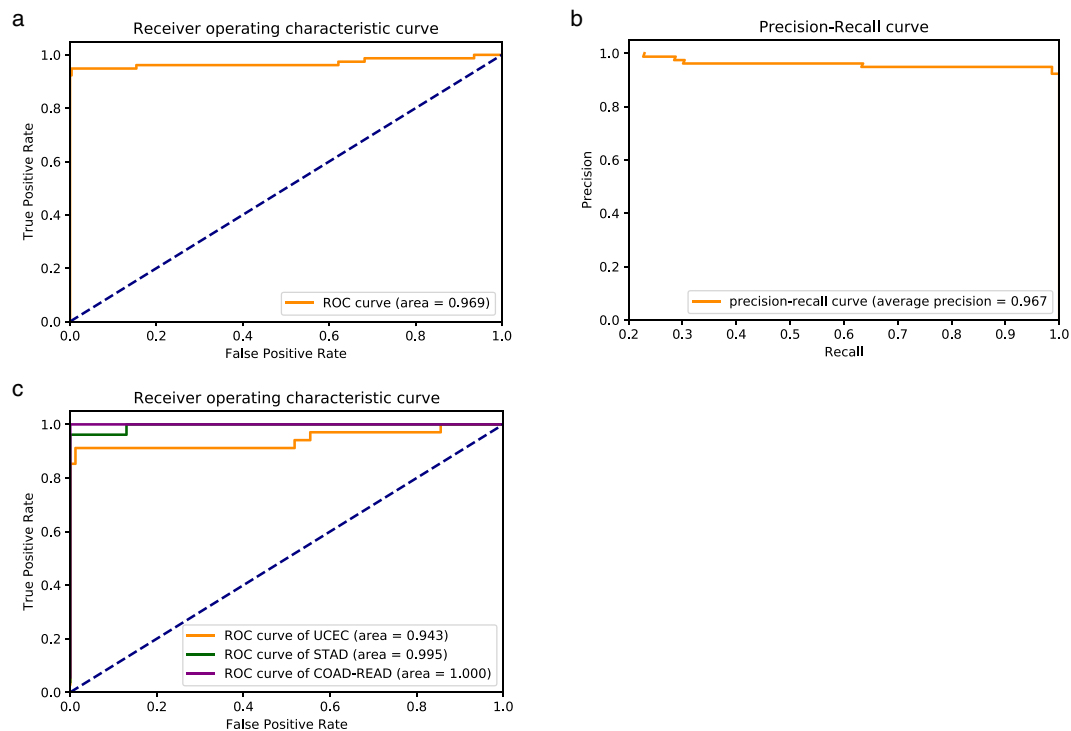
**Distributions of somatic mutational load in MSS and MSI-H tumors.** We calculated 22 features characterizing somatic mutational load of 1432 tumors of four types (see Table 1, COAD: colon adenocarcinoma, READ: rectal adenocarcinoma, STAD: stomach adenocarcinoma, and UCEC: uterine corpus endometrial carcinoma) using mutation data from tumor MAF files. Of all 22 features, 1–9, previously used by MSIsseq<sup>12</sup>, describe a general type (single nucleotide variants or micro-indels) of a mutation. 10–22 describe detailed classification of a mutation based on its translational and/or regulatory effects, which can provide critical information about how deleterious a mutation is (see Supplementary Table S1). All of these 22 features displayed skewed distributions in total 1432 tumors (see Supplementary Fig. S1). As a recent study revealed that MSI-L tumors were consistent with MSS tumors in terms of their MSI burden<sup>13</sup>, we combined MSS and MSI-L tumors together, and denoted them as MSS in this study. Referring to tumor clinical data provided by TCGA, 1123 of 1432 tumors were determined as MSS using MSI-PCR, the other 309 tumors were determined as MSI-H (see Table 2). MSI-H tumors displayed significantly higher counts of mutations (both single nucleotide variants and micro-indels) per megabase than MSS tumors throughout exome sequences and in simple sequence repeat regions. MSI-H tumors also tended to possess more deleterious mutations, such as mutations causing shifts in open reading frames, missense mutations, and nonsense mutations (Fig. 1). By pairing these 22 features, a clear separation between MSS tumors and MSI-H tumors can be observed (see Supplementary Fig. S2). Wilcoxon rank-sum tests of each of these 22 features between MSS and MSI-H tumors were performed, and results indicated that all these 22 features were differently distributed between MSS and MSI-H tumors.

**Analysis of 22 features used for mutational load characterization.** 1432 tumors were then randomly split into two datasets, training and testing, using 3:1 ratio within each tumor type. Accordingly, the training set incorporated 1074 tumors of all four types, and the testing set incorporated 358 tumors of four types. Detailed information about these two datasets were presented in Tables 1 and 2. In order to gain some insights about the contributions of these features to MSI classification, we first trained a random forest classifier using all 22 features of 1074 tumors belonging to the training set. The importance scores of all these 22 features were then



**Figure 1.** MSS and MSI-H tumors possessed different somatic mutational load. Boxplots for eight somatic mutational load features of MSS (a combination of MSI-L and MSS) tumors and MSI-H tumors were given. X axis of each subplot denotes two MSI status and Y axis denotes counts (per megabase (Mb)) of a mutational load feature. P-values obtained by Wilcoxon rank sum test were labeled on each subplot. From (a–h), each subplot denotes following features: SNP, INDEL, SNP\_R, INDEL\_R, Frame\_Shift\_Del, Frame\_Shift\_Ins, Missense\_Mutation, and Nonsense\_Mutation, respectively.

obtained (see Supplementary Table S2) from the random forest classifier to evaluate contributions of these features<sup>23</sup>. According to their importance scores, those 13 new features other than the 9 features selected by MSIseq<sup>12</sup> greatly impact the classification results, especially Frame\_Shift\_Del and Frame\_Shift\_Ins. Besides, all 22 features contributed to the classification of MSS and MSI-H tumors in terms of their importance scores obtained from the random forest classifier. Thus, we decided to keep all 22 features for further implementation.



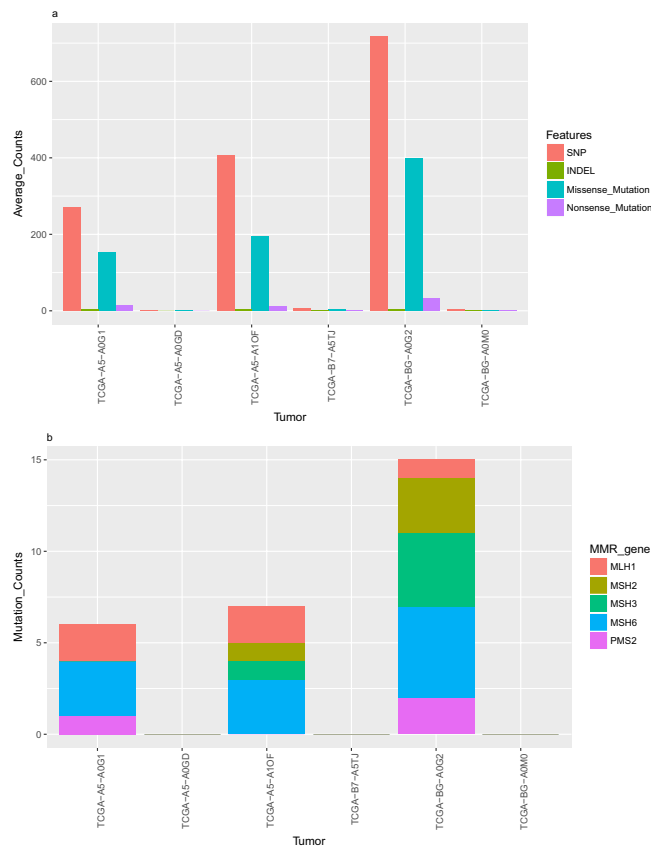
**Figure 2.** Performance of MSIpred. **(a)** Receiver operating characteristic (ROC) curve of MSIpred. The blue dash line denotes a line with slope of 1 and intercept of 0. **(b)** Precision-recall curve of MSIpred. Both ROC curve and precision-recall curve were obtained by applying MSIpred on the 358-tumor testing set. **(c)** ROC curves of MSIpred for different tumor types. Endometrial tumors (UCEC): orange, 117 tumors; stomach tumors (STAD): green, 103 tumors; colorectal tumors (COAD-READ): purple, 138 tumors.

**Classification of MSI using Support Vector Machine.** We then developed a pan-tumor support vector machine (SVM) classifier with a radial basis function kernel for MSI classification using aforementioned 22 features of all tumors from the training set. After the SVM classifier was trained, we released a python package called MSIpred (Microsatellite Instability Predictor) implementing that SVM classifier and a bioinformatics pipeline to compute those required features from MAF files to perform MSI auto classification.

**MSIpred is accurate for MSI classification.** We applied MSIpred to that 358-tumor testing set and evaluated its performance by finding concordances between MSIpred predicted MSI status and MSI-PCR determined MSI status provided by TCGA, using several different metrics. In this study, we denoted MSS tumors as negative samples while MSI-H tumors as positive ones. MSIpred finally achieved a sensitivity (recall) of 0.936, specificity of 0.996, precision of 0.986, G mean of 0.966, F1 score of 0.961, and overall accuracy of 0.983. ROC curve (Fig. 2a) of MSIpred achieved area under curve of 0.969. Precision-recall curve (Fig. 2b) of MSIpred achieved average precision of 0.967. These results indicated that MSIpred is a robust tool for MSI classification, and revealed that MSS tumors and MSI-H tumors possessed distinct somatic mutational loads.

**MSIpred is robust across different tumor types.** MSIpred was designed to perform pan-tumor MSI classification since it didn't incorporate tumor type as a feature when predicting MSI status. It is reasonable to evaluate performance of MSIpred on each tumor type individually since different types of tumors may hold different signatures of somatic mutational load. Hence, we then evaluated performance of MSIpred on each tumor type individually using the 358-tumor (UCEC: 117 tumors; STAD 103 tumors; COAD and READ :138 tumors) testing set. ROC curves for endometrial, stomach, and colorectal tumors were plotted (Fig. 2c). For colorectal tumors and stomach tumors, MSIpred achieved area under curve of 1.0 and 0.995, respectively. Area under curve for endometrial tumor is slightly lower with a value of 0.943. These results indicated that MSIpred didn't over-fit a particular tumor type and MSIpred was robust and generic on different tumor types.

**Analysis of discrepant cases.** We further investigated somatic mutational load (characterized by aforementioned 22 features) for those tumors whose MSIpred predicted MSI status and MSI-PCR determined MSI status were discrepant. In the 358-tumor testing set, 5 (4 endometrial, 1 stomach) of 78 MSI-PCR determined MSI-H tumors were classified as MSS by MSIpred (false negative). 1 (endometrial) of 280 MSI-PCR determined MSS tumors was classified as MSI-H (false positive) by MSIpred. As shown in Fig. 3a, three (TCGA-B7-A5TJ, TCGA-A5-A0GD, TCGA-BG-A0M0) of five false negative tumors possessed quite low somatic mutational load, indicating their MMR system might still be intact. The other two false negative tumors possessed high somatic mutational load. The only false positive tumor (TCGA-BG-A0G2, see Fig. 3a) also possessed relatively high



**Figure 3.** Analysis of discrepant cases. **(a)** Bar plot of four features (SNP, INDEL, Missense\_Mutation, Nonsense\_Mutations) characterizing somatic mutational load of six discrepant tumors **(b)** Mutation profiles of six discrepant tumors in terms of genes in MMR system.

somatic mutational load, indicating that its MMR system might have been inactivated or affected. We then analyzed their corresponding MAF files with special foci on five genes involved in MMR system including *MLH1*, *MSH2*, *MSH3*, *MSH6* and *PMS2*. For those three false negative tumors (TCGA-B7-A5TJ, TCGA-A5-A0GD, TCGA-BG-A0M0) with low somatic mutational load, no mutation was found in these five MMR genes. For the other two false negative tumors with high somatic mutation load, mutations were detected in MMR genes: TCGA-A5-A1OF possessed mutations in four (*MLH1*, *MSH2*, *MSH3*, *MSH6*) of five MMR genes; TCGA-A5-A0G1 possessed mutations in three (*MLH1*, *MSH6*, *PMS2*) of five MMR genes. The only false positive tumor, TCGA-BG-A0G2, possessed mutations in all five MMR genes (Fig. 3b). This result suggests that the MSI-PCR results might not be very accurate for these cases.

**Validation of MSIPred using non-TCGA MAF data.** As both of the training set and testing set data were obtained from TCGA, the performance of MSIPred on the aforementioned testing set might be overoptimistic. We further decided to validate the performance of MSIPred with another 390-tumor testing set whose data were generated by non-TCGA projects. This 390-tumor testing set consisted of MAF files of 368 colorectal tumors<sup>24</sup> and 22 stomach tumors<sup>25</sup>. These MAF files and their corresponding binary (cBioPortal directly provided binary MSI status) MSI status determined by MSI-PCR were all obtained from cBioPortal<sup>26,27</sup>. 90 of 390 tumors were determined as MSI-H and the rest 300 were determined as MSS by MSI-PCR. With this non-TCGA testing set, MSIPred achieved a sensitivity (recall) of 0.778, specificity of 0.990, precision of 0.959, G mean of 0.877, F1 score of 0.859, and overall accuracy of 0.941. Compared to the performance of MSIPred on the previous 358-tumor TCGA testing data, the sensitivity of MSIPred on non-TCGA testing data was lower, but still reached a good overall accuracy.

**Comparison of MSIPred with MSIsq.** Among previously published software tools, MSIsq<sup>12</sup> is the only one that also predicts MSI status solely from tumor MAF files. Our tool MSIPred also utilized nine mutation features that were first proposed by MSIsq. Therefore, we compared performances of MSIsq with MSIPred in this study. The 390-tumor non-TCGA testing set that was utilized to validate performance of MSIPred was also used to evaluate MSIsq with its default decision tree classifier. MSIsq gave overall accuracy of 0.918, sensitivity of 0.667, precision of 0.968 and specificity of 0.993 (see Table 3). These results revealed that MSIsq was good at recognizing negative samples (MSS) but poor at recognizing positive samples (MSI-H). Thus, MSIPred achieved higher sensitivity and accuracy than MSIsq.

Tools	non-TCGA testing set size							
	MSS	MSI-H	Se (recall)	Sp	Pre	G mean	F1	ACC
MSIpred (22 features)	300	90	0.778	0.990	0.959	0.877	0.859	0.941
MSIpred (9 features)			0.622	0.997	0.982	0.787	0.762	0.910
MSIseq			0.667	0.993	0.968	0.814	0.789	0.918

**Table 3.** Performances of original MSIpred, modified MSIpred, and MSIseq on 390-tumor non-TCGA testing set.

**13 new features improved performance of MSIpred.** Of all 22 features utilized by MSIpred, 13 (see Supplementary Table S1, row 10–22) were new features for MSI classification. These 13 features characterized detailed translational or regulatory effects of mutations. We validated the positive impacts of these 13 new features on MSI classification by comparing performances of two different versions of MSIpred on the aforementioned 390-tumor non-TCGA testing set: the original MSIpred that utilized all 22 (9 inherited from MSIseq plus 13 new features) features and a modified MSIpred that only utilized 9 features proposed by MSIseq (see Supplementary Table S1, row 1–9). Results showed that by adding 13 new features, performance of MSIpred was improved (see Table 3). The sensitivity of MSIpred was improved from 0.667 to 0.778 and overall accuracy was improved from 0.910 to 0.941.

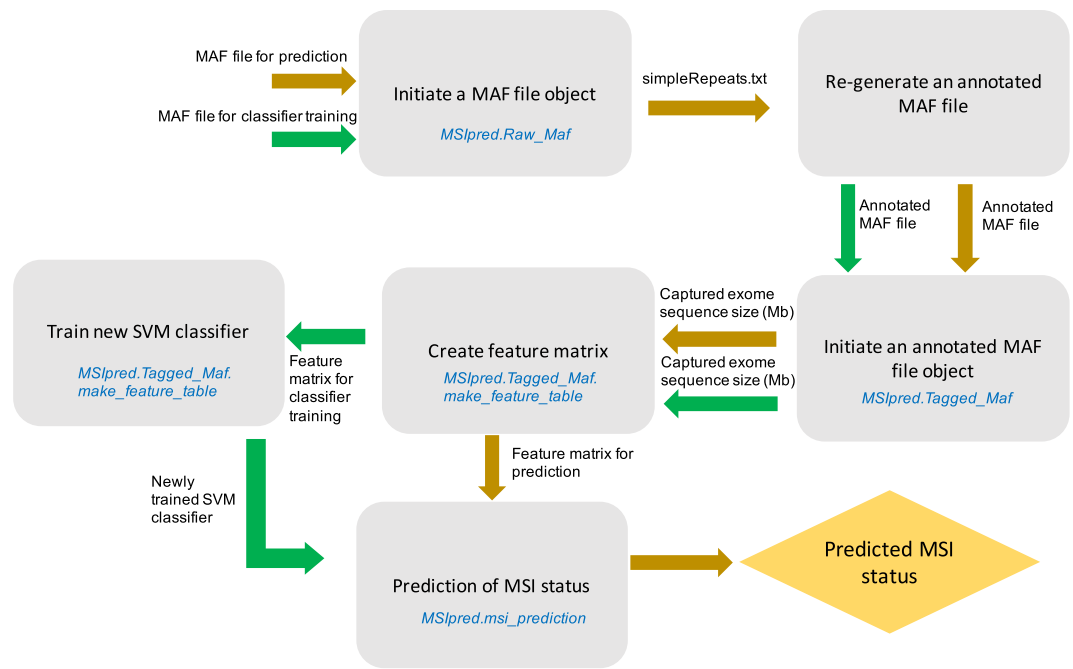
## Discussion

We utilized 22 features, which can be computed from tumor MAF files, to characterize somatic mutational load of tumors by considering both mutation types and their detailed translational/regulatory effects. All these 22 features were differentially distributed in MSI-PCR determined MSS and MSI-H tumors, indicating that these features could serve as useful predictors for MSI classification. Further analysis on these 22 features using a random forest classifier revealed that all these 22 features contributed to the classification of MSS and MSI-H tumors. We then developed a python package called MSIpred, which implemented a SVM classifier trained by those 22 features of a 1074-tumor training set, and a bioinformatics pipeline to compute those required features from MAF files to perform automatic binary MSI classification. If more data is available for training, MSIpred allows redo of the training step in order to get a better classifier. Those MAF files that MSIpred utilized as inputs were generated from paired tumor-normal exome sequencing data. An evaluation of MSIpred by finding concordances with MSI-PCR determined MSI status on a 358-tumor testing set showed that MSIpred is a robust pan-tumor MSI classification tool with high accuracy, sensitivity, and specificity.

Using MSI-PCR determined MSI status as the experimental reference, 6 discrepancies including five false negative cases and one false positive case were observed in MSIpred classification results. Further analysis of their somatic mutational load and mutation profiles in terms of genes belonging to MMR system revealed that three of five false negative tumors were likely to possess intact MMR systems, while the other two false negatives were likely to possess inactivated or insufficient MMR systems. The only false positive case was also likely to possess an inactivated or insufficient MMR system. These results all together revealed that some of the discrepancies may be due to misclassification of MSI-PCR method when considering tumor mutation profiles (see Fig. 3). For those three false negatives with intact MMR systems and the only false positive case, their MSIpred predicted MSI status were more persuasive while for the rest two false negative cases, MSI-PCR termed MSI status were more likely to be correct. Thus MSIpred can serve as a reliable complementary tool for MSI diagnosis beyond the conventional MSI-PCR method. Besides, as MSIpred only requires MAF files for MSI classification, it enables researchers to better utilize NGS data from public domain.

Several computational tools for MSI classification based on NGS data were published. MSIsensor<sup>10</sup> and mSI-NGS<sup>11</sup> both examine aligned reads in “BAM” format, and try to derive MSI status by measuring prevalence of unstable microsatellite loci throughout genome. MOSAIC<sup>13</sup> is another software that utilizes sequencing data in “BAM” format for MSI classification. However, MOSAIC doesn’t directly compare distributions of unstable microsatellite loci in paired tumor-normal data, instead, it implemented a weighted tree classifier to perform MSI classification using average gain of novel microsatellite alleles. BAM files are often very large and need large amount of computational resources to manipulate. Mutation calling has become a routine analysis for NGS data and yields tumor mutation data in MAF format. MSIseq<sup>12</sup> and MIRMMR<sup>14</sup>, two MSI classification tools that either solely or partially utilize MAF data for MSI classification have been released. MSIseq implemented a decision tree classifier using features computed from MAF files as well as a feature of tumor types, which would potentially limit its application to some rare tumors. In addition, MSIseq only makes use of general types of mutations (single nucleotide variants or micro-indels) without considering their translational and regulatory effects. Unlike MSIseq, which only requires MAF data, MIRMMR also requires additional tumor methylation data to predict MSI status by a logistic regression classifier. In our study, MSIpred incorporated 9 of 10 features (eliminated tumor type) proposed by MSIseq, and added another 13 features that take potential translational and regulatory effects of mutations into account. By applying MSIpred and MSIseq to another 390-tumor non-TCGA testing set, the good performance of MSIpred was further validated, though a decline in its sensitivity was observed. This underestimation of MSI using non-TCGA data could potentially be explained by that MMR genes can also be repressed by epigenetic mechanisms. Apart from that, MSIpred did not consider individual microsatellite locus when doing classification, but individual microsatellite locus might be informative in terms of MSI classification. Compared to MSIseq, MSIpred still reached a better sensitivity and overall accuracy. A modified MSIpred that only utilized 9 (the features previously utilized by MSIseq) of 22 features was also applied to the same non-TCGA





**Figure 4.** Workflow for MSIpred. MSIpred provides a python class of MAF file (*MSIpred.Raw\_Maf*), which requires a tumor MAF file to create a MAF file object. After a MAF file object has been created, a method (*Raw\_Maf.create\_tagged\_maf*) associated with this object will re-generate an annotated MAF file by adding one extra column called “In\_repeats”, which indicates whether mutation events happen in simple repeats region or not, given a reference file indicating loci of simple repeats regions throughout genome (*simpleRepeats.txt* file). Then, a class of annotated MAF file provided by MSIpred (*MSIpred.Tagged\_Maf*) takes the aforementioned annotated MAF file to create an object of annotated MAF file. A method (*Tagged\_Maf.make\_feature\_table*) associated with annotated MAF file object takes the size (Mb) of captured exome sequences used for exome sequencing to calculate 22 features for somatic mutational load characterization of all tumors embedded in the very first MAF file, and returns a feature matrix as a pandas dataframe. Finally, *MSIpred.msi\_prediction*, a function provided by MSIpred, takes that feature matrix, and gives the final predicted MSI status for all tumors with the help of an implemented SVM classifier. MSIpred also allows users to train their own SVM classifier by a function called *MSIpred.svm\_training* using newly obtained MAF data and known MSI status of tumors. The newly trained classifier can be utilized in *MSIpred.msi\_prediction* function for MSI prediction and classification.

testing set, resulting a lower sensitivity and overall accuracy. It implied the positive informative effects of those 13 new features in MSI classification.

We finally conclude that MSIpred is a robust tool for pan-tumor MSI classification from tumor MAF data, and is available as a python 2 package. We anticipate that MSIpred will have a wide usage in MSI clinical diagnosis.

## Methods

**Input and output.** MSIpred requires MAF files containing tumor somatic mutation annotation information derived from paired tumor-normal whole exome sequencing data, a reference file indicating loci of simple sequence repeats throughout genome, and length (Mb) of overall captured sequences used for exome sequencing. MSIpred then computes 22 features characterizing somatic mutational loads of tumors with the implemented bioinformatics pipeline. Subsequently, these features are utilized for prediction of MSI status. The final output of MSIpred is a pandas dataframe containing predicted MSI status of tumors.

**Implementation.** MSIpred is written and tested under python 2 (version 2.7.12) and is freely available as a python package. It requires pandas (version 0.20.3)<sup>28</sup>, intervaltree (version 2.1.0)<sup>29</sup>, and scikit-learn (0.19.1)<sup>30</sup> packages to work properly. First, a MAF file is annotated by adding an additional column “In\_repeats”, indicating whether mutation events took place in simple sequence repeats region or not. The annotated MAF file generated by last step is then utilized to extract 22 features characterizing somatic mutational load for all tumors embedded in this annotated MAF file. These 22 features of all tumors are then combined into a feature matrix as a pandas dataframe. Finally, MSIpred calls the implemented SVM classifier to accurately predict tumor MSI status using the aforementioned feature matrix, and returns results as a pandas dataframe. MSIpred also allows users to train new SVM classifier given new MAF data and known MSI statuses of tumors. Subsequently, the newly trained SVM classifier can be utilized for MSI prediction. A general workflow of MSIpred is presented in Fig. 4.

**Somatic mutation data and clinical MSI status of tumors.** MAF files of somatic mutation data created by MuTect mutation calling pipeline were collected from four projects (COAD: colon adenocarcinoma,

READ: rectal adenocarcinoma, STAD: stomach adenocarcinoma, and UCEC: uterine corpus endometrial carcinoma) of TCGA<sup>31–33</sup>.

MSI-PCR determined MSI status of all tumors were collected from tumor clinical data provided by TCGA. To retrieve these data from TCGA, a R package called TCGAbiolinks (version 2.5.12)<sup>34</sup> was used for data query in this study. In total, MAF files together with MSI-PCR determined MSI status of 1432 tumors (COAD: 377, READ: 131, STAD: 437, UCEC: 487) were used in this study.

**Feature selection.** We incorporated 9 features (see Supplementary Table S1, row 1–9) proposed by MSIseq<sup>12</sup>. These nine features were calculated based on information of ‘Variant\_Type’ columns of MAF files. These nine features take account mutations in simple sequence repeats region specially, thus we need a reference file for loci of simple sequence repeats region in human genome. We retrieved loci of simple sequence repeats for GRCh38 (Genome Reference Consortium Human Reference 38) from a table called “SimpleRepeats” in UCSC genome annotation database (<http://hgdownload.cse.ucsc.edu/goldenPath/hg38/database/simpleRepeat.txt.gz>). Although, typical unit lengths of microsatellites vary from one to six base pairs, but a recent study pointed out that microsatellite loci with long unit lengths cannot be detected reliably by using 100 bp (which is the reads length utilized by TCGA) sequencing reads<sup>35</sup>. Thus we only retrieved loci of simple sequence repeats whose unit lengths are less than and equal to five base pairs. Besides, we utilized another 13 features (see Supplementary Table S1, row 10–22), which were calculated from information of ‘Variant\_Classification’ columns of MAF files. These 13 features gave more detailed information about potential translational and regulatory effects of mutations that can differentiate deleterious mutations from not deleterious ones. In general, all these 22 features denote counts of different kinds of mutations normalized by lengths (Mb) of captured exome sequences. For all 1432 tumors used in this study, their captured exome sequence lengths were obtained from their TCGA project marker papers (COAD: 44 Mb, READ: 44 Mb, STAD: 50 Mb, UCEC: 44 Mb)<sup>31–33</sup>.

**Training and testing set.** Features and corresponding MSI-PCR determined MSI status for all tumors were formulated into a large table (see Supplementary 1432\_tumor\_feature.csv). This table were then randomly split into two data sets, training and testing, according to a ratio of three to one for each tumor type, using a function called “train\_test\_split” provided by sklearn. Tables 1 and 2 shows the summary of these two datasets.

**Analysis of 22 features regarding their importance in MSI classification.** We used the random forest algorithm to perform comprehensive analysis of all 22 features in order to gain their importance in MSI classification<sup>23</sup>. A random forest classifier was trained by the 1074-tumor training set using a random forest classifier framework provided by the python package scikit-learn (sklearn, 0.19.1)<sup>30</sup>. Importance scores of all features were then obtained from the “feature\_importances\_” attribute of the trained random forest classifier.

**Categorization of MSI.** TCGA categorized MSI into three types using MSI-PCR: MSS (microsatellite stable), MSI-L (microsatellite instability low), and MSI-H (microsatellite instability high) in its clinical data. In this study, we grouped first two categories (MSS and MSI-L) together and denoted them as MSS because a recent study pointed out MSS and MSI-L tumors possess similar MSI burden<sup>13</sup>. We denoted MSI-H tumors as positive samples and MSS tumors as negative samples to make it more convenient for binary classification and model evaluation.

**Non-TCGA testing set.** MAF files of 390 tumors, including 368 colorectal tumors and 22 stomach tumors generated by two non-TCGA projects<sup>24,25</sup> together with their corresponding clinical data that embedded MSI-PCR determined binary (MSS and MSI-H) tumor MSI status, were directly downloaded from cBioPortal<sup>26,27</sup>. 22 features for mutational load characterization of each of these 390 tumors (see Supplementary non\_tcga\_390\_tumor.csv) were extracted by MSIpred from their MAF files. Corresponding captured exome lengths (colorectal: 67 Mb, stomach: 38 Mb) required by MSIpred were obtained from their corresponding papers<sup>24,25</sup>. This data set was denoted as non-TCGA testing set, which was aimed to validate performance of MSIpred, and to conduct comparisons among original MSIpred (22 features), modified MSIpred (9 features), and MSIseq (10 features).

**Support vector machine.** Support vector machines (SVMs) have been widely used in classification problems of computational biology due to high accuracy and flexibility in statistical modeling<sup>36</sup>. In this study, we utilized a SVM framework provided by the python package scikit-learn (sklearn, 0.19.1)<sup>30</sup> to build our binary (MSS and MSI-H) classifier. As support vector machine is sensitive to scaling of features, our classifier actually was designed to incorporate a two-step framework: the first step is a standardized scaler provided by StandardScaler class of sklearn for feature normalization; the second step is a support vector machine classifier. A radial basis function (RBF) kernel of SVM was chosen and hyper parameters, gamma and C were tuned with 1074-tumor training data using grid search paired with ten-fold cross validation, which was achieved using GridSearchCV function of sklearn. The hyper parameter set yielding the largest average accuracy during cross validation was chosen. The SVM framework with optimized hyper-parameters was then trained with all training data (1074 tumors), and this trained SVM classifier was embedded in MSIpred.

**Model assessment metrics.** As our 358-tumor testing set is imbalanced (280 MSS (negative) and 78 MSI-H(positive)), it is not sufficient to evaluate performance of our classifier by just using overall accuracy. Thus, we incorporated several other metrics for model evaluation. Equations of these metrics are given with following abbreviations: sensitivity (*Se*), specificity (*Sp*), precision (*Pre*), accuracy (ACC), FP (False Positive), FN (False Negative), TP (True Positive), TN (True Negative).



$$Se = \frac{TP}{TP + FN}$$

$$Sp = \frac{TN}{TN + FP}$$

$$Pre = \frac{TP}{TP + FP}$$

$$Acc = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

$$G\ Mean = \sqrt{Se * Sp}$$

$$F1\ Score = 2 * \frac{Pre * Se}{Pre + Se}$$

Receiver operating characteristic (ROC) curve and precision-recall curve were plotted to visualize the performance of MSIpred. Coordinate data used for plotting precision-recall curve and ROC curve was obtained by “precision\_recall\_curve” function and “roc\_curve” function provided by sklearn, respectively. Visualizations were achieved using Matplotlib package.

**Comparisons among original MSIpred, modified MSIpred and MSIseq.** We compared performances of our original MSIpred (22 features) with a modified version of MSIpred (9 features) and MSIseq (10 features) using the aforementioned 390-tumor non-TCGA testing dataset. The simple sequence repeats reference file required by these three tools was based on the same human reference genome (i.e., GRCh37) that the testing MAF files were based on. The modified MSIpred implemented a MSI classifier that was trained by just 9 (see Table S1 row 1–9) of 22 features of all tumors from the 1074-tumor training set. Performance comparisons of the original MSIpred, modified MSIpred and MSIseq were conducted by using aforementioned metrics.

**Statistical analysis.** Wilcoxon rank sum tests were conducted using a built-in R (version 3.4.3) function wilcox.test. Data visualization was achieved using R package ggplot2 (version 2.2.1)<sup>37</sup>, python package Matplotlib (version 2.0.2)<sup>38</sup>, and python package Seaborn (version 0.8.0)<sup>39</sup>.

### Availability

MSIpred is written in python 2. The package and detailed usage documents are freely available on GitHub: (<https://github.com/bioinfolabmu/MSIpred>). All the datasets generated and analyzed during the current study are available from the corresponding author on reasonable request.

### References

- Zane, L., Bargelloni, L. & Patarnello, T. Strategies for microsatellite isolation: A review. *Molecular Ecology* **11**, 1–16 (2002).
- Jiricny, J. The multifaceted mismatch-repair system. *Nature Reviews Molecular Cell Biology* **7**, 335–346 (2006).
- Murphy, K. M. *et al.* Comparison of the microsatellite instability analysis system and the Bethesda panel for the determination of microsatellite instability in colorectal cancers. *J. Mol. Diagnostics* **8**, 305–311 (2006).
- Umar, A. *et al.* Revised Bethesda Guidelines for hereditary nonpolyposis colorectal cancer (Lynch syndrome) and microsatellite instability. *J. Natl. Cancer Inst.* **96**, 261–268 (2004).
- Boland, R. C. & Goel, A. Microsatellite Instability in Colorectal Cancer. *Gastroenterology* **138**, 2073–2087 (2010).
- Hendriks, Y. M. C. *et al.* Diagnostic approach and management of Lynch syndrome (hereditary nonpolyposis colorectal carcinoma): a guide for clinicians. *CA. Cancer J. Clin.* **56**, 213–25 (2006).
- Gryfe, R. *et al.* Tumor microsatellite instability and clinical outcome in young patients with colorectal cancer. *N. Engl. J. Med.* **342**, 69–77 (2000).
- Pritchard, C. C. *et al.* Complex MSH2 and MSH6 mutations in hypermutated microsatellite unstable advanced prostate cancer. *Nat. Commun.* **5** (2014).
- Boland, C. R. *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 Research* **58**, 5248–5257 (1998).
- Niu, B. *et al.* MSIsensor: Microsatellite instability detection using paired tumor-normal sequence data. *Bioinformatics* **30**, 1015–1016 (2014).
- Salipante, S. J., Scroggins, S. M., Hampel, H. L., Turner, E. H. & Pritchard, C. C. Microsatellite instability detection by next generation sequencing. *Clin. Chem.* **60**, 1192–1199 (2014).
- Ni Huang, M. *et al.* MSIseq: Software for assessing microsatellite instability from catalogs of somatic mutations. *Sci. Rep.* **5** (2015).
- Hause, R. J., Pritchard, C. C., Shendure, J. & Salipante, S. J. Classification and characterization of microsatellite instability across 18 cancer types. *Nat. Med.* **22**, 1342–1350 (2016).
- Foltz, S. M., Liang, W.-W., Xie, M. & Ding, L. MIRMMR: binary classification of microsatellite instability using methylation and mutations. *Bioinformatics* **33**, 3799–3801 (2017).
- Veenstra, R., Kostine, M., Cleton-Jansen, A. M., De Miranda, N. F. C. C. & Bovée, J. V. M. G. Immune checkpoint inhibitors in sarcomas: In quest of predictive biomarkers. *Lab. Investig.* **98**, 41–50 (2018).
- Roth, A. *et al.* JointSNVMix: A probabilistic model for accurate detection of somatic mutations in normal/tumour paired next-generation sequencing data. *Bioinformatics* **28**, 907–913 (2012).

17. Larson, D. E. *et al.* Somaticsnpiper: Identification of somatic point mutations in whole genome sequencing data. *Bioinformatics* **28**, 311–317 (2012).
18. Saunders, C. T. *et al.* Strelka: Accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics* **28**, 1811–1817 (2012).
19. Koboldt, D. C. *et al.* VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576 (2012).
20. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).
21. Shiraishi, Y. *et al.* An empirical Bayesian framework for somatic mutation detection from cancer genome sequencing data. *Nucleic Acids Research* 1–10. <https://doi.org/10.1093/nar/gkt126> (2013).
22. Weinstein, J. N. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–20 (2013).
23. Genuer, R., Poggi, J.-M. & Tuleau-Malot, C. Variable selection using random forests. *Pattern Recognit. Lett.* **31**, 2225–2236 (2010).
24. Giannakis, M. *et al.* Genomic Correlates of Immune-Cell Infiltrates in Colorectal Carcinoma. *Cell Rep.* **15**, 857–865 (2016).
25. Wang, K. *et al.* Exome sequencing identifies frequent mutation of ARID1A in molecular subtypes of gastric cancer. *Nat. Genet.* **43**, 1219–1223 (2011).
26. Cerami, E. *et al.* The cBio Cancer Genomics Portal: An open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2**, 401–404 (2012).
27. Gao, J. *et al.* Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal. *Sci. Signal.* **6**, (1–34) (2013).
28. McKinney, W. Data Structures for Statistical Computing in Python. *Proc. 9th Python Sci. Conf.* **1697900**, 51–56 (2010).
29. Halbert, C.-L. & Tretyakov, K. Intervaltree (2013).
30. Pedregosa, F. & Varoquaux, G. *Scikit-learn: Machine learning in Python. ... of Machine Learning ...* **12** (2011).
31. Cancer Genom Atlas. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337 (2012).
32. Bass, A. J. *et al.* Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* **513**, 202–209 (2014).
33. Kandoth, C. *et al.* Integrated genomic characterization of endometrial carcinoma. *Nature* **497**, 67–73 (2013).
34. Colaprico, A. *et al.* TCGAAbiolinks: An R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.* **44**, e71 (2016).
35. Cortes-Ciriano, I., Lee, S., Park, W. Y., Kim, T. M. & Park, P. J. A molecular portrait of microsatellite instability across multiple cancers. *Nat. Commun.* **8**, 1–12 (2017).
36. Ben-Hur, A., Ong, C. S., Sonnenburg, S., Schölkopf, B. & Rätsch, G. Support vector machines and kernels for computational biology. *Plos Comput. Biol.* **4** (2008).
37. Wilkinson, L. ggplot2: Elegant Graphics for Data Analysis by WICKHAM, H. *Biometrics* **67**, 678–679 (2011).
38. Hunter, J. D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **9**, 99–104 (2007).
39. Waskom, M. *et al.* mwaskom/seaborn: v0.8.0 (July 2017) (Version v0.8.0). Zenodo, <https://doi.org/10.5281/zenodo.824567> (2017, July 8).

## Acknowledgements

This project is funded by the Committee on Faculty Research (CRF) Program, the Office for the Advancement of Research & Scholarship (OARS), the Department of Biology, and Cell, Molecular and Structural (CMSB) Graduate Program, Miami University, Oxford, Ohio, USA.

## Author Contributions

C.W. and C.L. conceived and designed MSIPred. C.W. implemented MSIPred in python, tested MSIPred, analyzed the data, and prepared all figures and tables. C.W. and C.L. wrote the main manuscript text. All authors read and approved the final manuscript.

## Additional Information

**Supplementary information** accompanies this paper at <https://doi.org/10.1038/s41598-018-35682-z>.

**Competing Interests:** The authors declare no competing interests.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2018