Multiple gene-specific DNA methylation in blood leukocytes and colorectal cancer risk: a case-control study in China

SUPPLEMENTARY METHODS

Sample size

We estimated the sample size according to a logistic regression test assessing whether aberrant DNA methylation in blood leukocytes was associated with the risk of colorectal cancer (CRC). A sample size of 652 participants was needed to achieve 90% power (at the 5% level of statistical significance) to detect odds ratios (ORs) of 1.8 or more with a 20% prevalence in the control group. In addition, taking into consideration incomplete questionnaires and the failure rate for MS-HRM detection, we included about 20% more samples and finally targeted a total sample size of 800 participants.

DNA extraction

After the peripheral blood samples were centrifuged at 1,600 g for 10 minutes to separate plasma, leukocyte-derived DNA was extracted from buffy coats using the QIAamp DNA Blood Mini Kit (Qiagen, Hilden, Germany) according to the manufacturer’s protocol. Briefly, buffy coats were resuspended in 400 μl of phosphate-buffered saline. After the addition of 40 μl of proteinase K and 400 μl of buffer AL (Qiagen), the mixture was incubated at 56°C for 10 min, treated with 400 μl of ethanol (99%), and mixed. The mixture was transferred to a QIAamp spin column and centrifuged at 6,000 g at room temperature for 1 minute, followed by washing with buffers AW1 and AW2 and centrifugation. DNA was eluted in 150 μl of buffer AE and was quantified using a NanoDrop 2000c bioanalyzer (Thermo-Fisher, USA). Then, the DNA was divided into aliquots and stored at -80°C.

Bisulfite modification

From each subject, 500 ng of DNA was bisulfite-modified using the EpiTect Plus DNA Bisulfite Kit (Qiagen) according to the manufacturer’s protocol. The bisulfite-treated DNA was eluted in 30 μl of elution buffer and quantified using a NanoDrop 2000c bioanalyzer set to RNA as the sample type [1]. Then, the bisulfite-modified DNA sample was diluted to a final concentration of 10 ng/μl and was distributed into three aliquots for storage (-20°C).

Pyrosequencing verification

To verify the results of MS-HRM, we performed pyrosequencing in a subset of the samples. For DAPK1, pyrosequencing was performed for quantitative methylation analyses of 48 HRM-positive samples (36 CRC case and 12 control samples that were selected from 106 cases and 43 controls, respectively) and 48 HRM-negative samples (24 CRC case and 24 control samples that were selected from 322 cases and 385 controls, respectively); for MLH1, 24 HRM-positive samples (16 CRC case and 8 control samples) and 48 HRM-negative samples (24 CRC case and 24 control samples that were selected from 412 cases and 420 controls, respectively).

The regions analyzed via MS-HRM included 7 and 21 CpG dinucleotide sites for DAPK1 and MLH1, respectively. Subsequently, we designed pyrosequencing primer sets to analyze the corresponding regions for DAPK1 and MLH1 using PyroMark Assay Design Software (version 2.0.1.15, Qiagen). PyroPCR was performed using a PyroMark PCR Kit (Qiagen) according to the manufacturer’s protocol. The primer sets used are listed in Supplementary Table 11 (the reverse primer in each set was biotinylated). The standard conditions for PyroPCR were as follows: 12.5 μl of PyroMark PCR Master Mix, 2.5 μl of CoralLoad Concentrate, 0.5 μl of each primer (10 pmol/μl) and 2 μl (nearly 20 ng) of bisulfite-modified DNA in a final volume of 25 μl. The PyroPCR thermocycling protocol was as follows: initial PCR activation (95°C for 15 min), 45 3-step cycles (94°C for 30 s, 56°C for 30 s, and 72°C for 30 s) and final extension (72°C for 10 min). For each sample, 10 μl of PyroPCR product was used for pyrosequencing reactions and methylation quantification, which were performed using a PyroMark Q24 Advanced System version 2.0.6 (Qiagen) according to the manufacturer’s protocol. The pyrosequencing results were analysed using PyroMark software version 3.0.0 (Qiagen) to obtain a percentage of methylation at each CpG site. In the MS-HRM analysis, we could detect only the mean methylation percentage of the analyzed sequences and could not precisely quantify the methylation percentage of individual CpG sites. Therefore, the mean methylation percentage of all CpG sites based on pyrosequencing was calculated for each analyzed gene, and this value was used for comparison with the results of MS-HRM. The concordance between the methylation levels derived from the two DNA
methylation assessment methods was assessed using Spearman correlation coefficients, ROC curves and AUC analyses. Additionally, we used the Bland-Altman plot as a graphical method to compare the two methods.

**Definitions of questionnaire-derived variables**

Questionnaire-derived variables considered included: age (<60, ≥60), gender, BMI (<24, ≥24), family history of any cancer other than CRC in first-degree relatives (yes, no), occupational physical activity (blue-collar, white-collar), smoking status (no, yes), and consumption of coarse grains (<50, ~100, ~200, ≥200 g/week), dairy (<1, ≥1 times/week), fish stewed with brown sauce (<1, ≥1 times/week), fried food (<1, ≥1 times/month), fruits (<2, ≥2 times/week), green vegetables (<100, ≥100 g/day), leftovers (<1, ≥1 times/week), pork (<250, ≥250 g/week) and soybeans (<1, ≥1 times/week).

**Missing data analysis and imputation**

All questionnaire-derived variables were analyzed via missing value analysis (Little’s missing completely at random (MCAR) test) to assess whether the pattern of missing data was related to the observed data (MCAR or missing at random (MAR)). The MAR assumption fit our data. Therefore, missing data were imputed using multiple imputation (five imputations) via the expectation-maximization method, with all questionnaire-derived variables (including outcomes) included in the imputation model. We used the method of multiple imputation to impute the missing dataset rather than single imputation because the former is generally considered to be superior to the latter for solving the problem of missing data [2, 3].

**Statistical analysis**

The sample size was calculated using PASS version 11.0.7 (NCSS LLC., USA). The Bland-Altman plot analysis was performed using MedCalc software version 15.4 (Ostend, Belgium).

**SUPPLEMENTARY RESULTS**

**Missing data and imputation**

The characteristics of demographic variables and questionnaire-derived variables of the participants before and after multiple imputations are listed in Supplementary Table 1. Across the 856 subjects, missing data occurred in less than 5% for each questionnaire-derived variable, except for consumption of fruits, green vegetables and dairy. The environmental factors that are missing across the entire dataset included family history (values missing for 24 participants, 2.80% of the samples), occupational physical activity (13, 1.52%), dietary consumption of coarse grains (2, 0.23%), dairy products (206, 24.07%), fish stewed with brown sauce (11, 1.28%), fried food (5, 0.58%), fresh fruits (50, 5.84%), green vegetables (52, 6.07%), leftover (7, 0.82%), pork (24, 2.80%), and soybean products (1, 0.12%).

**Pyrosequencing verification**

For DAPK1 and MLH1, the methylation status obtained via MS-HRM was compared with the mean methylation level based on quantitative pyrosequencing. Using Bland-Altman plots, we found that for MLH1, all examinations of methylation fell within the limits of agreement except for one sample displaying a high methylation level, in which MS-HRM reported a lower methylation percentage than pyrosequencing (Supplementary Figure 1). For DAPK1, all samples fell within the limits of agreement except for three samples with relatively high methylation, in which MS-HRM tended to indicate lower levels of methylation than pyrosequencing. These results indicated that the MS-HRM results were well confirmed by the pyrosequencing results.

**REFERENCES FOR SUPPLEMENTARY METHODS**

1. Laurin N, Celestin F, Clark M, Wilkinson D, Yamashita B, Fregeau C. New incompatibilities uncovered using the Promega DNA IQ chemistry. Forensic science international. 2015; 257:134-141.
2. Lorenzo-Seva U, Ferrando PJ. POLYMAT-C: a comprehensive SPSS program for computing the polychoric correlation matrix. Behav Res Methods. 2015; 47:884-889.
3. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, Vickers AJ, Ransohoff DF, Collins GS. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Annals of internal medicine. 2015; 162:W1-73.
(Continued)
Supplementary Figure 1: Comparisons of the methylation levels of *DAPK1* and *MLH1* obtained by methylation-sensitive high-resolution melting (MS-HRM) and pyrosequencing. Profiles of *DAPK1* (A) and *MLH1* (B) methylation for each sample. Spearman correlation analyses for *DAPK1* (C) and *MLH1* (D). Receiver operating characteristic curves (ROC) and corresponding area under the curve (AUC) analyses for *DAPK1* (E) and *MLH1* (F). Bland-Altman plots for *DAPK1* (G) and *MLH1* (H).
Normalized Melting Curves

(Continued)
Supplementary Figure 2: Normalized melting curves of the MS-HRM assay for each gene. (A) APC; (B) CDH1; (C) CDKN2A; (D) DAPK1; (E) IGF2; (F) MGMT; (G) MINT31; (H) MLH1; (I) NEUROG1 and (J) WIF1. Data were analyzed using the software modules of Gene Scanning. A set of dilution standards (including 100%, 5%, 2%, 1% and 0% methylated DNA) were used for all the 10 genes. The blue line refers to the normalized melting curve of the 100% methylated DNA controls, red (5%), green (2%), purple (1%) and brown (unmethylated DNA control).
Supplementary Table 1: Demographic variables and questionnaire-derived variables of participants before and after multiple imputations

See Supplementary File 1
### Supplementary Table 2: Associations between DNA methylation and the risk of CRC according to gender

| DNA methylation | Male | Female |
|-----------------|------|--------|
|                 | OR\_adjusted\(^a\) | 95% CI | \(P\)-value | OR\_adjusted\(^b\) | 95% CI | \(P\)-value |
| APC             | 1.0079 | 0.3274-3.1027 | 0.9891 | 0.9307 | 0.2685-3.2265 | 0.9098 |
| CDH1            | 0.5460 | 0.2591-1.1504 | 0.1115 | 0.5729 | 0.2496-1.3147 | 0.1887 |
| CDKN2A          | 3.8048 | 0.7690-18.8248 | 0.1014 | 3.5483 | 0.6330-19.8891 | 0.1499 |
| DAPK1           | 3.1068 | 1.8621-5.1835 | \(<0.0001\) | 3.4747 | 1.9710-6.1258 | \(<0.0001\) |
| IGF2            | 3.0470 | 1.7782-5.2209 | 0.0001 | 2.8244 | 1.5732-5.0709 | 0.0005 |
| MGMT            | 1.8404 | 1.0633-3.1855 | 0.0293 | 1.8789 | 1.0329-3.4179 | 0.0388 |
| MINT31          | 16.4418 | 2.1239-127.2812 | 0.0073 | 19.1759 | 2.3319-157.6892 | 0.0061 |
| MLH1            | 1.5214 | 0.5808-3.9854 | 0.3931 | 1.4222 | 0.5097-3.9685 | 0.5012 |
| NEUROG1         | 2.7349 | 1.4759-5.0681 | 0.0014 | 2.5590 | 1.3239-4.9465 | 0.0052 |
| WIF1            | 2.4800 | 1.4249-4.3163 | 0.0013 | 2.6647 | 1.4321-4.9582 | 0.0020 |
| MRS\_10-M       | 4.0606 | 2.7080-6.0889 | \(<0.0001\) | 4.6556 | 2.9603-7.3216 | \(<0.0001\) |
| MRS\_10-H       | 9.9474 | 4.6758-21.1623 | \(<0.0001\) | 10.2057 | 4.4860-23.2179 | \(<0.0001\) |

CI: confidence interval; CRC: colorectal cancer; MRS: methylation risk score; OR: odds ratio.

\(^a\) ORs adjusted for age and BMI.

\(^b\) ORs adjusted for age, BMI, occupational physical activity, smoking, and consumption of coarse grains, fish stewed with brown sauce, fried food, leftovers and pork.

\(P\) values < 0.05 are in bold.
**Supplementary Table 3: Associations between DNA methylation and the risk of CRC according to age**

| DNA methylation | <60 years | | | | | ≥60 years | | | |
|-----------------|-----------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
|                  | OR adjusted | 95% CI         | P-value | OR adjusted | 95% CI         | P-value | OR adjusted | 95% CI         | P-value | OR adjusted | 95% CI         | P-value |
| APC              | 1.8991     | 0.4762-5.7526  | 0.3635 | 1.9450      | 0.5108-7.466   | 0.3639  | 1.7217     | 0.5895-5.0280  | 0.3204 | 1.8853     | 0.5669-6.295   | 0.3010 |
| CDH1             | 0.4675     | 0.1930-1.1321  | 0.0920 | 0.4518      | 0.1772-1.1518  | 0.0961  | 0.6847     | 0.3359-1.3957  | 0.2972 | 1.0010     | 0.4417-2.2685  | 0.9981 |
| CDKN2A           | 5.0497     | 0.5747-44.3708 | 0.1442 | 6.1125      | 0.6072-61.5854 | 0.1250  | 1.3169     | 0.2835-6.1170  | 0.7253 | 0.8220     | 0.1605-4.2093  | 0.8141 |
| DAPK1            | 3.4006     | 1.8816-6.1459  | 0.0001 | 3.2927      | 1.7756-6.1089  | 0.0002  | 2.6161     | 1.5493-4.4174  | 0.0003 | 2.8258     | 1.5608-5.1160  | 0.0006 |
| IGF2             | 2.0861     | 1.1520-3.7778  | 0.0152 | 2.1398      | 1.1412-4.0122  | 0.0177  | 3.1931     | 1.8222-5.5957  | <0.001 | 3.0396     | 1.6413-5.6295  | 0.0004 |
| MGMT             | 2.3714     | 1.2580-4.4703  | 0.0076 | 2.3096      | 1.1921-4.4748  | 0.0131  | 1.3931     | 0.7624-2.5457  | 0.2811 | 1.5256     | 0.7835-2.9709  | 0.2141 |
| MINT31           | 7.3221     | 1.5990-33.5293 | 0.0103 | 8.4610      | 1.7952-39.8787 | 0.0070  | 1.9767     | 0.4943-7.9052  | 0.3353 | 1.7715     | 0.4289-7.3173  | 0.4294 |
| MLH1             | 3.0382     | 0.7986-11.5590 | 0.1031 | 2.5814      | 0.6463-10.3109 | 0.1803  | 1.2971     | 0.3968-4.2407  | 0.6669 | 1.2666     | 0.3394-4.7259  | 0.7250 |
| NEUROG1          | 2.1655     | 1.0437-4.4929  | 0.0380 | 2.1039      | 0.9806-4.5137  | 0.0562  | 2.9836     | 1.5818-5.6275  | 0.0007 | 3.1101     | 1.5646-6.1819  | 0.0012 |
| WIFI             | 2.2057     | 1.2066-4.0321  | 0.0102 | 2.1298      | 1.1321-4.0069  | 0.0190  | 2.3500     | 1.2724-4.3403  | 0.0063 | 2.9545     | 1.4648-5.9591  | 0.0025 |
| MRS_10-M         | 3.3651     | 2.1487-5.2703  | <0.0001 | 3.7130      | 2.2915-6.0163  | <0.0001 | 4.0251     | 2.5361-6.3881  | <0.0001 | 4.2716     | 2.5302-7.2116  | <0.0001 |
| MRS_10-H         | 8.8417     | 3.7590-20.7968 | <0.0001 | 9.3211      | 3.8057-22.8295 | <0.0001 | 5.2290     | 2.7089-10.0936 | <0.0001 | 4.9945     | 2.4213-10.3019 | <0.0001 |

CI: confidence interval; CRC: colorectal cancer; MRS: methylation risk score; OR: odds ratio.

* ORs adjusted for gender and BMI.

b ORs adjusted for gender, BMI, occupational physical activity, smoking, and consumption of coarse grains, fish stewed with brown sauce, fried food, leftovers and pork.

P values < 0.05 are in bold.
| DNA methylation | <24 |          |          | ≥24 |          |          |
|----------------|-----|----------|----------|-----|----------|----------|
|                | OR\(^a\) | 95% CI   | P-value  | OR\(^b\) | 95% CI   | P-value  |
| APC            | 1.2586 | 0.4141-3.8258 | 0.6851 | 1.3584 | 0.4184-4.4103 | 0.6103 |
| CDH1           | 0.8938 | 0.4304-1.8561 | 0.7634 | 1.0214 | 0.4619-2.2583 | 0.9584 |
| CDKN2A         | 4.2718 | 0.5095-35.8148 | 0.1808 | 4.2958 | 0.4804-38.4095 | 0.1948 |
| DAPK1          | 3.1913 | 1.7938-5.6776 | 0.0001 | 3.5033 | 1.8923-6.4859 | 0.0001 |
| IGF2           | 2.7907 | 1.5962-4.8789 | 0.0003 | 2.7255 | 1.5170-4.8967 | 0.0008 |
| MGMT           | 1.6104 | 0.8450-3.0694 | 0.1476 | 1.6306 | 0.8301-3.2031 | 0.1558 |
| MINT31         | 2.1544 | 0.5734-8.0942 | 0.2557 | 2.7438 | 0.7069-10.6498 | 0.1447 |
| MLH1           | 1.7643 | 0.5397-5.7621 | 0.3477 | 1.7262 | 0.4877-6.1090 | 0.3973 |
| NEUROG1        | 2.6286 | 1.3289-5.1933 | 0.0055 | 2.9123 | 1.4306-5.9287 | 0.0032 |
| WIF1           | 2.2053 | 1.1815-4.1163 | 0.0130 | 2.7322 | 1.3958-5.3483 | 0.0034 |
| MRS_10-M       | 3.7159 | 2.3484-5.8797 | <0.0001 | 4.5709 | 2.7525-7.5908 | <0.0001 |
| MRS_10-H       | 6.0499 | 2.9690-12.3281 | <0.0001 | 6.7389 | 3.1595-14.3733 | <0.0001 |

CI: confidence interval; CRC: colorectal cancer; MRS: methylation risk score; OR: odds ratio.

\(^a\) ORs adjusted for age, gender and BMI.

\(^b\) ORs adjusted for gender, BMI, occupational physical activity, smoking, and consumption of coarse grains, fish stewed with brown sauce, fried food, leftovers and pork.

\(P\) values < 0.05 are in bold.
Supplementary Table 5: Reclassification table of individuals of predicted risk using the combined EF and MRS_10 model versus the EF-only model

| Basic model | Combined EF and MRS_10 model |
|-------------|-----------------------------|
| EF-only model | ≤0.5 | >0.5 | RC% |
| ≤0.5 | | | |
| Total | 358 | 77 | 18 |
| Case | 97 | 57 | 37 |
| Control | 261 | 20 | 7 |
| >0.5 | | | |
| Total | 88 | 333 | 21 |
| Case | 31 | 243 | 11 |
| Control | 57 | 90 | 39 |

NRI (95% CI) | 0.1472 (0.0892-0.2052), \( P<0.0001 \)
IDI (95% CI) | 0.0930 (0.0736-0.1123), \( P<0.0001 \)
Delta-AUC (95% CI) | 0.0618 (0.0399-0.0854), \( P<0.0001 \)

AUC: area under the curve; CI: confidence interval; Delta-AUC: the difference of AUCs between the basic model and the extended model with the MRS; EF: environmental factors; IDI: integrated discrimination improvement; MRS: methylation risk score; NRI: net reclassification improvement; RC%: reclassified percent (%).

\( P \) values < 0.05 are in bold.
Supplementary Table 6: Relationship between environmental factors and DNA methylation

See Supplementary File 1

Supplementary Table 7: Sensitivity analysis for the associations between MRS_10 and CRC risk by omitting each gene at a time

See Supplementary File 1
Supplementary Table 8: Correlation between DNA methylation levels in leukocytes and matched colorectal tumour tissues in patients with CRC

| Gene     | APC   | CDKN2A | MGMT  | APBA1 | MINT31 | MLH1  |
|----------|-------|--------|-------|-------|--------|-------|
| Number of Cases | 211   | 213    | 211   | 205   | 204    | 205   |
| Spearman's Rho  | -0.1487 | 0.0080 | 0.0267 | 0.0695 | 0.0914 | 0.0003 |
| P-value     | 0.0301 | 0.9079 | 0.6995 | 0.3218 | 0.1934 | 0.9967 |
**Supplementary Table 9: Sensitivity analysis for the categorical net reclassification improvement at different cut-off points**

| Basic Model  | NRI at Different Cut-Off Points | Combined EF and MRS_10 |
|--------------|---------------------------------|------------------------|
| EF-Only      | NRI^{0.3}                        | 0.0911 (0.0437-0.1385), P=0.0002 |
|              | NRI^{0.4}                        | 0.0841 (0.0259-0.1423), P=0.0046 |
|              | NRI^{0.6}                        | 0.1238 (0.0633-0.1844), P<0.0001 |
|              | NRI^{0.7}                        | 0.1168 (0.0615-0.1721), P<0.0001 |

AUC: area under the curve; CI: confidence interval; EF: environmental factors; MRS: methylation risk score; NRI: net reclassification improvement.

*P* values < 0.05 are in bold.
Supplementary Table 10: Sequences of primer sets used for methylation-sensitive high-resolution melting analysis

See Supplementary File 1
**Supplementary Table 11: Sequences of primer sets used for pyrosequencing**

| Gene | Primer sequences: 5’-3’ | CpG Sites/Amplicon size (bp) | Annealing temperature (°C) |
|------|--------------------------|-----------------------------|-----------------------------|
| **DAPK1** | Forward PCR primer: **GCGCGGGAGTTGGGAGGAGT**<br>Reverse PCR primer: **Biotin-CTCCGAAACTACCTACCCAAACC**<br>Sequencing primer: **GCGCGGGAGTTGGGAGGAGT**<br>Analyzed sequence: **AGYGAGYGTYGTYGTAGAATTYGTAGYGTYGGTTTGGTAGGGTAGTT**<br>Dispensation order: **GATGTCAATCGTCAGTCATGTACGTAGTCGTACGGTACG** | 7/70 | 56 |
| **MLH1** | Forward PCR primer: **GGAGTGAAGGAGGTTACGGGTAAGT**<br>Reverse PCR primer: **Biotin-AAAAACGATAAAACCCTATACCTAATCTATC**<br>Sequencing primer_1: **GGAGTGAAGGAGGTTACGGGTAAGT**<br>Analyzed sequence_1: **YGGGTAAGTYGTTTTGAYGTAGAYGTTTTATTAGGGTYGYGYGTYGTYGTYGTTATATATYGTTYGTAGTATTYGTGTTAGGT**<br>Dispensation order_1: **ATCGTATGTCTGTTGTATCGTAGTATCGTTATATGTCTAGTCAGTCAGTCTAGTCGTAGTGATCTGTA**<br>Sequencing primer_2: **GTAGTATTAGTGTTTAGTTT**<br>Analyzed sequence_2: **YGTAGTGGYGTTTGAYGTYGTYGGGTAGTTAYGATGAGGYGGYGATAGATTAGGTATAGGGTTTTAT**<br>Dispensation order_2: **GTCGTAGTAGTCGTGTATCAGTCTAGTCAGTCTAGTCGTAGTGATCTGTA** | 22/182 | 56 |

*For MLH1, the size of the PyroMark PCR-amplified DNA template is 182bp, which is too long for pyrosequencing in a single pyrosequencing reaction; thus, we performed two pyrosequencing reaction using two sets of sequencing primers.*
Supplementary Table 12: Estimates of regression coefficients and odds ratios from the models

See Supplementary File 1