·基因组学与肿瘤风险预测·

基于常见遗传变异和传统风险因素的中国南方汉族人群结直肠癌风险预测模型研究

李娇元 常江 朱颖 杨洋 龚雅洁 柯俊涛 娄娇
钟荣 龚静 夏肖萍 缪小平

【摘要】 **目的** 基于结直肠癌全基因组关联研究(GWAS)发现的易感位点,联合传统风险因素建立中国南方汉族人群结直肠癌风险预测模型。**方法** 对1 066例结直肠癌患者和3 880例健康对照的21个GWAS候选位点进行基因分型,分析其与结直肠癌易感性之间的关联。通过遗传风险评分(GRS)和加权遗传风险评分(wGRS)计算显著候选位点的联合效应。以不同方式组合遗传风险评分和传统风险因素,构建结直肠癌风险预测模型,并绘制受试者工作特征曲线评价模型优劣性。**结果** 7个候选位点与结直肠癌易感性显著相关。随着风险评分的升高,人群患结直肠癌的风险也随之升高(GRS:P=0.002 6,wGRS:P<0.000 1),相比于四分位分组中最低一组,GRS和wGRS最高的一组*OR*值分别为1.33(95%*CI*:1.12~1.58,P=0.001 0)和1.76(95%*CI*:1.45~2.14,P<0.000 1)。联合传统风险因素和wGRS的模型为最优模型,其曲线下面积为0.593(95%*CI*:0.573~0.613)。**结论** 结直肠癌易感位点间存在显著的联合作用。相比于传统风险因素模型,传统风险因素结合加权遗传风险评分模型能更好预测结直肠癌的患病风险。

【关键词】 结直肠癌;单核苷酸多态性;风险预测

Risk prediction of colorectal cancer with common genetic variants and conventional non-genetic factors in a Chinese Han population

Li Jiaoyuan[1], *Chang Jiang*[2], *Zhu Ying*[1], *Yang Yang*[1], *Gong Yajie*[1], *Ke Juntao*[1], *Lou Jiao*[1], *Zhong Rong*[1], *Gong Jing*[1], *Xia Xiaoping*[3], *Miao Xiaoping*[1]. 1 *Department of Epidemiology and Biostatistics, School of Public Health*, 2 *Key Laboratory of Environment and Health, Ministry of Education and Ministry of Environmental Protection, Tongji Medical College, Huazhong University of Science and Technology, Wuhan* 430030, China; 3 *Clinical Laboratory of the Fourth Affiliated Hospital, Zhejiang University School of Medicine*

Corresponding authors: *Xia Xiaoping*, Email: xiaopingxia@sina.com; *Miao Xiaoping*, Email: miaoxp@hust.edu.cn

This work was supported by grants from the National Natural Science Foundation of China (No. 81171878, No. 81222038), *Fok Ying-Tong Education Foundation of China* (No. 131038), *Excellent Youth Foundation of Hubei Scientific Committee* (No. 2012FFA011) *and Public Project of Science and Technology Department, Zhejiang Province* (No. 2014C33264).

【**Abstract**】 **Objective** To understand the association between multiple genetic loci identified by genome-wide association studies (GWASs) and colorectal cancer (CRC) risk, and whether these genetic factors, along with traditional risk factors, could contribute to the colorectal cancer risk prediction in a Chinese Han population. **Methods** A case-control study (1 066 CRC cases and 3 880 controls) was initially conducted to assess the association between 21 recently discovered single-nucleotide polymorphisms (SNPs) and CRC risk. Genetic risk score (GRS) and weighted genetic risk score (wGRS) were calculated to evaluate the joint effects of selected loci. Multiple models combining

DOI:10.3760/cma.j.issn.0254-6450.2015.10.003

基金项目:国家自然科学基金(81171878,81222038);教育部霍英东青年教师基金(131038);湖北省杰出青年基金(2012FFA011);浙江省科技厅公益项目(2014C33264)

作者单位:430030 武汉,华中科技大学同济医学院公共卫生学院流行病与卫生统计学系(李娇元、朱颖、杨洋、龚雅洁、柯俊涛、娄娇、钟荣、龚静、缪小平),环境与健康教育部重点实验室(常江);浙江大学医学院附属第四医院检验科(夏肖萍)

通信作者:夏肖萍,Email:xiaopingxia@sina.com;缪小平,Email:miaoxp@hust.edu.cn

genetic and non-genetic factors were established and receiver operating characteristic curve analysis was used to compare the discriminatory power of different predictive models. **Results** There were 7 SNPs significantly associated with CRC susceptibility. As the GRS or wGRS increased, the risk of CRC also increased (trend P=0.002 6 for GRS, trend P<0.000 1 for wGRS). The *OR*s for highest versus lowest quartile of GRS and wGRS were 1.33 (95% *CI*: 1.12–1.58, P=0.001 0) and 1.76 (95% *CI*: 1.45–2.14, P<0.000 1), respectively. The model incorporating wGRS and traditional risk factors, including sex, age, smoking and drinking, was the best one to predict CRC risk in this population, with an area under curve of 0.593 (95% *CI*: 0.573–0.613). **Conclusion** Multiple genetic loci identified by GWASs jointly influenced the CRC risk. The combination of genetic factors and conventional non-genetic factors improved the performance of risk predictive model for colorectal cancer.

【**Key words**】 Colorectal cancer; Single-nucleotide polymorphism; Risk prediction

结直肠癌是世界范围内常见的消化系统恶性肿瘤[1]。近年来我国结直肠癌的发病率和死亡率均呈平稳上升趋势[2]。有效的结直肠癌风险预测模型是用于结直肠癌高危人群筛查，实现早发现、早预防和早治疗的有效手段。已有的结直肠癌预测模型主要探讨环境因素在风险预测中的作用[3-4]。而结直肠癌的发生是多种环境和遗传因素共同作用的结果，后者在肿瘤形成和发展中也扮演着重要角色[5]。在全基因组关联研究（GWAS）成功应用于肿瘤遗传研究的背景下，GWAS发现的遗传易感位点赋予了风险预测模型新的含义。为此本研究旨在利用GWAS发现的中国南方汉族人群结直肠癌易感位点，联合性别、年龄、吸烟、饮酒等常见风险因素建立结直肠癌风险预测模型，为构建适合中国人群的结直肠癌早期预警预测体系提供理论参考依据。

对象与方法

1. 研究对象：共纳入4 946人，其中结直肠癌1 066例（病例组），正常对照3 880例（对照组）。分别来自2009年1月至2012年12月在华中科技大学同济医院、协和医院确诊的结直肠癌病例和同期随机抽取协和医院的健康体检人群。病例组纳入标准为汉族，并经组织病理学确诊的结直肠癌患者，排除非原发性结直肠癌及其他恶性肿瘤者。对照组均通过基本体检确认无肿瘤病史和症状。研究对象的人口学资料（包括性别、年龄、吸烟和饮酒状态）来自病历或体检报告。从未吸或平均每天吸<1支且吸烟时间<1年者定义为非吸烟者，否则为吸烟者；一生中每周饮酒>2次且持续>1年为饮酒者，否则为非饮酒者。研究对象在参加该项目前均签署知情同意书，该项目经华中科技大学伦理委员会批准。

2. 遗传变异位点的选择与基因分型：通过查阅NHGRI GWAS数据库中相关文献[6]，去除连锁不平衡的位点，选择了21个在东亚人群中与结直肠癌相关的常见单核苷酸多态性位点（SNP）作为风险预测模型的候选位点[7-9]。在TaqMan基因分型系统（ABI 7900HT Real Time PCR System, Applied Biosystems）完成基因分型。PCR反应为5 μl体系，反应条件为95 ℃预变性10 min；95 ℃ 15 s和60 ℃ 1 min，共40个循环。由于21个候选位点中2个位点（rs10774214 和rs1665650）基因分型失败，最终19个位点纳入后续研究。排除基因分型成功率较低（<90%）的研究对象，共1 065例结直肠癌患者和3 873名对照纳入研究。

3. 遗传位点与结直肠癌易感性的关联分析：运用logistic回归中的加性模型（additive model）计算19个位点与结直肠癌易感性之间的关联。模型中校正了年龄、性别、吸烟及饮酒状态，并以P<0.05为差异有统计学意义的标准筛选纳入风险模型的位点。在19个遗传位点中，7个位点与结直肠癌具有显著关联，因而列入风险模型。

4. 结直肠癌风险预测模型的构建：本研究采用遗传风险评分（genetic risk score, GRS）计算7个显著位点的联合作用。该方法是将每个SNP位点看作独立的结直肠癌危险因素，利用0、1、2三个线性数值分别代表某个体携带某一SNP风险等位基因的个数。而位点联合作用的计算方法包括GRS和权重遗传风险评分（weighted genetic risk score, wGRS）[10]。GRS假设每个位点对结直肠癌易感性有相同贡献，个体的GRS即为其携带风险等位基因个数之和。wGRS假设每个位点对结直肠癌易感性的影响不同，但与其*OR*值有关，因此wGRS为每个SNP在logistic回归中β值加权后的平均风险等位基因个数。在此基础上，为了避免个别缺失值的影响以及便于比较，本文将wGRS分别除以β值之和的2倍并乘以SNP等位基因个数。GRS和wGRS均根据其在研究对象中的分布进行了四分位划分，分别命名为QGRS和QwGRS。所有风险评分与结直肠

癌易感性的关联均在调整了性别、年龄、吸烟、饮酒状态的情况下由非条件logistic回归计算得出。

本研究利用遗传因素（7个SNP位点）和传统风险因素（性别、年龄、吸烟和饮酒）分别构建了5种风险模型。模型1仅纳入性别、年龄、吸烟和饮酒因素；模型2在纳入以上因素的基础上加入GRS；模型3在纳入传统风险因素的基础上加入wGRS；模型4在纳入传统风险因素的基础上加入QGRS；模型5在纳入传统风险因素的基础上加入QwGRS。并采用赤池信息准则（Akaike's Information Criterion, AIC）作为评价以上模型拟合优良性的指标。利用受试者工作特征（ROC）曲线及计算曲线下面积（AUC）评价模型的预测效果。

5. 统计学方法：采用PLINK软件完成候选位点的关联分析和GRS[11]；风险模型构建在SAS 9.1软件中完成；利用SPSS 13.0软件完成ROC及AUC计算。

结　果

1. 遗传位点与结直肠癌易感性的关联：以$P<0.05$为初筛标准，19个SNP位点中7个（rs647161、rs10505477、rs6983267、rs10795668、rs7229639、rs4939827、rs2423279）与结直肠癌易感性存在显著关联（表1）。其中最显著的位点为20号染色体上的rs2423279（$OR=1.24$, 95%CI: 1.12～1.37, $P=2.77\times10^{-5}$）。为了进一步分析显著位点的联合作用，计算了7个位点的GRS和wGRS。显示病例组和对照组的GRS分布略有不同（图1），相比对照组，病例组的GRS分布偏向右移；随着GRS的升高，病例组所占的相对比例也逐渐升高。对GRS和wGRS根据四分位进行分组，并计算各组人群结直肠癌的患病风险。结果显示，随着GRS或wGRS的增加，人群罹患结直肠癌的风险也逐渐增加（表2）。以GRS≤3的人群为参照，评分4～5、6和≥7的人群罹患结直肠癌的风险分别为$OR=1.11$（95%CI: 0.95～1.35），$P=0.311\ 6$；$OR=1.02$（95% CI: 0.83～1.26），$P=0.844\ 4$和$OR=1.33$（95%CI: 1.12～1.58），$P=0.001\ 0$；趋势检验：$P=0.002\ 6$。以wGRS＜0.158的人群为参照，wGRS为0.159～0.291的人群患结直肠癌风险的$OR=1.09$（95% CI: 0.89～1.34），$P=0.408\ 6$；wGRS为0.292～0.417的人群$OR=1.40$（95% CI: 1.14～1.71），$P=0.001\ 1$；wGRS≥0.418的人群$OR=1.76$（95%CI: 1.45～2.14），$P<0.000\ 1$；趋势检验：$P<0.000\ 1$。

2. 风险预测模型构建及模型间的比较：比较5种不同方式建模的AIC值，选出拟合优度最优者（表3）。以仅纳入性别、年龄、吸烟、饮酒4种传统风险因素模型的AIC值为参考（AIC=5 091.580），表明以不同方式联合环境因素和遗传因素构建的模

表1 19个GWAS易感位点与结直肠癌易感性的关联

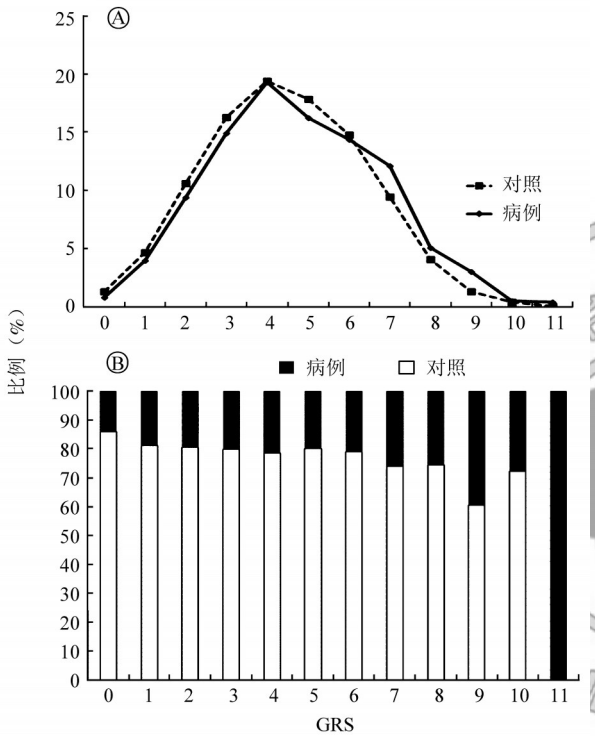
染色体号	物理位置	rs	等位基因	次要等位基因频率		*OR*值(95%*CI*)	*P*值
				病例组	对照组		
1	222164948	rs6687758	A＞G	0.222	0.212	1.05(0.93～1.18)	0.427 2
5	134489324	rs647161	A＞G	0.330	0.302	1.14(1.03～1.26)	0.014 1
8	128407443	rs10505477	G＞A	0.442	0.415	1.11(1.01～1.22)	0.036 0
8	128413305	rs6983267	A＞C	0.442	0.416	1.10(1.00～1.21)	0.049 5
8	128424792	rs7014346	G＞A	0.331	0.308	1.10(0.99～1.22)	0.063 8
10	8701219	rs10795668	G＞A	0.335	0.363	0.89(0.80～0.98)	0.023 1
10	80819132	rs704017	A＞G	0.288	0.300	0.94(0.85～1.05)	0.270 0
10	114726843	rs11196172	A＞G	0.296	0.306	0.94(0.85～1.05)	0.287 1
11	61597972	rs1535	A＞G	0.421	0.409	1.04(0.94～1.15)	0.418 9
11	61552680	rs174537	C＞A	0.422	0.410	1.04(0.94～1.15)	0.431 9
11	61571478	rs174550	A＞G	0.421	0.410	1.04(0.94～1.15)	0.437 0
11	61564299	rs4246215	C＞A	0.417	0.407	1.03(0.93～1.14)	0.551 1
12	6385727	rs10849432	A＞G	0.184	0.196	0.92(0.82～1.05)	0.211 8
17	800593	rs12603526	A＞G	0.267	0.255	1.06(0.95～1.19)	0.268 3
18	46448129	rs7229639	G＞A	0.178	0.155	1.18(1.04～1.35)	0.010 2
18	46453463	rs4939827	G＞A	0.267	0.244	1.12(1.01～1.25)	0.036 5
19	41869392	rs2241714	A＞G	0.503	0.486	1.07(0.97～1.18)	0.157 8
19	41860296	rs1800469	A＞G	0.504	0.488	1.07(0.97～1.18)	0.188 7
20	7832380	rs2423279	G＞A	0.375	0.325	1.24(1.12～1.37)	2.77×10^{-5}

注：*OR*值为调整性别、年龄、吸烟和饮酒状态

表2　GRS四分位分组（Q1～Q4）与结直肠癌易感性的关联

项目	GRS	均值	病例/对照	*OR*值(95%*CI*)[a]	*P*值	wGRS	均值	病例/对照	*OR*值(95%*CI*)[a]	*P*值
Q1	≤3	2.29	309/1 271	1.00	–	≤0.158	0.070	216/1 034	1.00	–
Q2	4	4.00	205/753	1.11(0.91～1.35)	0.311 6	0.159～0.291	0.236	236/1 014	1.09(0.89～1.34)	0.408 6
Q3	5,6	5.00	173/693	1.02(0.83～1.26)	0.844 4	0.292～0.417	0.353	277/925	1.40(1.14～1.71)	0.001 1
Q4	≥7	6.82	377/1 156	1.33(1.12～1.58)	0.001 0	≥0.418	0.526	336/900	1.76(1.45～2.14)	<0.000 1

注：[a]同表1；趋势检验*P*值：GRS为0.002 6，wGRS为<0.000 1



注：Ⓐ病例组和对照组GRS分布；Ⓑ不同GRS下病例组和对照组所占比例

图1　结直肠癌显著易感位点的遗传风险评分在病例和对照中的分布

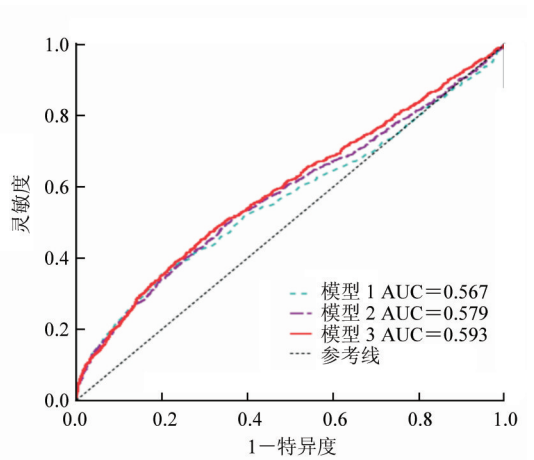


图2　不同结直肠癌风险预测模型的ROC曲线

型均优于单独纳入传统风险因素的模型，其中以传统风险因素联合wGRS的模型为拟合度最好模型，AIC=5 051.451。以不同方式建模的ROC曲线见图2，相比于仅纳入传统风险因素（模型1），在这些风险因素基础上加入GRS的模型（模型2），其AUC提升了1.2%（0.579 *vs*. 0.567）。而联合传统风险因素和wGRS的模型AUC进一步提升至0.593（95%*CI*：0.573～0.613）。综合以上结果，传统风险因素联合wGRS进入预测模型具有最优的拟合性和最好的预测效果，因此应以此方式建立结直肠癌的风险预测模型。

表3　不同风险预测模型的AIC及AUC比较

模型	纳入变量	AIC	AUC值(95%*CI*)
1	年龄, 性别, 吸烟, 饮酒	5 091.580	0.567(0.545～0.588)
2	年龄, 性别, 吸烟, 饮酒, GRS	5 075.576	0.579(0.558～0.600)
3	年龄, 性别, 吸烟, 饮酒, wGRS	5 051.451	0.593(0.573～0.613)
4	年龄, 性别, 吸烟, 饮酒, QGRS	5 084.907	0.573(0.552～0.594)
5	年龄, 性别, 吸烟, 饮酒, QwGRS	5 053.854	0.592(0.572～0.612)

讨　论

本文采用病例对照研究分析了经GWAS发现与结直肠癌易感性关联的21个亚洲人群阳性位点，显示7个SNP位点与中国南方汉族人群结直肠癌显著相关。为此利用这7个位点计算个体GRS，并联合传统风险因素以5种不同方式构建了结直肠癌风险预测模型。结果表明，基于易感位点计算得出的GRS与结直肠癌易感性显著相关。传统风险因素联合遗传因素的预测模型优于传统环境因素模型。其中性别、年龄、吸烟、饮酒4个常见风险因素和wGRS组成最优的结直肠癌风险预测模型。

目前基于GWAS的风险预测模型已经应用于2型糖尿病、前列腺癌和乳腺癌等复杂性疾病的风险预测[12-14]。本研究基于GWAS的研究结果构建了中国南方汉族人群结直肠癌的风险预测模型。值得注意的是，本研究中19个GWAS发现的易感位点只有7个验证出了阳性关联。可能的原因是样本人群存在的异质性和病例组样本量的局限所致。遗传因素联合传统风险因素构建的结直肠癌风险模型优于传统风险因素模型是本研究的主要发现之一。然而，即使是最优模型（由性别、年龄、吸烟、饮酒和wGRS组成），相比于传

统风险因素模型，其AUC仅仅提高了约3%。该结果与乳腺癌的预测模型极其相似[14-15]。原因可能是GWAS发现的易感位点均为常见变异，对疾病易感性的效应亦极其微弱（大部分GWAS发现的肿瘤易感位点的*OR*值均＜2.0，仅能解释一小部分遗传易感性）[16]。本研究中，与wGRS四分位分组中最低的一组相比，wGRS值最高的一组仅提升了约76%的结直肠癌风险，说明少数几个易感位点间的联合作用远远不足以代表全部的遗传风险，更多的结直肠癌易感位点还有待进一步探讨。

本研究存在不足。病例对照研究在收集人口学及环境暴露资料时可能存在信息偏倚。影响结直肠癌的环境因素多种多样[17-20]，而本研究只探讨了性别、年龄、吸烟及饮酒4种常见因素。由于GWAS的局限性，纳入的遗传因素也非常有限。此外，本研究人群为南方汉族人群，其存在的人群异质性使得结果的外推受到一定限制。因此在大规模多中心的前瞻性队列研究中，构建多种环境因素和遗传因素组成的风险预测模型将是今后的研究方向。

参 考 文 献

[1] Torre LA, Bray F, Siegel RL, et al. Global cancer statistics, 2012 [J]. CA Cancer J Clin, 2015, 65(2): 87-108.

[2] Chen Q, Liu ZC, Cheng LP, et al. An analysis of incidence and mortality of colorectal cancer in China, 2003-2007 [J]. Chin Cancer, 2012, 21(3): 179-182. (in Chinese)
陈琼，刘志才，程兰平，等. 2003—2007年中国结直肠癌发病与死亡分析[J]. 中国肿瘤，2012，21(3)：179-182.

[3] Ma EB, Sasazuki S, Iwasaki M, et al. 10-year risk of colorectal cancer: development and validation of a prediction model in middle-aged Japanese men [J]. Cancer Epidemiol, 2010, 34(5): 534-541.

[4] Steffen A, MacInnis RJ, Joshy G, et al. Development and validation of a risk score predicting risk of colorectal cancer [J]. Cancer Epidemiol Biomarkers Prev, 2014, 23(11): 2543-2552.

[5] Brenner H, Kloor M, Pox CP. Colorectal cancer [J]. Lancet, 2014, 383(9927): 1490-1502.

[6] Hindorff LA, Sethupathy P, Junkins HA, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits [J]. Proc Natl Acad Sci USA, 2009, 106(23): 9362-9367.

[7] Zhang B, Jia WH, Matsuda K, et al. Large-scale genetic study in East Asians identifies six new loci associated with colorectal cancer risk [J]. Nat Genet, 2014, 46(6): 533-542.

[8] Jia WH, Zhang B, Matsuo K, et al. Genome-wide association analyses in East Asians identify new susceptibility loci for colorectal cancer [J]. Nat Genet, 2013, 45(2): 191-196.

[9] Zhang B, Jia WH, Matsuo K, et al. Genome-wide association study identifies a new SMAD7 risk variant associated with colorectal cancer risk in East Asians [J]. Int J Cancer, 2014, 135(4): 948-955.

[10] Cornelis MC, Qi L, Zhang CL, et al. Joint effects of common genetic variants on the risk for type 2 diabetes in U. S. men and women of European ancestry [J]. Ann Intern Med, 2009, 150(8): 541-550.

[11] Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses [J]. Am J Hum Genet, 2007, 81(3): 559-575.

[12] Meigs JB, Shrader P, Sullivan LM, et al. Genotype score in addition to common risk factors for prediction of type 2 diabetes [J]. N Engl J Med, 2008, 359(21): 2208-2219.

[13] Hsu FC, Sun JL, Zhu Y, et al. Comparison of two methods for estimating absolute risk of prostate cancer based on single nucleotide polymorphisms and family history [J]. Cancer Epidemiol Biomarkers Prev, 2010, 19(4): 1083-1088.

[14] Wacholder S, Hartge P, Prentice R, et al. Performance of common genetic variants in breast-cancer risk models [J]. N Engl J Med, 2010, 362(11): 986-993.

[15] Mealiffe ME, Stokowski RP, Rhees BK, et al. Assessment of clinical validity of a breast cancer risk model combining genetic and clinical information [J]. J Natl Cancer Inst, 2010, 102(21): 1618-1627.

[16] Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases [J]. Nature, 2009, 461(7265): 747-753.

[17] Botteri E, Iodice S, Bagnardi V, et al. Smoking and colorectal cancer: a meta-analysis [J]. JAMA, 2008, 300(23): 2765-2778.

[18] Ben QW, An W, Jiang Y, et al. Body mass index increases risk for colorectal adenomas based on meta-analysis [J]. Gastroenterology, 2012, 142(4): 762-772.

[19] Bardou M, Barkun AN, Martel M. Obesity and colorectal cancer [J]. Gut, 2013, 62(6): 933-947.

[20] Boyle T, Keegel T, Bull F, et al. Physical activity and risks of proximal and distal colon cancers: a systematic review and meta-analysis [J]. J Natl Cancer Inst, 2012, 104(20): 1548-1561.

（收稿日期：2015-06-15）
（本文编辑：张林东）