Deletion 占所有 Deletion 的比例, PI 是指微卫星区域 Insertion 占所有 Insertion 的比例. MSIseq 基于决策树算法的分类器在单核苷酸替代 (Single Nucleotide Substitution, SNS) 和 Indel 短突变的高级别数据上进行模型训练和测试, 该方法在测试时仅使用微卫星区域小片段碱基插入与删除的比例这一特征. MSIpred 利用支持向量机对从突变数据中提取的 12 个特征进行模型训练和测试. MSIseq、MSIpred 等是从突变注释格式 (Mutation Annotation Format, MAF) 中提取特征用于模型训练和测试, 但 MAF 是由肿瘤与配对正常组织组织的测序数据产生. 因此, 提供一个直接对肿瘤组织测序数据的 BAM 文件进行分析的检测工具十分必要.

基于肿瘤与正常组织成对测序数据的 MSIsensor 1.1 和 MANTIS[16]可以直接对 BAM 文件进行分析. MANTIS 将样本的肿瘤组织与配对的正常组织在每个微卫星位点的等位基因分布看作两个向量, 定义这两个向量的 L_1 范数作为衡量该位点稳定程度的度量, 最后通过统计该样本所有位点的 L_1 范数的平均值获得该样本的 MSI 分数. MSIsensor 通过卡方检验比较肿瘤组织和配对的正常组织的相同微卫星位点的等位基因分布是否显著不同判断该位点是否稳定, 最后统计不稳定的位点占总位点的比例判断样本的 MSI 状态[17,18].

通过对 MSIsensor1.1 和 MANTIS 软件之间的对比, 本研究发现 MSIsensor1.1 具有以下两方面优点: (1) 设计基于 C++实现、计算效率高; (2) 通过使用全基因组范围内的微卫星位点信息的探测和卡方检验方法判断微卫星状态, 准确性非常高. 因此, 本研究选择该软件进行优化.

MSIsensor1.1 基于卡方检验的探测模型包含以下两个功能模块:

(1) 获取位点基本信息模块. 通过对参考基因组文件 (Reference Genome.fa) 执行扫描 (scan) 命令, 构建一个由微卫星位点基本信息组成的序列字典 (microsatellites.list), 其基本信息包括: 微卫星位点所在染色体号及其在染色体上的起始位置、微卫星重复单元的大小及重复次数以及微卫星位点的左右翼信息;

(2) 探测样本 MSI 状态模块. 通过对肿瘤组织测序数据文件 (Sample.tumor.bam)、正常组织测序数据文件 (Sample.normal.bam) 和 microsatellites.list 执行 msi 命令获得待测样本的 MSI 分数 (MSIscore). 该模块共有四个结果文件, 分别为: MSI 分数文件 (Output 文件)、微卫星位点等位基因分布文件 (Output_dis 文件)、germline 位点文件 (Output_germline 文件) 和 somatic 位点文件 (Output_somatic 文件). 该模块具体步骤如下: ① 计算每个位点在样本肿瘤与正常组织成对测序数据上微卫星的长度分布, 并筛选在肿瘤和正常组织测序数据中测序深度均大于 20 的微卫星位点; ② 利用卡方检验比较符合条件①的微卫星位点在肿瘤与正常组织中的分布差异, 若显著不同, 则判定该位点为不稳定的微卫星位点; ③ 通过统计不稳定位点占总位点的比例获得样本的 MSIscore, 进而判断该样本的 MSI 状态.

2 实验方法

目前, 基于肿瘤与正常组织成对测序数据的 MSI 探测方法, 利用配对的正常组织测序数据为参考, 对肿瘤组织的微卫星位点的不稳定情况进行评判, 最终达到判断样本 MSI 状态的目标. 对于缺少配对的正常组织测序数据做参照的情况, 为了避免使用大量 MSS 样本的正常组织测序数据构建基准线而丢失特异性 MSI 位点, 本文提出一种新模型, 该模型根据肿瘤组织每个微卫星位点等位基因分布的混乱情况, 来判断该位点及样本的 MSI 状态.

在正常细胞中, 错配修复系统 (Mismatch Repair, MMR) 提供了一种高效的机制来纠正 DNA 复制过程中发生的错误. 当 MMR 受损时, 例如错配修复基因 MLH1、MSH2 和 MSH3 的失活, 将导致基因序列——特别是微卫星位点重复单元的错误插入或删除得不到纠正, 微卫星位点的重复单元的重复次数出现波动, 即 MSI 发生[7]. 从酵母的研究中发现, 在 DNA 错配修复蛋白突变后, 微卫星中短缺失的数量明显增加, 而微卫星中插入的数量没有明显变化[8]. 后续研究也证实了, 由于 MMRD, 微卫星区域会发生小片段碱基的插入与缺失. 受已有研究成果的启发, 我们考虑到, 与 MSS 样本相比, MSI 样本肿瘤组织的微卫星位点的重复单元的重复次数的分布会因为 MMRD 呈现比较混乱的状态. 图 1 为 MSS 样本与 MSI 样本的肿瘤组织在同一微卫星位点的等位基因分布. 从图中可以看出, MSI 样本的微卫星位点等位基因分布与 MSS 样本相比较混乱. 鉴于 MSI 样本肿瘤组织微卫星位点测序数据这一特性, 本研究提出使用信息熵理论来探测样本的 MSI 状态.

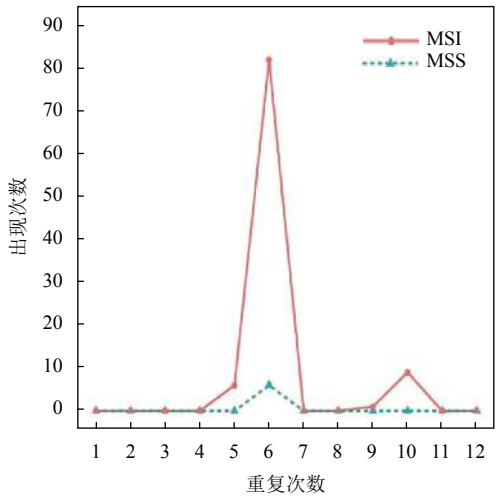

图 1 MSI 样本与 MSS 样本的同一微卫星位点等位基因分布比较

信息熵是描述“混乱”程度的量度, 系统越有序, 信息数据越集中的地方, 熵值越小; 系统越混乱, 信息数据越分散的地方, 熵值越大[19]. 扩展后的 MSIsensor 工作流程如图 2 所示.

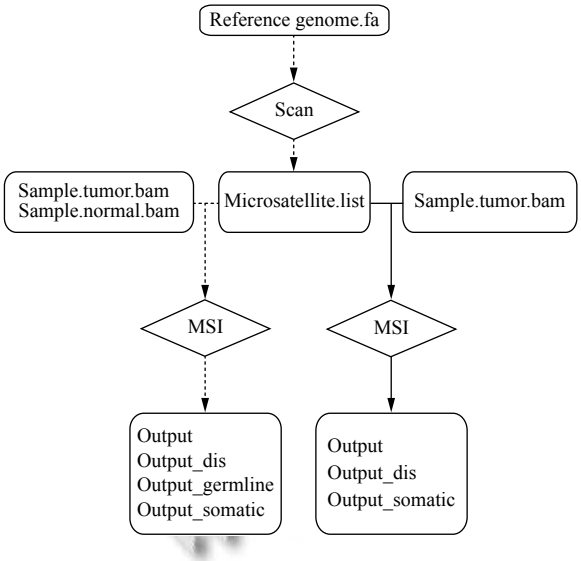

图 2 MSIsensor 工作流程图

图中虚线所指是 MSIsensor1.1 已有模块, 实线所指功能是本文在实现探测样本 MSI 状态功能中提出的基于信息熵理论在样本单肿瘤组织测序数据进行探测的新模型. 该模型共有三个输出, 分别为: Output 文件、Output_dis 文件、Output_somatic 文件. 新模型主要包含以下子模块:

(1) 目标位点筛选功能. 通过计算每个微卫星位点在肿瘤组织测序数据上的分布, 可筛选出测序深度大于 20(默认) 的微卫星位点.

(2) 目标位点过滤功能. 由于测序过程中可能会带来背景噪声信息, 因此, 对 (1) 筛选出的目标位点集进行过滤. 过滤规则为: 过滤掉肿瘤组织在该位点支持 reads 数小于 3 的等位基因 (被认为是噪声数据), 接下来以剩下的等位基因数作为样本在该位点的等位基因数.

(3) 加权信息熵值计算功能. 根据步骤 (2) 得到的每个位点等位基因分布情况, 计算加权信息熵值. 信息熵值计算公式如下:

$$H(U) = -\sum_{n=1}^{n} p_i log p_i \quad (1)$$

其中, p_i 指的是针对位点的每个重复次数, 其不为零支持 reads 数占该位点支持 reads 数总和的比例. 支持 reads 数不为零的重复次数的个数会较大程度影响每个位点的信息熵值, 即每个位点的信息熵值会因支持 reads 数不为零的重复次数的个数变化而跨越幅度较大. 为了方便后续信息熵值阈值 (*cutoff*) 的确定, 对每个位点的熵值赋予一定的权重 (*weight*). 若该位点的 $H(U)$*$weight$ 大于等于 0.3(默认), 则判定该位点不稳定. 其中, *weight* 等于该位点支持 reads 数不为零的重复次数的个数的倒数. 在 TCGA 数据集的不同癌种上使用不同的阈值进行多次实验, 通过比较得到, 当使用阈值 0.3 时, 在不同癌种上 Accuracy 都比较高. 图 3 给出了在 TCGA 数据集上随着 *cutoff* 变化时的 Accuracy 趋势.

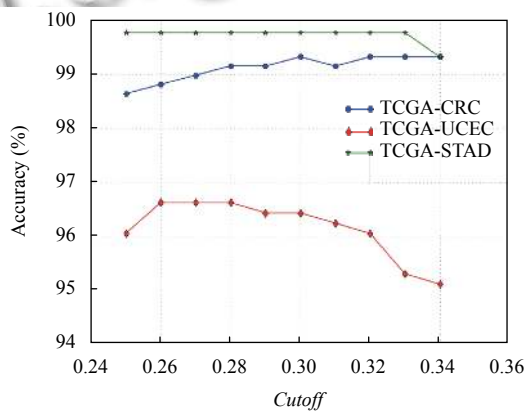

图 3 不同 *cutoff* 下的 Accuracy 比较

(4) MSIscore 计算功能. 通过统计由 (3) 得到不稳定位点数占满足条件 (1) 的微卫星总位点数的比例, 获得该样本的 MSIscore, 进而判断样本的 MSI 状态.

3 实验结果与讨论

本研究使用的数据来自国际癌症基因组图谱 (The Cancer Genome Atlas, TCGA) 计划[20]、欧洲基因组-表型组档案 (European Genome-phenome Archive, EGA)[21,22]和北京肿瘤医院 (Beijing Cancer Hospital), 包含的癌症类型有胃癌 (STomach ADenocarcinoma, STAD)、子宫内膜癌 (Uterine Corpus Endometrial Carcinoma, UCEC)、结肠癌 (COlon ADenocarcinoma, COAD) 和直肠腺癌 (REctum ADenocarcinoma, READ), 具体情况如表 1 所示.

表 1 TCGA、EGA 和 Beijing Cancer Hospital 数据集

数据集	癌症类型	MSS	MSI	Total
TCGA	STAD	356	85	441
	CRC (COAD、READ)	510	78	588
	UCEC	364	168	532
EGA	STAD	37	9	46
	COAD	57	14	71
Beijing Cancer Hospital	STAD	198	3	201

为了衡量 MSIsensor 扩增的使用单肿瘤组织测序数据探测 MSI 模块的表现性能, 本研究将其与 MSI sensor 基于卡方检验的 MSI 探测模块在表 1 所示数据上的测试结果进行对比. 由于不同的癌症类型或不同测序平台得到的测序序列会稍有不同, 本文不建议使用统一的 MSIscore 阈值划分 MSS 样本和 MSI 样本. 针对基于卡方检验在肿瘤与正常组织成对测序数据上探测 MSI 的模块, 本研究在 TCGA STAD、TCGA CRC、TCGA UCEC、EGA STAD、EGA COAD 和 Beijing Cancer Hospital 数据集上采用的 MSIscore 阈值分别为 3、10、3、3、10、和 15, 其各项性能指标表现如表 2 所示, 其中, 准确率 (Accuracy)表示预测状态正确的样本数占待检测样本总数的百分比; 精确率 (Precision) 表示软件预测的真实 MSI 样本数占软件预测出的 MSI 样本总数的百分比; 灵敏度 (Sensitivity) 表示软件预测的真实 MSI 样本数占 MSI 总样本数的百分比; 特异性 (Specificity) 表示软件预测的真实 MSS 样本数占 MSS 总样本数的百分比; F-分数 (F-score) 表示 Sensitivity 和 Precision 两个性能度量的调和平均数, 用来综合评估软件的性能. 具体分类效果如图 4 所示. 针对基于信息熵在单肿瘤组织上探测 MSI 的模块, 本研究在 TCGA STAD、TCGA CRC、TCGA UCEC、EGA STAD、EGA COAD 和 Beijing Cancer Hospital 数据集上采用的阈值分别为 13、20、13.5、11、15 和 20. 该模型的各项性能指标表现如表 3 所示, 具体分类效果如图 5 所示.

表 2 基于卡方检验探测模块的测试结果

数据集	性能指标				
	Accuracy(%)	Precision(%)	Sensitivity(%)	Specificity(%)	F1-score (%)
TCGA STAD	99.77	100	98.82	100	99.41
TCGA CRC	97.79	91.14	92.31	98.63	91.72
TCGA UCEC	96.99	94.19	96.43	97.25	95.29
EGA STAD	100	100	100	100	100
EGA COAD	100	100	100	100	100
Beijing Cancer Hospital	100	100	100	100	100

以上结果证明, 在 TCGA 上确定的对每个位点判定是否稳定的加权信息熵阈值 0.3, 在 EGA、Beijing Cancer Hospital 数据集上同样适用. 同时, 该结果也表明本文提出的基于信息熵理论在单肿瘤组织测序数据上探测样本 MSI 状态的方法是可行的.

基于信息熵的 MSI 探测方法只需要对单肿瘤组织测序数据进行分析, 因此, 在探测 MSI 的前期准备工作中不需要对正常组织进行测序, 可以节省测序成本; 且与 MSIsensor1.1 在肿瘤与正常组织测序数据上探测 MSI 的方法相比较, 其运行效率提高了一倍.

4 结论与展望

本研究针对样本单肿瘤组织的外显子或全基因组测序数据的 MSI 探测问题, 选择从高通量测序数据入手, 基于已有软件 MSIsensor1.1, 增加采用信息熵理论探测 MSI 状态的模块. 扩增后的软件不仅可针对肿瘤与正常组织成对测序数据探测样本 MSI 状态, 而且当无配对的正常组织测序数据做输入参照物和不严重影响精度的情况下, 可满足样本单肿瘤组织测序数据的 MSI 探测需求, 是一套功能完备、兼容多种测序数据模式、具备实用性和前瞻性的 MSI 探测软件.

后续我们将围绕以下两方面工作展开: (1) 基于此

工作深入挖掘每个位点信息并筛选与肿瘤发生发展相关的基因, 这样不但可以节省测序的成本, 而且可以提高检测的正确率; (2) 探究基于血浆的 MSI 探测方法. 癌细胞具有扩散性, 当癌细胞侵犯血管进入血液时, 血液中也会有 MSI 信号, 但此时信号很微弱, 不易检测. 目前大多数 MSI 检测局限于组织测序数据, 但是对于不方便提取组织的患者, 开发基于血浆的 MSI 探测方法具有重要的意义. 下一步, 我们将围绕此工作展开, 探求更强、更符合实际应用需求的 MSI 探测软件.

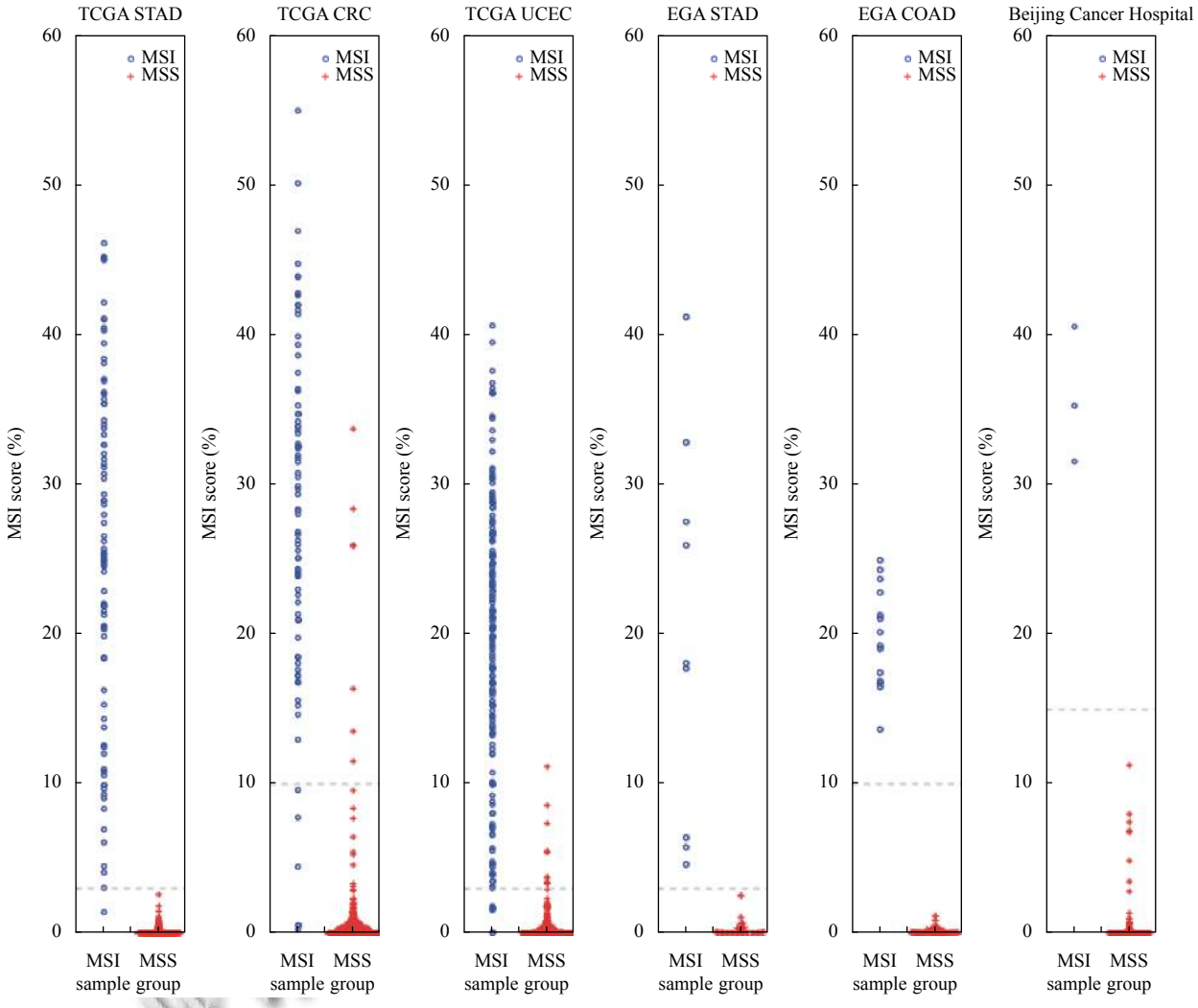

图 4 基于卡方检验探测模块的测试结果

表 3 基于信息熵理论探测模块的测试结果

数据集	性能指标				
	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F1-score (%)
TCGA STAD	99.77	100	98.82	100	99.41
TCGA CRC	99.32	100	94.87	100	97.37
TCGA UCEC	96.43	93.57	95.24	96.98	94.40
EGA STAD	95.65	88.89	88.89	97.30	88.89
EGA COAD	100	100	100	100	100
Beijing Cancer Hospital	100	100	100	100	100

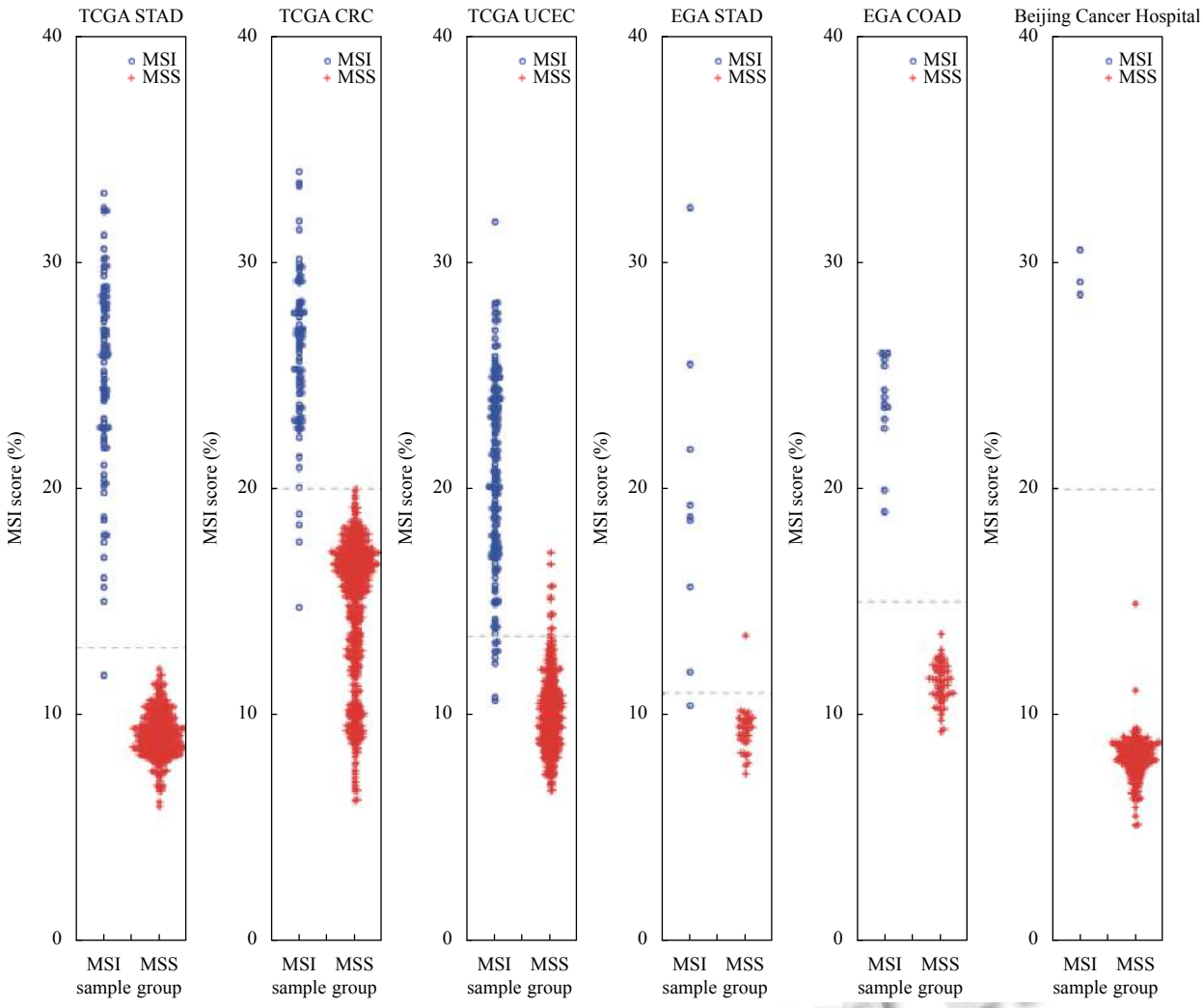

图 5 基于信息熵理论探测模块的测试结果

参考文献

1 Mardis ER. A decade's perspective on DNA sequencing technology. Nature, 2011, 470(7333): 198–203. [doi: 10.1038/nature09796]

2 Ribic CM, Sargent DJ, Moore MJ, *et al*. Tumor microsatellite-instability status as a predictor of benefit from fluorouracil-based adjuvant chemotherapy for colon cancer. New England Journal of Medicine, 2003, 349(3): 247–257. [doi: 10.1056/NEJMoa022289]

3 Pino MS, Chung DC. Microsatellite instability in the management of colorectal cancer. Expert Review of Gastroenterology & Hepatology, 2011, 5(3): 385–399.

4 Murphy KM, Zhang SL, Geiger T, *et al*. Comparison of the microsatellite instability analysis system and the Bethesda panel for the determination of microsatellite instability in colorectal cancers. The Journal of Molecular Diagnostics, 2006, 8(3): 305–311. [doi: 10.2353/jmoldx.2006.050092]

5 FDA News Release. FDA approves first cancer treatment for any solid tumor with a specific genetic feature. https://www.fda.gov/news-events/press-announcements/fda-approves-first-cancer-treatment-any-solid-tumor-specific-genetic-feature, 2017-05-23.

6 Aaltonen LA, Peltomaki P, Leach FS, *et al*. Clues to the pathogenesis of familial colorectal cancer. Science, 1993, 260(5109): 812–816. [doi: 10.1126/science.8484121]

7 Shia J. Immunohistochemistry versus microsatellite instability testing for screening colorectal cancer patients at risk for hereditary nonpolyposis colorectal cancer syndrome: Part I. The utility of immunohistochemistry. The Journal of Molecular Diagnostics, 2008, 10(4): 293–300. [doi: 10.2353/jmoldx.2008.080031]

8 Beamer LC, Grant ML, Espenschied CR, *et al*. Reflex

immunohistochemistry and microsatellite instability testing of colorectal tumors for Lynch syndrome among US cancer programs and follow-up of abnormal results. Journal of Clinical Oncology, 2012, 30(10): 1058–1063. [doi: 10.1200/JCO.2011.38.4719]

9 Sankila R, Aaltonen LA, Järvinen HJ, *et al*. Better survival rates in patients with MLH1-associated hereditary colorectal cancer. Gastroenterology, 1996, 110(3): 682–687. [doi: 10.1053/gast.1996.v110.pm8608876]

10 Niu BF, Ye K, Zhang QY, *et al*. MSIsensor: Microsatellite instability detection using paired tumor-normal sequence data. Bioinformatics, 2014, 30(7): 1015–1016. [doi: 10.1093/bioinformatics/btt755]

11 Kelkar YD, Strubczewski N, Hile SE, *et al*. What is a microsatellite: A computational and experimental definition based upon repeat mutational behavior at A/T and GT/AC repeats. Genome Biology and Evolution, 2010, 2(1): 620–635.

12 Salipante SJ, Scroggins SM, Hampel HL, *et al*. Microsatellite instability detection by next generation sequencing. Clinical Chemistry, 2014, 60(9): 1192–1199. [doi: 10.1373/clinchem.2014.223677]

13 Lu YH, Soong TD, Elemento O. A novel approach for characterizing microsatellite instability in cancer cells. PLoS One, 2013, 8(5): e63056. [doi: 10.1371/journal.pone.0063056]

14 Huang MN, McPherson JR, Cutcutache I, *et al*. MSIseq: Software for assessing microsatellite instability from catalogs of somatic mutations. Scientific Reports, 2015, 5: 13321. [doi: 10.1038/srep13321]

15 Wang C, Liang C. MSIpred: A python 2 package for the classification of tumor microsatellite instability from tumor mutation annotation data using a support vector machine. Nature Scientific Reports, 2018, 12(6): 1835. [doi: 10.1038/s41598-018-35682-z]

16 Kautto EA, Bonneville R, Miya J, *et al*. Performance evaluation for rapid detection of pan-cancer microsatellite instability with MANTIS. Oncotarget, 2017, 8(5): 7452–7463.

17 Karran P. Microsatellite instability and DNA mismatch repair in human cancer. Seminars in Cancer Biology, 1996, 7(1): 15–24. [doi: 10.1006/scbi.1996.0003]

18 Sia EA, Kokoska RJ, Dominska M, *et al*. Microsatellite instability in yeast: Dependence on repeat unit size and DNA mismatch repair genes. Molecular and Cellular Biology, 1997, 17(5): 2851–2858. [doi: 10.1128/MCB.17.5.2851]

19 沈世镒, 吴忠华. 信息论基础与应用. 北京: 高等教育出版社, 2004,

20 Weinstein JN, Collisson EA, Mills GB, *et al*. The cancer genome atlas pan-cancer analysis project. Nature Genetics, 2013, 45(10): 1113–1120. [doi: 10.1038/ng.2764]

21 Liu JF, McCleland M, Stawiski EW, *et al*. Integrated exome and transcriptome sequencing reveals ZAK isoform usage in gastric cancer. Nature Communications, 2014, 5: 3830. [doi: 10.1038/ncomms4830]

22 Seshagiri S, Stawiski EW, Durinck S, *et al*. Recurrent R-spondin fusions in colon cancer. Nature, 2012, 488(7413): 660–664. [doi: 10.1038/nature11282]