PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2021, 48(6): 677~687
www.pibb.ac.cn

MethyScan：一种甲基化特异性PCR引物设计及评估工具

曹英豪*
（中国医学科学院基础医学研究所 & 北京协和医学院基础学院，北京 100730）

摘要 DNA甲基化是重要的表观遗传现象，对基因表达发挥重要调控功能. 大量研究表明，基因DNA甲基化是重要的临床诊断生物标志物. 在临床上，实施快速、准确的DNA甲基化状态检测是诊断应用的前提和关键. 甲基化特异性PCR（methylation specific PCR，MSP）通过将两种引物与甲基化、非甲基化模板各自特异性结合和扩增，实现基因甲基化状态的区分，是切实可行、简单便捷的临床诊断实验技术. 但是，不同于常规PCR，MSP主要存在如何强化引物-甲基化/非甲基化模板特异性结合、降低引物序列*T*m值差异、去除假阳性扩增及提高敏感性等四大难点. 尽管大多数MSP引物设计软件对上述难题都提出了各自解决办法，但在引物设计影响因素考虑、设计与评估并行处理及特异性扩增预测等方面仍然存在较大缺陷. 为此，本研究通过对MethPrimer、MSPPrimer、MethBlast、BiSearch等现有MSP引物设计软件原理的深入探究，以及对Bowtie、SAMtools和BEDTools等工具的有效综合整合，基于图形库Matplotlib和第三方Python功能库BioPython与Primer3-py实现了具有系列优点的甲基化特异性PCR引物设计与评估可视化工具MethyScan. 它具有引物设计、基因组索引、引物评估等三大完整功能模块，不仅可快速进行MSP引物设计，实现巢式（Nested）引物适配，还可基于4种基因组碱基转换模板分析引物结合信息，图形化展示非特异性扩增与目的片段差异，从而综合评估引物特异性-非特异性扩增. 同时，对食管癌、结直肠癌等多种恶性肿瘤中6个潜在生物标志物TFPI-2、NDRG4、CDKN2A、CD44、CASP8和SDHD的甲基化引物设计对比结果表明，MethyScan不仅可获得更多CpG位点的检测引物，而且所获得MSP引物位置与其他软件结果相同或相近，且引物间Tm值差值更小. 总之，作为首个图形化展示特异性-非特异性扩增差异MSP引物设计工具，MethyScan可有效提高甲基化引物设计准确性，为临床DNA甲基化检测项目开展、检测试验实施及诊断试剂盒研发提供有力支撑. MethyScan工具下载地址：https://github.com/bioinfo-ibms-pumc/MethyScan.

关键词 引物设计，DNA甲基化，引物评估，MSP引物
中图分类号 Q344，Q341
DOI：10.16476/j.pibb.2020.0413

表观遗传现象是联接基因型与表型之间的一种生物学桥梁，是指DNA序列虽未改变，仍引起基因和染色体出现最终表型变化结果[1]. DNA甲基化是重要的表观遗传现象，对基因表达发挥重要调控功能. 在脊椎动物中，DNA甲基化主要表现为5-羟甲基胞嘧啶（5mC），即在DNA甲基化转移酶作用下，甲基基团（—CH_3）转移至胞嘧啶（5'-CpG-3'）第5位碳原子上[2]. 人类基因组约70%~80%的CpG双核甘酸通常处于甲基化状态[3]，其余非甲基化CpG双核苷酸则多以密集簇状出现于基因启动子区，形成高GC含量CpG岛[2, 4]. 值得注意的是，启动子区CpG岛甲基化状态变化可对细胞产生重要影响. 例如，肿瘤抑制基因（tumor suppressor genes）启动子区高甲基化往往会导致肿瘤的发生[5]. 此外，在胚胎早期发育中，DNA甲基化具有时间点动态变化现象. 这种现象赋予了胚胎动态重编程的可能，极有可能为胚胎全能性发展提供必需条件[6].

不仅如此，大量研究表明，DNA甲基化具有重要的临床诊断价值，是潜在有效的生物标志物.

* 通讯联系人.
Tel：010-69156963，E-mail：yhcao@ibms.pumc.edu.cn
收稿日期：2020-11-25，接受日期：2021-02-04

例如，Cologuard和Epi proColon两家公司基于NDRG4、BMP3、SEPT9等基因高甲基化位点开发出可应用于临床的结直肠癌检测试剂盒[7-8]. 在食管癌相关研究中，有报道证实TFPI-2基因启动子区在癌组织中呈现高甲基化，并且与该肿瘤分化显著相关[9]. 另外，在食管鳞癌中研究发现，HSPB2基因表达显著低于正常组织，并且该基因启动子区甲基化是影响基因表达的主要因素，是食管鳞癌发生的致病因子（casual factor）之一. 因此，HSPB2基因启动子区甲基化被认为是潜在的食管鳞癌生物标志物[10]. 而在宫颈癌研究中，目前普遍认为高危人乳头瘤病毒（hrHPV）持续感染是导致宫颈癌的根本原因. 由于宫颈上皮内瘤变（cervical intraepithelial neoplasia，CIN）是与宫颈浸润癌密切相关的癌前病变，因此临床上可通过CIN分期分型（triage）来实现宫颈癌发生发展的连续变化跟踪和CIN-hrHPV阳性个体鉴定[11]. 在具体临床操作中，除了hrHPV病毒分型手段外，基于DNA甲基化标志物的分类筛选同样是可能的有效手段[12]. 例如，对GynTect和QIAsure Methylation Test两种DNA甲基化诊断试剂盒的评估研究表明，采用GynTect可实现高特异性CIN2、CIN3的分期分型[5]. 另外，研究表明，DNA甲基化不仅可用于临床诊断，还可应用于预后和存活预测等[13].

在临床上，实施快速、准确的DNA甲基化状态检测是诊断应用的前提和关键. Herman等[14]开发的甲基化特异性PCR技术（methylation specific PCR，MSP），通过将M引物和U引物与甲基化模板、非甲基化模板的各自特异性结合与扩增，从而实现甲基化状态的区分，是切实可行、简单便捷、备受关注的临床诊断实验技术[15-20]. 由于MSP技术简单稳定，基于MSP技术还衍生出了巢式MSP（nested MSP）、定量MSP（quantitative methylation-specific PCR，qMSP）、多重定量MSP（quantitative multiplex methylation-specific PCR，QM-MSP）和微滴式数字MSP（methylation-specific droplet digital PCR，ddMSP）等技术，这为不同浓度下的DNA甲基化状态定量检测提供了多种技术选择[21-24].

不同于常规PCR，MSP技术有多个难点：a. 如何确保两种引物与甲基化、非甲基化模板的对应特异结合；b. 如何通过减少引物类型不同所带来的*T*m值差异，进而降低PCR反应体系的优化难度；c. 如何避免全基因组非特异性扩增造成的假阳性扩增条带；d. 如何解决样本制备时经亚硫酸氢盐处理造成的大量损耗，从而避免敏感性的降低. 目前，针对上述难题，大多数MSP引物设计软件都提出了各自的解决办法. 例如，MethPrimer[25]对目的模板进行甲基化、非甲基化两种模板序列转换并依此设计M引物和U引物：首先，将序列中除5mC外所有C均转换为T用于获取甲基化模板序列，并将序列中所有C转换为T用于获取非甲基化模板序列. 随后，根据Gardiner-Garden等[26]对CpG岛的定义，MethPrimer将进行CpG岛预测，并进而利用Primer3工具进行CpG岛内MSP引物设计. 值得注意的是，在引物设计中，MethPrimer会通过限定*T*m差异最大值来降低引物*T*m值差异，并设定系列约束条件（包括“引物3'端需包含CpG位点、增加引物中CpG数目”）来强化M、U引物与甲基化、非甲基化模板的对应特异结合. 而MSPPrimer[27]则利用滑动窗口来进行MSP引物设计，并通过缩短窗口长度来限定*T*m值的设定. 而且，MSPPrimer还分别采用巢式引物设计和引物特异性分值（即引物3端自由能与不同转换模板差异计算）来提高引物的敏感性和特异性，从而有助于引物的进一步筛选、降低实际扩增中的假阳性. 其他软件，如BiSearch[28]在完成引物设计的同时还加入了引物评估功能. 它通过预定义字符串匹配查找，完成引物在全基因组序列上的非特异性扩增信息分析. MethBlast[29]作为评估MSP引物特异性工具，整合了传统BLAST算法，通过引物与转换后基因组序列比对实现引物-全基因组非特异性扩增信息检测.

尽管上述软件针对MSP技术难题提出了各自的解决办法，但在引物设计影响因素考虑、设计与评估并行处理及特异性扩增预测等方面仍然存在较大缺陷. 例如，MethPrimer允许输入序列最长为5 000 bp，无法同时完成巢式引物设计和引物评估功能. MSPPrimer尽管可进行特异性分值计算，但它仅仅从引物本身理化性质进行引物评估，无法完成全基因组非特异性扩增筛查. BiSearch虽然可利用正则表达式字符串进行全基因组匹配查找，但其检测能力有限，并且在进行字符串匹配时需要提前进行定义. MethBlast仅含有引物评估功能，虽然可通过整合BLAST算法实现全基因组非特异性引物扩增信息检测，但由于所检测引物序列长度相对较短，加之采用传统的BLAST算法，其*E*值（*E*-value）、字段大小（word size）等比对参数并不能

完全适合引物评估. 而且，遗憾的是，MSPPrimer和MethBlast所提供的网页链接目前已经失效，相应工具已不可再用. 总之，上述这些工具尽管从特异性-非特异性扩增信息方面对MSP技术中的引物设计难题进行不断尝试，但仍然存在这样的缺陷：引物设计影响因素考虑不周，引物设计与评估无法并行处理，引物与目标片段其特异性-非特性性预测信息不足，可视化图形信息缺乏影响引物设计的直观、清晰判断等.

针对MSP引物设计难点，并为了综合解决上述软件中各种不足之处，本研究通过对Bowtie[30]、SAMtools[31]和BEDTools[32]等工具的有效综合整合，基于图形库Matplotlib和第三方Python功能库BioPython与Primer3-py实现了MSP引物设计与评估可视化工具MethyScan：具有引物设计、基因组索引、引物评估等三大完整功能模块，用户可通过MethyScan建立多种规则，从而快速进行MSP引物设计和巢式引物适配. 而且，用户还可通过MethyScan提供的4种基因组碱基转换模板对引物结合信息进行有效分析，从而以图形化可视方式获取非特异性扩增与目的片段之间差异信息，进而综合评估引物全基因组特异性-非特异性扩增情况. MethyScan工具的下载地址为：https://github.com/bioinfo-ibms-pumc/MethyScan.

1　数据与方法

1.1　测试数据的获取

在本研究中，以人类基因组GRCh38版本为例，提取了食管癌潜在标志物TFPI-2基因5'端5 000 bp序列，用于MSP引物设计及后续非特异性扩增评估. 为了进一步比较MethyScan与MethPrimer设计引物差异，除TFPI-2基因外，还提取了5个常见肿瘤标志物CDKN2A、CD44、SDHD、NDRG4和CASP8基因上游启动子区1 000 bp序列用于MSP引物设计.

1.2　MethyScan分析流程

MethyScan分析流程整体包括引物设计、基因组索引、引物评估三大模块（图1）. 每个模块均可以单独执行. 为了完成各模块功能，MethyScan建立了引物设计、基因组转换和引物评估等多种规

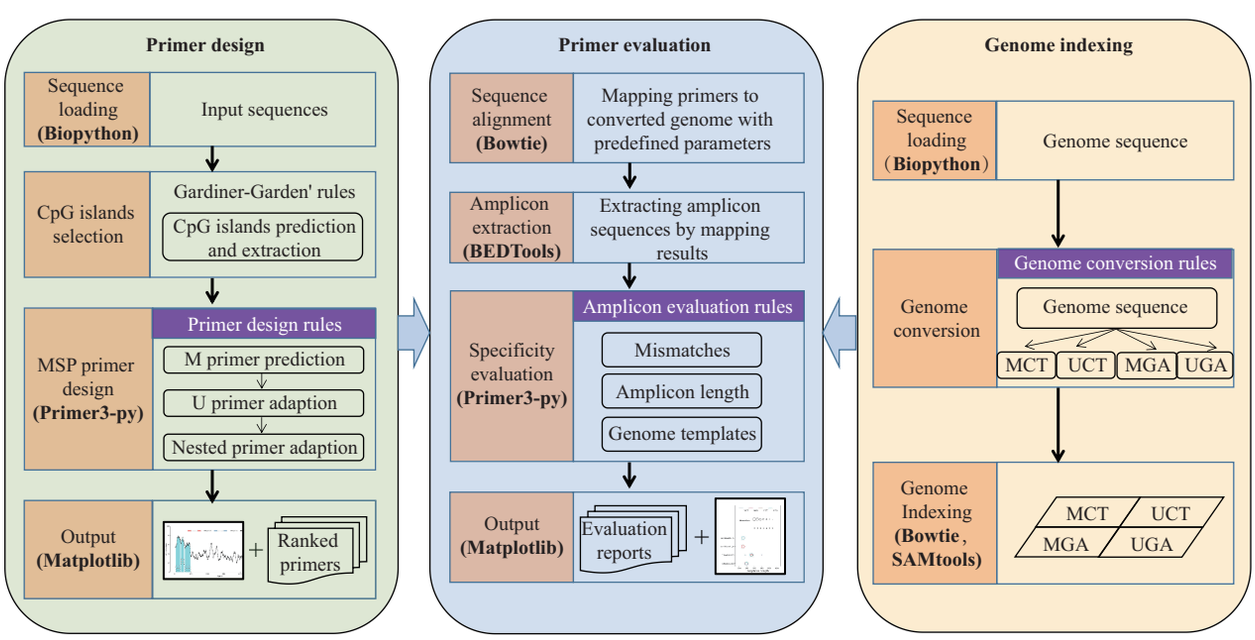

Fig. 1　The workflow of MethyScan

The workflow of MethyScan consists of three modules including primer design module, genome indexing module, and primer evaluation module. For each module, the crucial rules are labeled with purple color. Python packages and NGS tools integrated in the workflow are listed and enclosed by brackets in each main function. For a certain sequence, CpG islands will be scanned first to obtain regions of potential amplicons. All primers are designed by following the primer design rules of MethyScan. After the prediction of M primers, the U primers and Nested primers will be adapted according to the locations of M primers. Primer groups will be ranked by a series of primer properties, such as CpG number and *T*m difference. To evaluate non-specific amplification, genome sequence will be converted and indexed according to the methylation status and genome strand information. Then, Bowtie will be employed to accomplish sequence alignments. After extraction of amplicon sequences by BEDTools, primer evaluation will be performed to generate both a comprehensive result and a graphic report to show the non-specific amplification information.

则. 比如，MethyScan对输入序列进行CpG岛预测，确定引物预测区域后，根据引物设计规则进行M引物设计和U引物与巢式引物适配，并利用*T*m值、CpG数目等引物特征进行排序，最终输出引物序列及引物可视化展示. 在建立基因组索引时，MethyScan通过基因组转换规则对基因组模板进行4种序列转换，进而利用Bowtie及SAMtools构建基因组索引. 根据引物序列及基因组索引，MethyScan利用Bowtie进行序列比对，并通过BEDTools提取扩增子序列，结合引物错配碱基、扩增子长度及结合模板等信息进行分析评估，最终给出引物基因组非特异性扩增评估报告.

1.2.1　MethyScan引物设计规则及流程

针对M引物、U引物及巢式引物设计与筛选，MethyScan提出了单引物和引物组合考虑规则（表1）. 就单条引物筛选而言，规则主要包括3'端是否含有CpG序列、最大polyT数目和引物*T*m值区间. 而引物组合筛选，其规则主要包括扩增子长度区间及引物组合间*T*m差值. 值得说明的是，经亚硫酸氢盐处理后基因组模板序列会发生C→T碱基转变，此时序列复杂性变低，导致实际扩增中引物二聚体（dimer）与发夹结构（hairpin）较为常见，考虑到引物设计时难以避免上述结构，MethyScan对引物二聚体和发夹结构特征的预测并未用于引物筛选，只提供累积得分和引物输出排序.

Table 1　MethyScan design rules

Checking rules	M primer	U primer	Nested primer
3' end CpG boundary	+	–	–
Maximum polyT	+	+	+
*T*m interval	+	+	+
Amplicon length interval	+	–	+
*T*m difference for primer pair	+	+	–

MethyScan进行引物设计流程如下：

a. 基于Gardiner-Garden定义进行CpG岛预测[26].

b. 基于CpG岛进行M引物设计. 为了能与U引物明确区分，在设计M引物时，要求3' 端序列中最末3个碱基须含有CpG，以保证引物退火时与模板的特异性结合.

c. 进行U引物适配设计. 利用上述M引物设计中获得的CpG序列位置信息，根据表1所示U引物设计规则进行长度调整和适配. 在调整U引物长度时，为了能与M引物明确区分，要求保留相应M引物中CpG位置序列，以完成U引物设计.

d. 在扩增子之外25~100 bp序列区，针对每对M、U引物组合进行巢式引物适配. 除表1所示巢式引物设计规则外，在巢式适配过程中，还需确保巢式引物序列中不含有CpG. 若无法适配此规则，可在此基础上进行简并引物设计，以实现甲基化、非甲基化模板的扩增.

e. 进行引物特征排序和M、U和巢式引物序列信息输出与引物可视化展示. 在排序时，主要包含的引物特征有：M引物序列中的CpG数目；M、U和巢式引物结构特征累积得分；M引物组合*T*m值差值；M、U和巢式引物组合最大*T*m值差值以及巢式引物中所包含的CpG总数.

1.2.2　MethyScan基因组转换规则

在建立基因组索引前，需要考虑经亚硫酸氢盐处理后基因组碱基发生的变化. 因此，MethyScan建立了4种基因组转换规则用以获取转换后基因组，包括正链上由C转换为T的甲基化基因组（MCT）、正链上由C转换为T的非甲基化基因组（UCT）、负链上由C转换为T的甲基化基因组（MGA）和负链上由C转换为T的非甲基化基因组（UGA）. 由图2可以看出，MCT与UCT的差别在

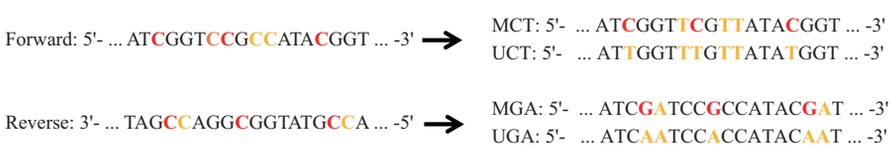

Fig. 2　Genome conversion rules in MethyScan

According to the methylation status and genome strand information, the genome sequence will be converted into 4 sequences including MCT, UCT, MGA, and UGA genome sequences. Unlike the UCT genome sequence in which all cytosines converted to thymines, methylated cytosines with red color in the MCT genome sequence will be retained and only unmethylated cytosines with orange color will be converted. Additional step for the MGA and UGA genome sequence will be performed by transforming to the forward strand after similar conversions like the MCT and UCT genome sequences, respectively.

于CG和TG的区别，MGA与UGA的差别在于CG和CA的区别. MCT和UCT反映了正链上的转换，而MGA和UGA反映了负链上的转换. 经过转换后，MethyScan调用Bowtie和SAMtools建立基因组索引.

1.2.3　MethyScan引物评估流程

MethyScan通过序列比对完成引物全基因组非特异性扩增评估，评估流程如下：

a. MethyScan通过二代测序比对软件Bowtie完成引物与4种转换后基因组序列比对. 为了获取更优比对结果，MethyScan调整比对参数“-y -l 5 -n 3 -e 90”，可以有效找出单条引物3个及以下错配比对信息.

b. MethyScan解析引物序列比对结果，获取扩增子及相应基因组坐标信息，并利用BEDTools提取扩增子序列.

c. MethyScan提取引物错配碱基数目、扩增子长度及基因组模板等信息用于引物扩增评估，同时计算引物*T*m值、GC含量、引物二聚体、发夹结构等基本信息用于最终评估报告输出.

d. MethyScan分别以文本及图形格式输出引物扩增评估报告.

1.3　MethyScan下载安装及命令行示例说明

1.3.1　MethyScan下载安装

MethyScan可运行于安装了Python3环境的Linux操作系统. 安装过程如下.

首先，采用git命令从GitHub克隆MethyScan全部源代码，终端/命令窗口执行命令为：

git clone https://github. com/bioinfo-ibms-pumc/MethyScan.git

上述克隆命令将获得一个源代码目录，目录中包含一个程序入口文件代码（MethyScan.py）及3个库文件代码（formatdb.py、utils.py和searchdb.py）及其他示例序列信息文件等.

获得上述文件目录后，可在克隆目录下执行Designprimer、Formatdb和Searchdb命令，从而分别实现引物设计、基因组索引和引物评估功能操作. 相应的命令帮助可通过“-h”进行获取，即：

python MethyScan.py designprimer -h

python MethyScan.py formatdb -h

python MethyScan.py searchdb -h

除此之外，https://github. com/bioinfo-ibms-pumc/MethyScan提供了网页版详细帮助文档及命令行示例信息.

由于MethyScan运行依赖Bowtie、SAMtools和BEDTools，及Python图形库Matplotlib与第三方Python库BioPython和Primer3-py，因此，还需进行如下的操作：

其中，Bowtie、SAMtools和BEDTools可通过Conda软件包管理系统安装，相应的命令如下：

conda install bowtie samtools bedtools

而第三方Python库可通过pip管理工具进行安装，命令如下：

pip3 install biopython matplotlib primer3-py

1.3.2　Designprimer引物设计命令示例说明

Designprimer命令主要用于M、U和巢式引物设计，包含参数涉及引物长度、*T*m值区间和引物*T*m差值等. 不仅如此，Designprimer命令还支持Fasta格式DNA序列输入文件. 下面的命令展示了对TFPI-2基因的引物设计示例. 其中，输入文件为tfpi2.fa，输出文件前缀名为result：

python3 MethyScan.py designprimer -i tfpi2.fa -o result

上述命令执行后，获得结果输出文件共包含3种：（1）result.txt，该文件以制表符分隔来呈现引物名称、5'引物序列和3'引物序列. 其中，引物名称由引物序号、CpG起始终止位置和引物类型三部分组成，诸如“p2_CpG-73-340_U”（第二组U引物p2，其中CpG岛起始位置为73 bp，终止位置为340 bp）；（2）result.detail.txt，该文件包含每个CpG岛中引物及扩增子序列详细信息；（3）xxx.pdf，该文件为图形文件，其命名含有基因名称，用以标识引物在xxx序列中位置信息.

1.3.3　Formatdb数据库索引命令示例说明

Formatdb命令主要用于参考基因组序列转换. 该命令支持Fasta格式的参考基因组序列文件，并利用Bowtie子命令bowtie-build和SAMtools子命令faidx建立基因组索引文件. 下面的命令展示了基于人类参考基因组序列建立索引的过程. 其中，输入文件为hg19.fasta，输出文件前缀名为mydb，参数-t表示多线程加速：

python3 MethyScan.py formatdb -i hg19.fasta -o mydb -t 12

上述命令执行后，获得结果输出包括MCT、MGA、UCT、UGA四种类型基因组序列索引文件类型，其中每种类型有7个文件，一共28个文件. 以MCT文件类型为例，包括mydb.MCT.1.ebwt、mydb. MCT. 2. ebwt、 mydb. MCT. 3. ebwt、 mydb.

MCT.4.ebwt、mydb.MCT.rev.1.ebwt、mydb.MCT.rev.2.ebwt和mydb.MCT.fai文件. 在评估引物非特异性扩增时，只需指定前缀名mydb即可，程序将自动搜索相应索引文件.

1.3.4　Searchdb引物评估命令示例说明

Searchdb命令主要用于评估引物在基因组上特异性-非特异性扩增信息，包含参数涉及扩增子长度和引物序列错配碱基数目等. Searchdb命令支持输入格式为Designprimer命令产生的文件格式，即以制表符分隔的包含引物名称、5'引物序列和3'引物序列三列信息文本文件. 下面的命令展示了利用**1.3.3**中人类参考基因组索引，对**1.3.2**中引物进行评估的示例. 其中，输入引物文件为result.txt，索引文件为mydb，输出文件前缀名为primer_res，参数-t表示多线程加速，-k表示转换后不同类型参考基因组：

python MethyScan.py searchdb -i result.txt -d mydb -o primer_res -t 12 -k ALL

上述命令执行后，获得结果输出文件一共有两个，一个为primer_res.txt文件，包括引物扩增子概括信息（以Amplicon Summary为开始的文本内容）和详细扩增信息（以Amplicon Detail为开始的文本内容）；另一个为primer_res.pdf图形文件，用以比较评估引物非特异性扩增.

2　结　　果

2.1　TFPI−2基因引物设计

TFPI-2基因甲基化被认为是早期食管癌潜在标志物之一. MethyScan在TFPI-2基因序列中预测出3个CpG岛，并对每个CpG岛设计了3组M、U和巢式引物（图3）. 引物序列详细信息见网络版附件（PIBB20200413Sup1_Tables S1−3.xlsx）.

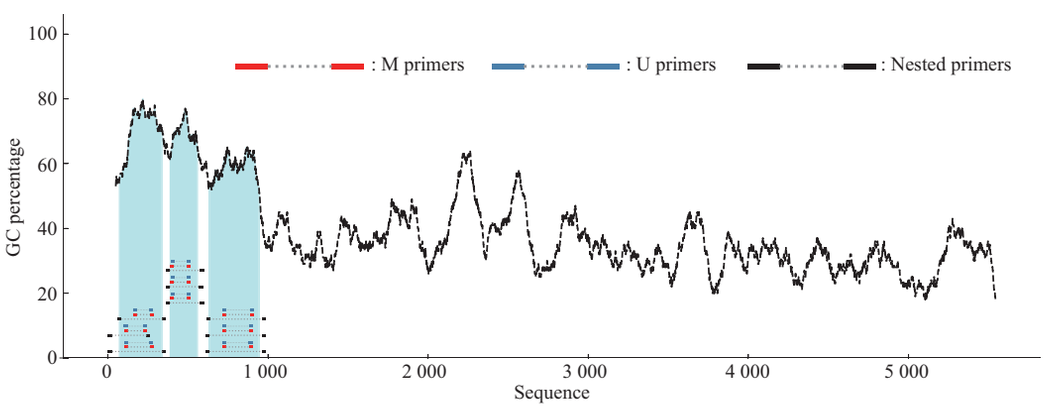

Fig. 3　The primers for TFPI−2 gene designed by MethyScan

Three CpG islands predicted in the TFPI-2 gene promoter were displayed with cyan color. For each island, three primer groups were designed by MethyScan. The M primers, U primers, and Nested primers in a primer group were shown as red, blue and black rectangles linked with dotted lines.

2.2　TFPI−2基因引物非特异性扩增评估

亚硫酸氢盐处理后的基因组序列复杂度降低，这是导致引物在基因组其他区域形成非特异性扩增，并影响目的条带判定和PCR结果解读的主要原因之一. 在引物设计中，有效预估非特异性扩增信息将为目的基因扩增和甲基化鉴定带来实际价值. 为直观展示TFPI-2基因引物在全基因组上的非特异性扩增信息，通过提取所有引物在4种转换基因组中的扩增子序列（长度位于500 bp之内），MethyScan可将错配信息、基因组模板类型和扩增子长度进行图形可视化展示. 如图4所示，扩增子以圆圈表示. 若引物匹配至MCA、UCA模板，表明扩增子序列位于基因组正链；若匹配至MGA、UGA模板，则表明扩增子序列位于基因组负链. 对于巢式引物而言，这种匹配还需要同时保证在甲基化、非甲基化模板的扩增. 利用图形可视化展示引物的非特异性扩增信息，可以从理论的角度给出引物选择及其后续实验优化策略. 例如：若观察到扩增子长度与目的片段长度有较大区别时，实验中需注意电泳胶图中条带的区分；若观察到扩增子长度与目的片段长度差异较小而错配数目却较大时，可以通过调节*T*m值去除非特异性结合. 总之，对引物进行选择时，首选非特异性扩增片段较少的引物组合，其次选择非特异性扩增与目的片段长度差异

较大、错配数目差异较大的引物组合，最后避免选择长度相同且错配数目差异较小的引物组合. 通过分析可知，图4中p2_CpG-73-340的M、U及巢式引物组合为最优组合.

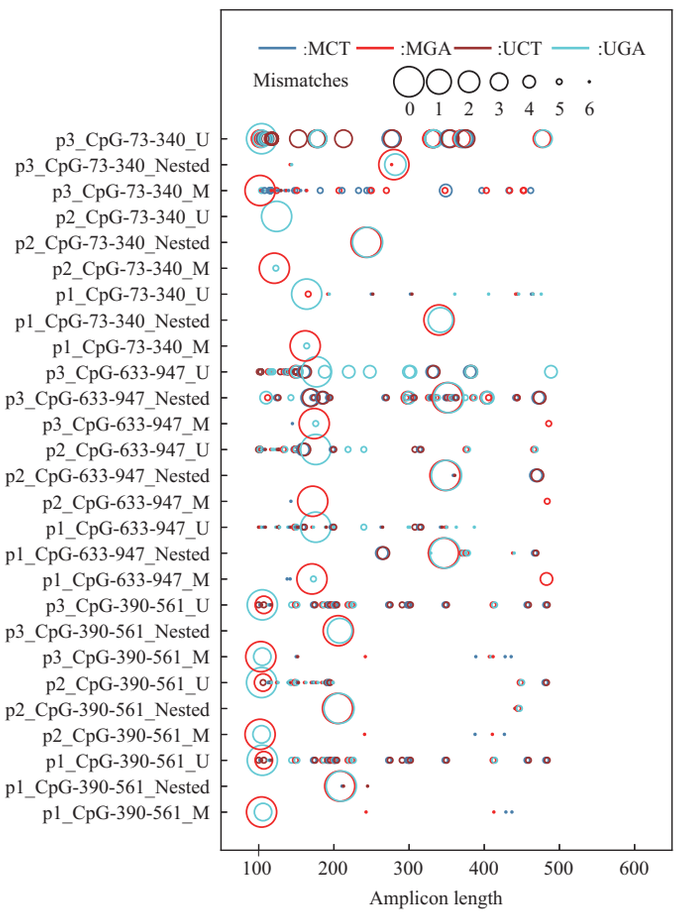

Fig. 4　The evaluation of primer sequences for TFPI−2 gene designed by MethyScan

Mismatch, genome template, and amplicon length information were evaluated by MethyScan. Each circle represented a potential amplicon while the size and color of the circle represented mismatch and genome template information, respectively. As a typical primer group, p2_CpG-73-340 gained the best specificity for U primers and Nested primers, though only a non-specific amplification might happen on the UGA genome sequence with 5 mismatches for M primers.

2.3　MethyScan与其他MSP引物设计软件的比较

2.3.1　与MethPrimer引物设计比较

为有效检测MethyScan设计引物的性能及稳定性，本研究选取食管癌、结直肠癌等多种恶性肿瘤中6个潜在生物标志物（TFPI-2、CDKN2A、CD44、CASP8、NDRG4和SDHD）的基因启动子区域1 000 bp序列，用于MethPrimer和MethyScan在MSP引物设计方面的比较. 为了确保比较的公正性，研究对软件所涉及的*T*m值区间、扩增长度以及引物末端CpG数目等主要参数进行了相同设置.

通过分析，除了SDHD和NDRG4基因外，其余基因经MethyScan和MethPrimer设计的引物位置基本一致（详细信息见网络版附件PIBB20200413 Sup1_Tables S1−3.xlsx）. 但通过对引物*T*m值、GC含量及引物中CpG数目的分析（图5），可以看出，无论是引物*T*m值差值还是GC含量差值，MethyScan均小于MethPrimer. 而且，经MethyScan设计的M引物还含有更多CpG位点数目. 这不仅为后续实验调整带来了便利，还能够增加M引物与U引物与各自模板的特异性结合.

2.3.2　引物非特异性扩增的理论性验证

本研究选取了MethBlast、BiSearch研究论文所涉及的6对引物进行非特异性扩增的理论性验证

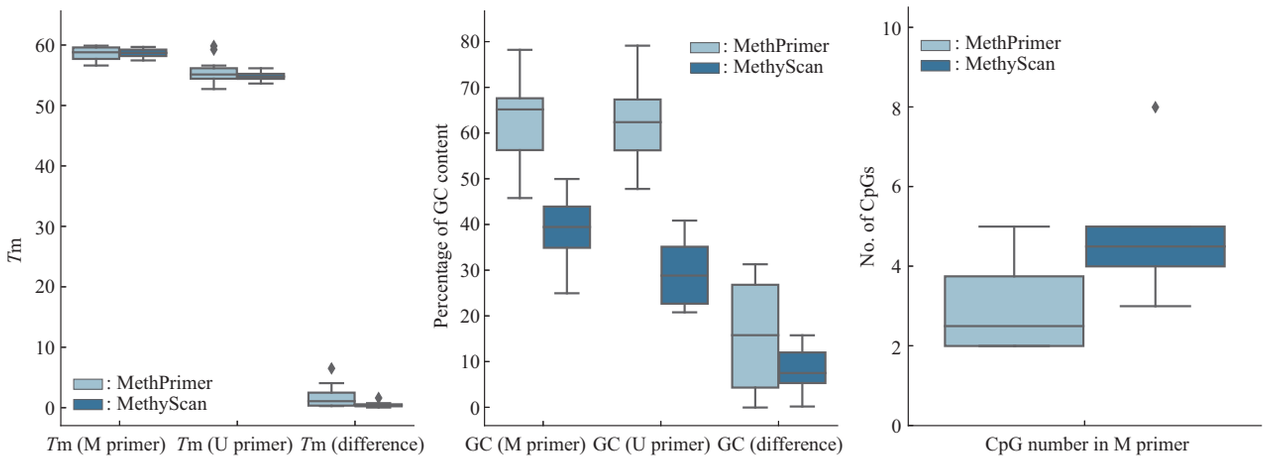

Fig. 5　The comparison of primer properties designed by MethyScan and MethPrimer

Both the M primers and U primers of six genes were designed by MethyScan and MethPrimer under the same parameters. Tm values and GC content properties of the primers were analyzed by using python package Primer3-py. The M primers and U primers were merged to calculate difference of *T*m values and GC content. The analysis of the CpG number in M primers was also included in the comparison.

（引物详细信息见网络版附件PIBB20200413Sup1_Tables S1−3.xlsx），引物组合名称分别为methblast11_U、methblast11_M、methblast15_U、methblast15_M、THbs114_308和THbs566_827. 经过MethyScan评估发现，methblast11_U、methblast11_M引物组合可在MGA和UGA模板上扩增唯一条带（图6），而methblast15_U、methblast15_M引物组合却无法扩增（由于引物无法与模板结合，因此图6中无结果显示）. 这一结果与MethBlast相应的论文研究结果一致. 另外，在MethyScan评估报告中显示，相较于THbs114_308引物扩增所获得的唯一条带（195 bp），THbs566_827除扩增目的条带（263 bp）外，还有其他非特异性扩增条带出现（包括282 bp、177 bp、187 bp和396 bp），形成BiSearch论文中所述弥散（smear）现象.

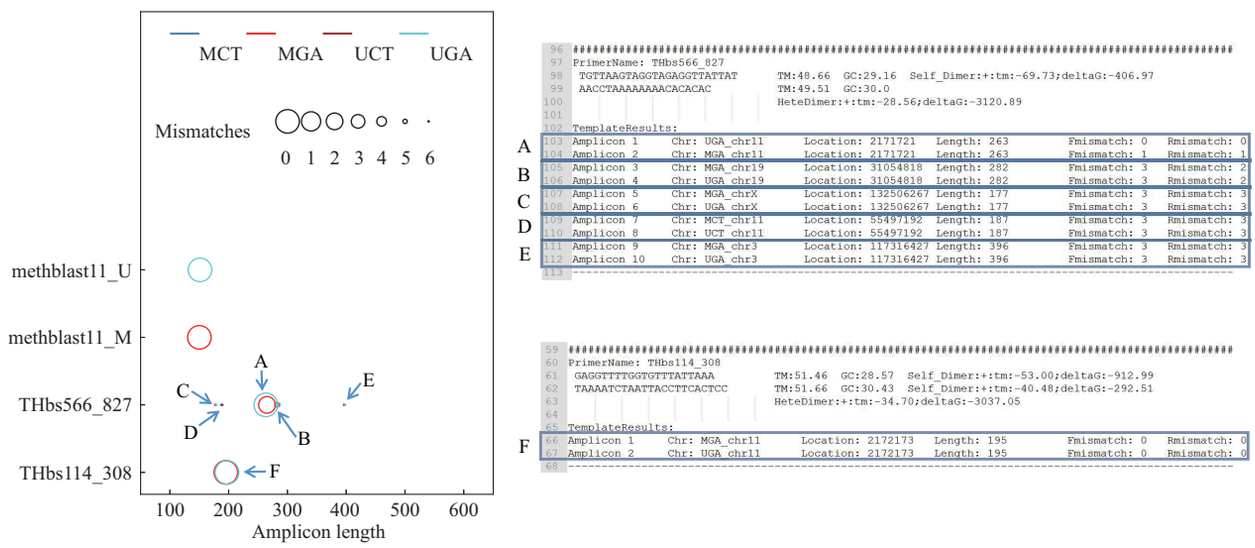

Fig. 6　Theoretical verification of primer non−specific amplification from Methblast and BiSearch articles

The non-specific amplification information of six primers from Methblast and BiSearch articles was analyzed by MethyScan. Unlike the specific amplification of the methblast11 primer group, the amplification of the methblast15 primer group was failed and not shown. Specific/non-specific amplifications were marked with an alphabet for primer group THbs566_827 and THbs114_308. Both the phenomenon of smear and distinct bands mentioned in the BiSearch article could be deduced by MethyScan.

3 结　论

本研究针对MSP引物设计中存在的引物影响考虑因素、设计与评估并行处理及特异性扩增预测等问题，开发了具有引物设计、基因组索引和引物评估三大功能的甲基化特异性PCR引物设计与评估可视化工具MethyScan. 通过CpG岛预测，M引物、U引物与巢式引物规则设计，以及*T*m值、引物二聚体和发夹结构等特征综合评估，MethyScan能够有效完成MSP引物设计和适配. 不仅如此，通过基因组转换规则的利用和基因组索引，MethyScan实现了对Bowtie、SAMtools及BEDTools等工具的有效综合整合，从而完成基因组引物扩增信息呈现和非特异性扩增信息与目的扩增片段差异信息的可视化展示，并最终实现MSP引物特异性-非特异性扩增的简明、准确评估. 与其他软件相比，MethyScan在引物包含CpG位点数目及*T*m值差异估算设计上更具优势，极大简化了实际PCR实验条件的优化调整. 作为首个特异性-非特异性扩增MSP引物设计图形化工具，MethyScan可有效提高甲基化引物设计准确性.

4 讨　论

DNA甲基化是重要的表观遗传现象，在基因表达中发挥重要调控作用. 研究表明，肿瘤抑制基因启动子的高甲基化与疾病密切相关，为此，基于MSP技术进行甲基化诊断试剂盒研发和开展基因DNA甲基化检测具有重要的临床诊断价值. 在早期肿瘤的甲基化检测中，由于目的基因甲基化片段浓度非常低，再加之样本经亚硫酸氢盐处理的损耗，这导致MSP技术的检测敏感性大大降低. 作为一种样本富集手段，巢式引物的加入是增加检测敏感性的重要方式. 本研究通过对6个常见恶性肿瘤甲基化标志基因的M引物、U引物及巢式引物设计，突显了MethyScan在MSP技术中的引物设计优势，并为基因甲基化临床诊断试剂盒的研发提供了参考.

此外，避免引物假阳性扩增是保证MSP实验成功的重要前提之一. 目前，大多数MSP引物设计软件基于引物与基因组序列比较并获取相应扩增信息，从而进行非特异性扩增评估. 在比较引物与基因组序列时主要包括字符串匹配和序列比对两种方式. 相对字符串匹配方式，序列比对能够通过复杂参数设定以获取更高敏感性. 但对于常规引物序列，由于其长度较短，使用二代测序比对软件Bowtie比传统BLAST更适合. 此外，常见引物评估软件对非特异性扩增信息的展示内容仅限于提供引物与扩增片段序列匹配文本信息，既没有扩增片段的比较，也没有图形化展示，很难得知更优引物组合. 本研究通过Bowtie进行序列比对并图形化显示错配碱基数目、不同转换模板及扩增长度等多种信息，可简单明了地观察引物非特异性扩增与目的片段的差异，从而选择更优引物组合，为后续PCR实验条件的调整及DNA甲基化检测项目的研发提供有力的支撑.

附件　PIBB20200413Sup1_Tables S1−3.xlsx见本文网络版(http://www.ibp.ac.cn或http://www.cnki.net)

参　考　文　献

[1] Goldberg A D, Allis C D, Bernstein E. Epigenetics: a landscape takes shape. Cell, 2007, **128**(4): 635-638

[2] Bird A P. CpG islands as gene markers in the vertebrate nucleus. Trends in Genetics, 1987, **3**(12): 342-347

[3] Ehrlich M, Gama-Sosa M A, Huang L H, *et al*. Amount and distribution of 5-methylcytosine in human DNA from different types of tissues of cells. Nucleic Acids Res, 1982, **10**(8): 2709-2721

[4] Law J A, Jacobsen S E. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nat Rev Genet, 2010, **11**(3): 204-220

[5] Baylin S B. DNA methylation and gene silencing in cancer. Nat Clin Pract Oncol, 2005, **S1**: S4-S11

[6] Wen L, Tang F. Human germline cell development: from the perspective of single-cell sequencing. Mol Cell, 2019, **76**(2): 320-328

[7] Devos T, Tetzner R, Model F, *et al*. Circulating methylated SEPT9 DNA in plasma is a biomarker for colorectal cancer. Clin Chem, 2009, **55**(7): 1337-1346

[8] Imperiale T F, Ransohoff D F, Itzkowitz S H, *et al*. Multitarget stool DNA testing for colorectal-cancer screening. N Engl J Med, 2014, **370**(14): 1287-1297

[9] Jia Y, Yang Y, Brock M V, *et al*. Methylation of TFPI-2 is an early event of esophageal carcinogenesis. Epigenomics, 2012, **4**(2): 135-146

[10] Chang X, Yamashita K, Sidransky D, *et al*. Promoter methylation of heat shock protein B2 in human esophageal squamous cell carcinoma. Int J Oncol, 2011, **38**(4): 1129-1135

[11] Dippmann C, Schmitz M, Wunsch K, *et al*. Triage of hrHPV-positive women: comparison of two commercial methylation-specific PCR assays. Clin Epigenetics, 2020, **12**(1): 171-177

[12] Steenbergen R D, Snijders P J, Heideman D A, *et al*. Clinical implications of (epi)genetic changes in HPV-induced cervical

precancerous lesions. Nat Rev Cancer, 2014, **14**(6): 395-405

[13] Hao X, Luo H, Krawczyk M, *et al*. DNA methylation markers for diagnosis and prognosis of common cancers. Proc Natl Acad Sci USA, 2017, **114**(28): 7414-7419

[14] Herman J G, Graff J R, Myohanen S, *et al*. Methylation-specific PCR: a novel PCR assay for methylation status of CpG islands. Proc Natl Acad Sci USA, 1996, **93**(18): 9821-9826

[15] Ramalho-Carvalho J, Henrique R, Jeronimo C. Methylation-specific PCR. Methods Mol Biol, 2018, **1708**: 447-472

[16] Huang Z, Bassil C F, Murphy S K. Methylation-specific PCR. Methods Mol Biol, 2013, **1049**: 75-82

[17] Ku J L, Jeon Y K, Park J G. Methylation-specific PCR. Methods Mol Biol, 2011, **791**: 23-32

[18] Licchesi J D, Herman J G. Methylation-specific PCR. Methods Mol Biol, 2009, **507**: 305-323

[19] Nuovo G J. Methylation-specific PCR *in situ* hybridization. Methods Mol Biol, 2004, **287**: 261-272

[20] Ohashi H. Methylation-specific PCR. Methods Mol Biol, 2002, **192**: 91-97

[21] Palmisano W A, Divine K K, Saccomanno G, *et al*. Predicting lung cancer by detecting aberrant promoter methylation in sputum. Cancer Res, 2000, **60**(21): 5954-5958

[22] Lo Y M, Wong I H, Zhang J, *et al*. Quantitative analysis of aberrant p16 methylation using real-time quantitative methylation-specific polymerase chain reaction. Cancer Res, 1999, **59**(16): 3899-3903

[23] Fackler M J, Mcveigh M, Mehrotra J, *et al*. Quantitative multiplex methylation-specific PCR assay for the detection of promoter hypermethylation in multiple genes in breast cancer. Cancer Res, 2004, **64**(13): 4442-4452

[24] Husseiny M I, Fahmy A, Du W, *et al*. Development of quantitative methylation-specific droplet digital PCR (ddMSP) for assessment of natural tregs. Front Genet, 2020, **11**(300): 1-11

[25] Li L C, Dahiya R. MethPrimer: designing primers for methylation PCRs. Bioinformatics, 2002, **18**(11): 1427-1431

[26] Gardiner-Garden M, Frommer M. CpG islands in vertebrate genomes. J Mol Biol, 1987, **196**(2): 261-282

[27] Brandes J C, Carraway H, Herman J G. Optimal primer design using the novel primer design program: MSPprimer provides accurate methylation analysis of the ATM promoter. Oncogene, 2007, **26**(42): 6229-6237

[28] Tusnady G E, Simon I, Varadi A, *et al*. BiSearch: primer-design and search tool for PCR on bisulfite-treated genomes. Nucleic Acids Res, 2005, **33**(1): e9

[29] Pattyn F, Hoebeeck J, Robbrecht P, *et al*. methBLAST and methPrimerDB: web-tools for PCR based methylation analysis. BMC Bioinformatics, 2006, **7**: 496-504

[30] Langmead B, Trapnell C, Pop M, *et al*. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol, 2009, **10**(3): R25

[31] Li H, Handsaker B, Wysoker A, *et al*. The Sequence Alignment/Map format and SAMtools. Bioinformatics, 2009, **25**(16): 2078-2079

[32] Quinlan A R, Hall I M. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics, 2010, **26**(6): 841-842

MethyScan：A Tool for Methylation Specific PCR Primer Design and Evaluation

CAO Ying-Hao*

(*Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences & School of Basic Medicine, Peking Union Medical College, Beijing* 100730, China)

Abstract　DNA methylation is an important epigenetic phenomena and plays crucial roles in the gene regulation. Many studies showed that DNA methylation can be used as clinical diagnostic biomarker. However, the ability to detect the DNA methylation status quickly and accurately is a prerequisite and key point for clinical application. By using two kinds of primers which can bind to methylated and unmethylated template respectively, methylation specific PCR (MSP) can distinguish DNA methylation status and prove to be a feasible and convenient diagnostic technique in clinical practice. Unlike traditional PCR, MSP mainly has four difficulties: how to enhance the specificity of binding to primer-methylated/unmethylated template, how to reduce the difference of *T*m value of primer sequences, how to remove false positive amplification and how to improve sensitivity. Though most MSP primer design tools have proposed various solutions for those difficulties, there are still some defects in consideration of primer influencing factors, multitasking, prediction of specific amplification in MSP primer design and evaluation. Therefore, in this study, after deep exploration of existing MSP primer design tools, a novel MSP primer design and graphic evaluation tool named MethyScan was developed based on the integration of Bowtie, SAMtools, and BEDTools with Python graphic library Matplotlib and third-party functional libraries Biopython and Primer3-py. Three functional modules were involved in MethyScan including primer design, genome indexing and primer evaluation. MethyScan not only has the ability to perform MSP primers design and Nested primers adaptation, but also can evaluate primers specific/non-specific amplification with the analysis of primer binding information on four converted genomic templates and graphically displaying of the difference between non-specific amplification and target. Meanwhile, the comparison of MSP primer design for six potential biomarkers TFPI-2, NDRG4, CDKN2A, CD44, CASP8, and SDHD in esophageal cancer, colorectal cancer, and other malignant tumors suggested that MethyScan can not only obtain primers with more CpG sites, but also obtain primers with same or similar locations to those of other softwares, and the difference of Tm values of primers is even smaller. As the first MSP primer design tool for graphically displaying specific/non-specific amplification differences, MethyScan can effectively improve the accuracy of methylation primer design and provides strong support for the development of clinical DNA methylation detection projects, tests and diagnostic kits. The download address of MethyScan is: https://github.com/bioinfo-ibms-pumc/MethyScan.

Key words　primer design, DNA methylation, primer evaluation, MSP primer

DOI：10.16476/j.pibb.2020.0413

* Corresponding author.

Tel：86-10-69156963，E-mail：yhcao@ibms.pumc.edu.cn

Received：November 25，2020　Accepted：February 4，2021