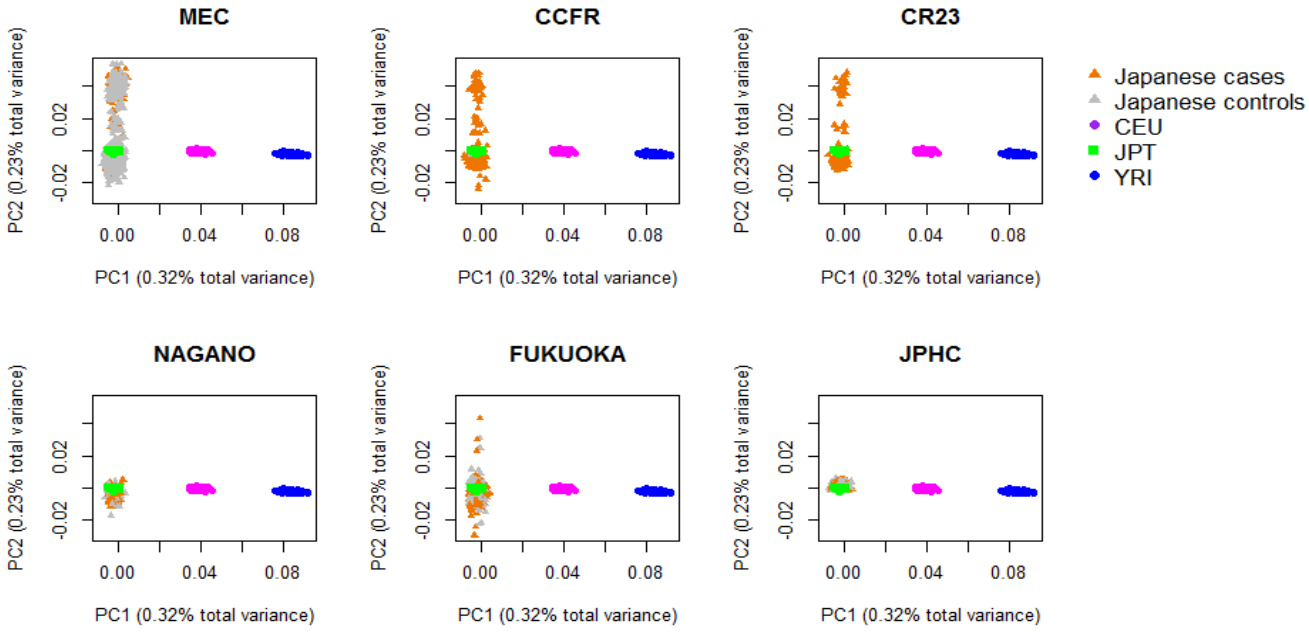
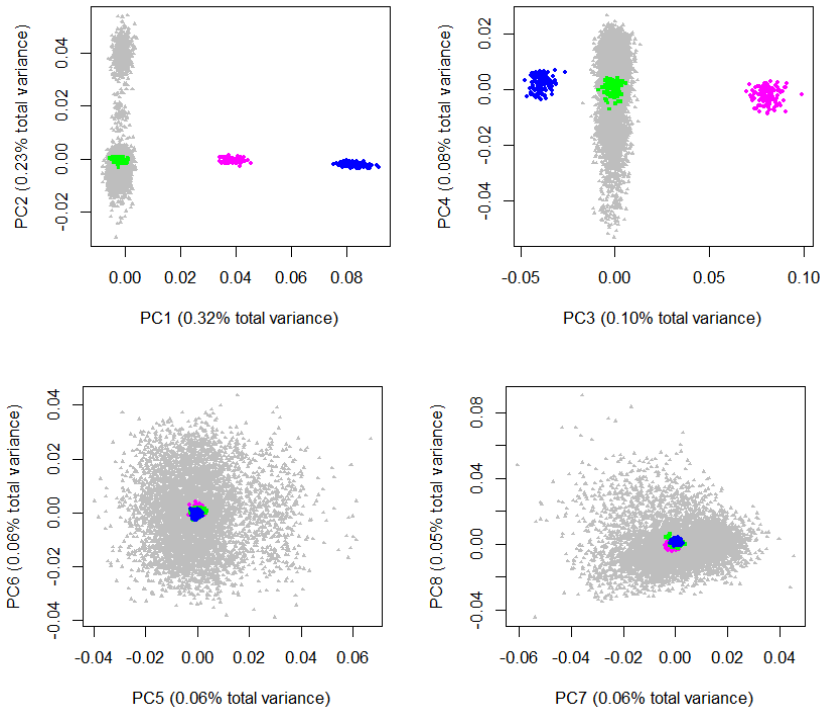
# Supplementary Information

## Supplementary Figures

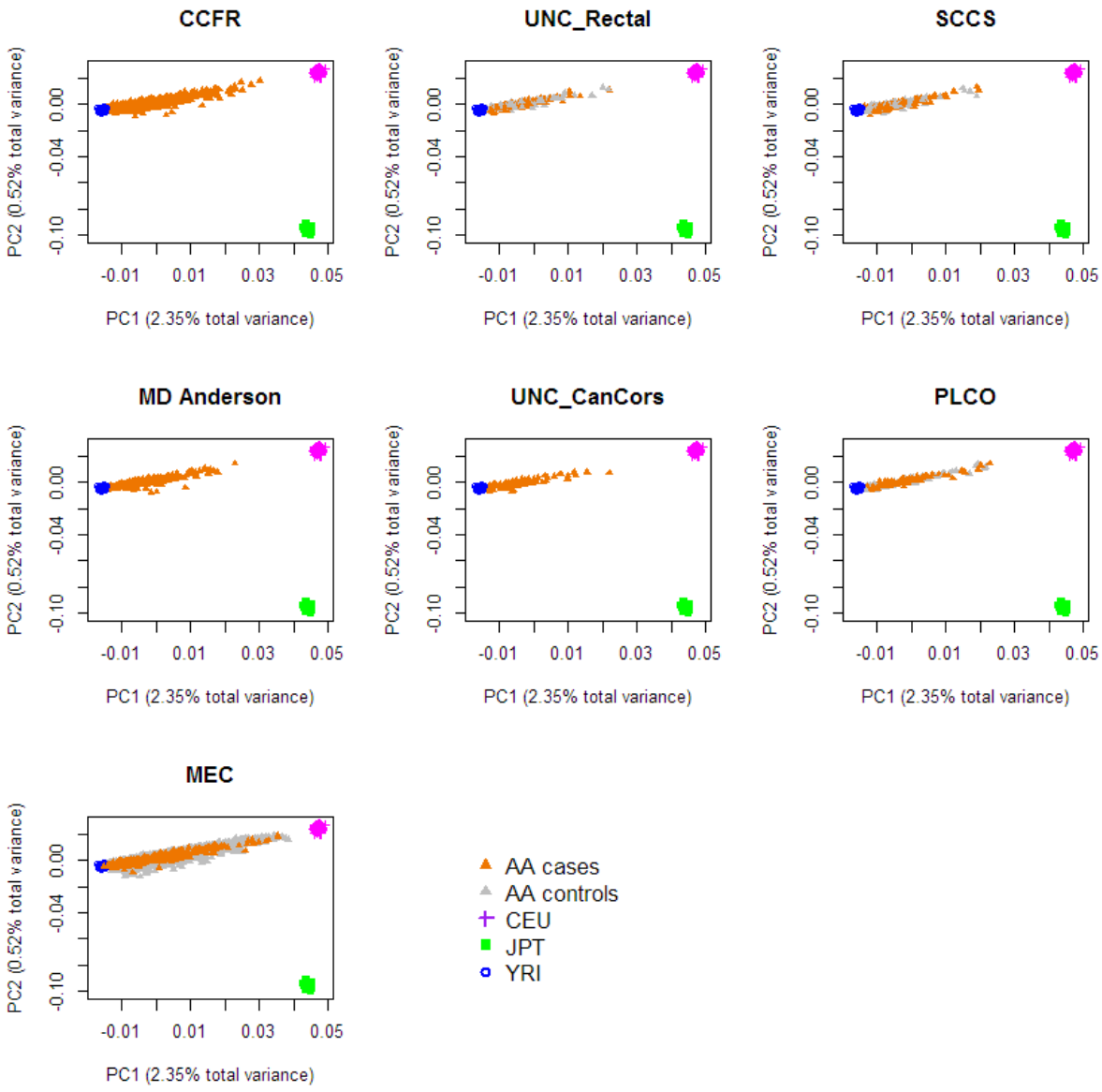
A)




B)

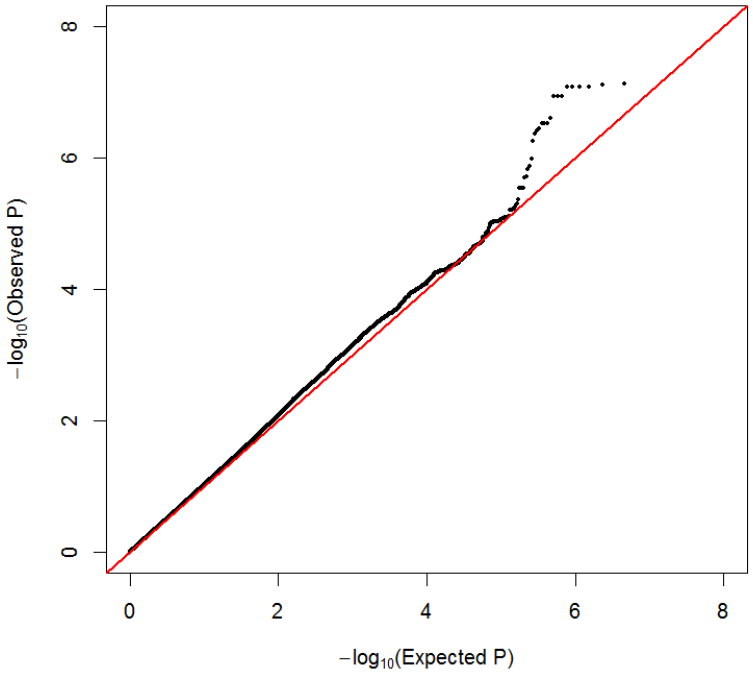



**Supplementary Figure 1.** Distribution of top PCs for the Japanese subjects and HapMap CEU (purple), JPT (green) and YRI (blue) populations. A) Plot of the first two PCs by study/center and disease status of the Japanese, where CRC cases are in orange and controls in grey B) Plot of the first 8 PCs, where Japanese subjects are all in grey.

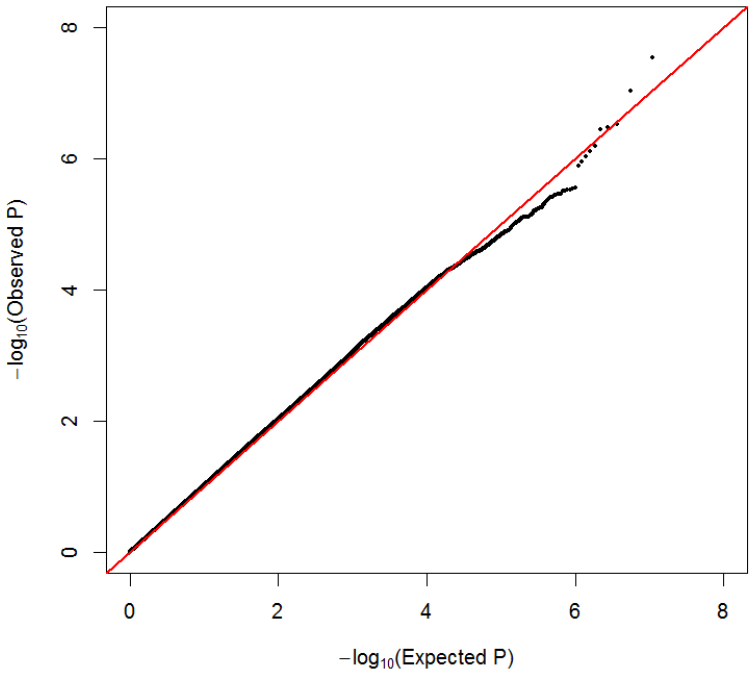



**Supplementary Figure 2.** Distribution of the first two PCs in African American (AA) subjects, by study/center and disease status and HapMap CEU (purple), JPT (green) and YRI (blue) populations. CRC cases are in orange and controls in grey.

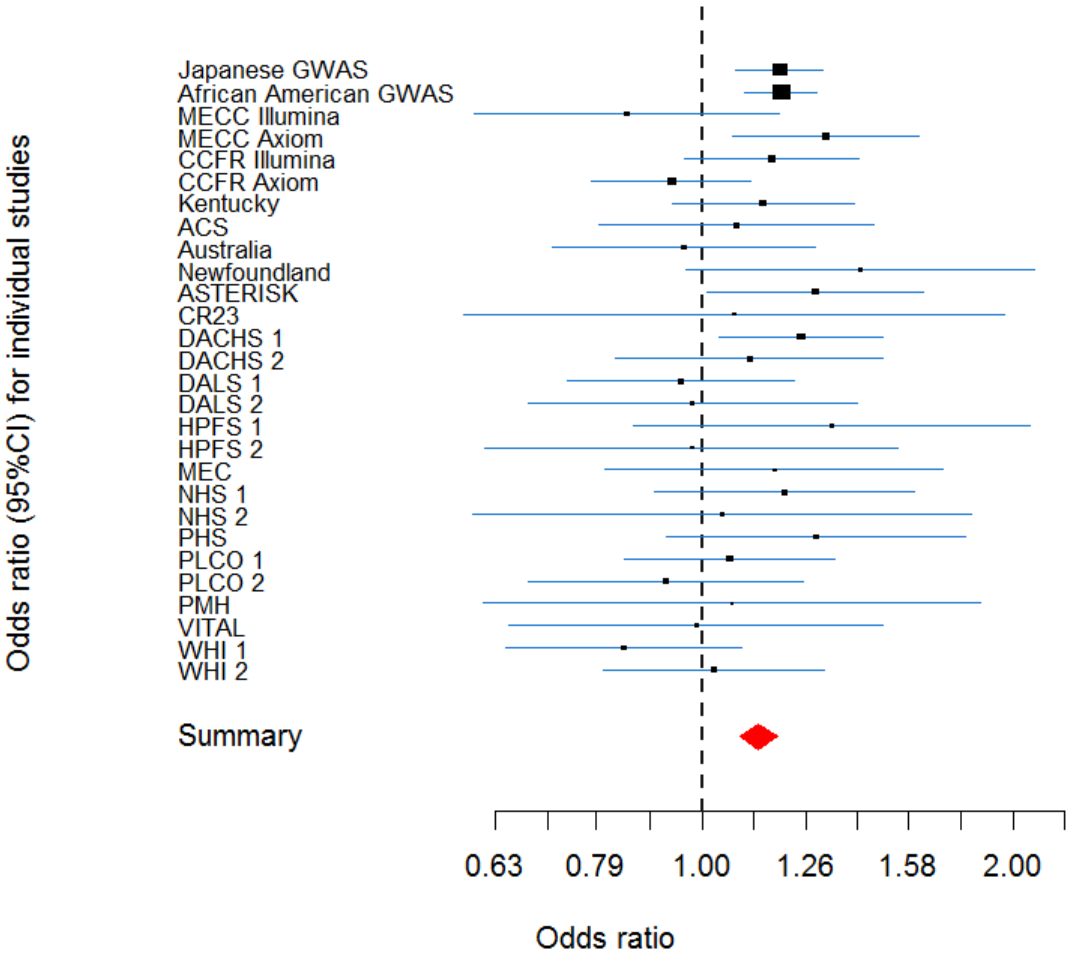
a)




b)

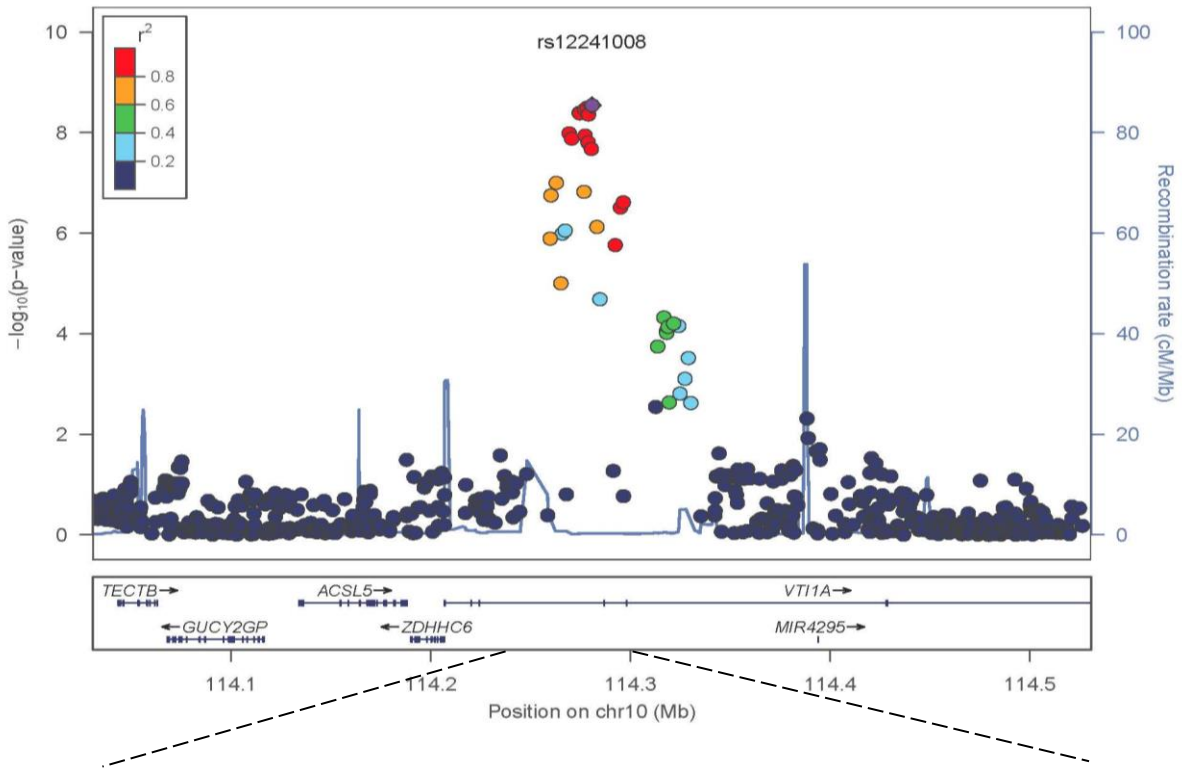



**Supplementary Figure 3.** Quantile-Quantile plot of genome-wide association results in the Japanese GWAS (a) and in the African American GWAS (b).

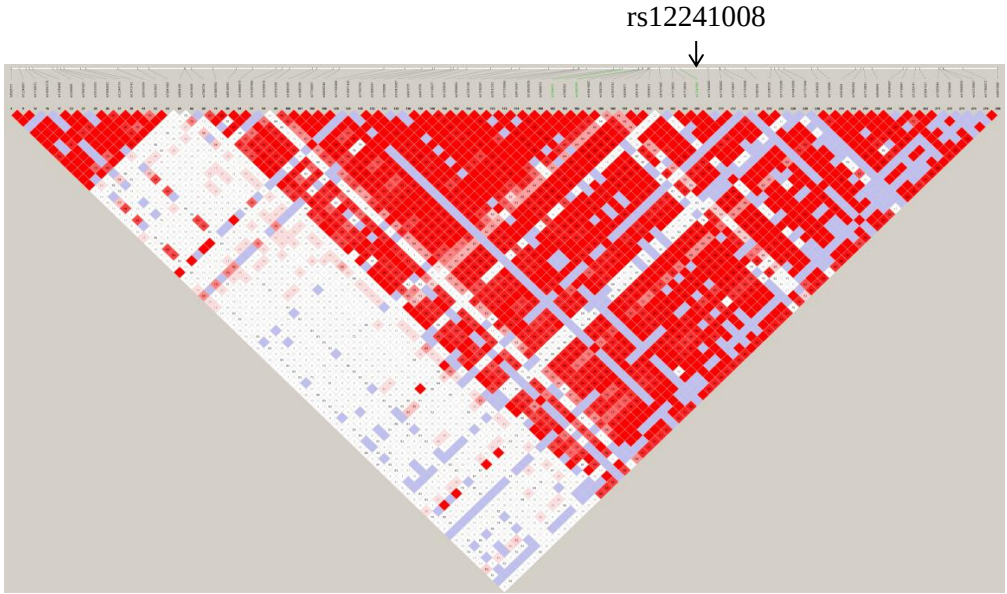



**Supplementary Figure 4.** Study-specific association results for rs12241008.

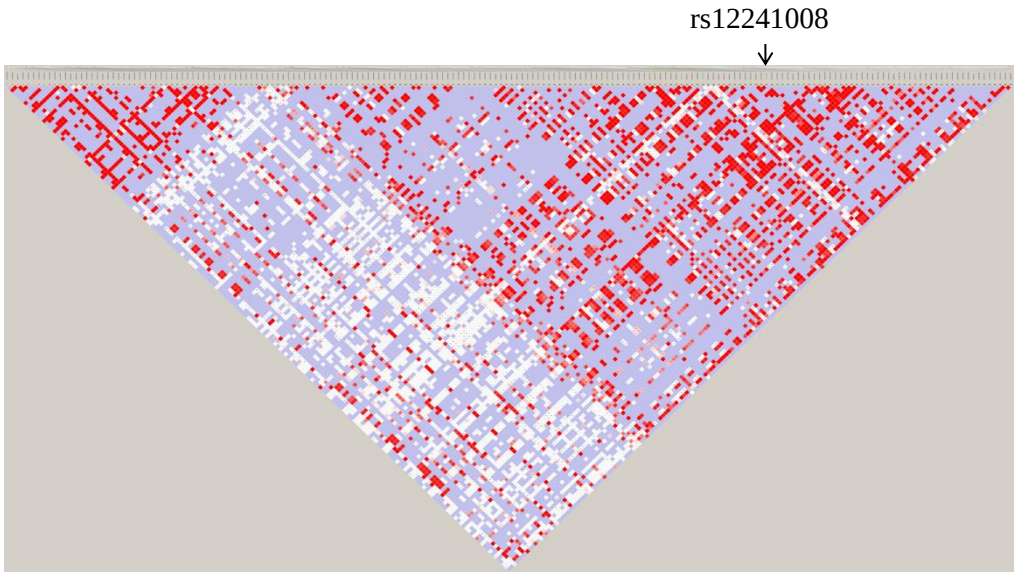
a)


rs12241008
r²
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
TECTB
ACSL5
VTI1A
GUCY2GP
ZDHHC6
MIR4295
114.1
114.2
114.3
114.4
114.5
Position on chr10 (Mb)

b)


rs12241008

c)




d)


**Supplementary Figure 5.** Regional P-value and LD plots for the newly identified 10q25 CRC risk locus. (a) P-value plot, where P was generated combining all Japanese, African Americas, CORECT and GECCO studies. The SNP with the smallest *p*-value, rs12241008, is shown as a purple diamond. $r^2$ is in relation to this SNP from the 1000 Genomes Project in Europeans. (b-d) LD pattern (generated by HaploView) for the region 114.24 Mb to 114.3 Mb using the 1000 Genomes data in East Asians (b), Africans (c) and Europeans (d), based on markers with minor allele frequency > 0.01. In all 3 populations, the top hit rs12241008 is in a LD region of roughly 50 kb long.

## Supplementary Tables

**Supplementary Table 1.** Basic characteristics of 6,424 Japanese subjects by study

| Study/Batch | | Genotyped Markers | Controls | Cases | % Female | | Mean Age (SD) | |
|---|---|---|---|---|---|---|---|---|
| | | | | | Controls | Cases | Controls | Cases |
| MEC | MEC1 | 786,901 | 744 | 754 | 47.1 | 40.0 | 69.3 (8.5) | 70.6 (8.6) |
| | MEC2 | 558,151 | 18 | 23 | | | | |
| | PrCa | 450,114 | 825 | 16 | | | | |
| | BrCa | 446,586 | 710 | 4 | | | | |
| CCFR | | 790,064 | 9 | 276 | 66.7 | 47.5 | 68.0 (10.5) | 61.4 (11.5) |
| CR2&3 | | | | 134 | na | 43.3 | na | 66.4 (13.0) |
| Fukuoka | | | 749 | 662 | 37.0 | 37.0 | 58.6 (10.6) | 60.5 (9.1) |
| Nagano | | | 102 | 105 | 35.3 | 36.2 | 59.1 (8.8) | 59.6 (8.9) |
| JPHC | | 782,650 | 640 | 653 | 48.0 | 48.2 | 56.2 (7.4) | 66.3 (8.0) |
| **Total** | | | 3797 | 2627 | 45.0 | 42.1 | 64.7 (10.5) | 65.3 (10.1) |

**Supplementary Table 2.** Basic characteristics of 6,597 African American subjects by study

| | ***n* (% Female)** | | **Mean Age (SD)** | |
|---|---|---|---|---|
| **Study** | **Controls** | **Cases** | **Controls** | **Cases** |
| CCFR | 19 (63.2) | 943 (50.5) | 55.5 (12.3) | 59.1 (10.5) |
| UNC-CanCORS | | 80 (52.5) | | 63.1 (9.9) |
| UNC-Rectal | 105 (42.9) | 109 (48.6) | 62.9 (10.0) | 62.2 (10.2) |
| MEC | 4328 (33.8) | 342 (41.5) | 68.6 (8.3) | 73.0 (6.7) |
| MD ANDERSON | | 185 (50.3) | | 56.6 (12.2) |
| PLCO | 94 (53.2) | 76 (52.6) | 66.0 (5.4) | 64.8 (5.5) |
| SCCS | 157 (57.3) | 159 (58.5) | 55.2 (8.4) | 55.0 (8.4) |
| **Total** | 4703 (35.2) | 1894 (49.6) | 61.6 (11.3) | 67.9 (8.7) |

**Supplementary Table 3.** Sample sizes and genotyping platforms for participating studies

| STUDY | | Cases | Controls | Total | Platform |
|---|---|---|---|---|---|
| **African American GWAS** | | 1,894 | 4,703 | 6,597 | Illumina 1M/2.5M |
| **Japanese GWAS** | | 2,627 | 3,797 | 6,424 | Illumina 1M/660k |
| | Sum | 4,521 | 8,500 | 13,021 | |
| | | | | | |
| **CORECT (European Ancestry)** | | | | | |
| MECC_Illumina | | 484 | 498 | 982 | Illumina Omni 2.5 |
| MECC Axiom | | 1,120 | 820 | 1,940 | Affymetrix Axiom |
| CCFR_Illumina | | 1,977 | 999 | 2,976 | Illumina 1M/1M-duo, Omni |
| CCFR Axiom | | 1,660 | 1,393 | 3,053 | Affymetrix Axiom |
| Kentucky | | 1,038 | 1,134 | 2,172 | Affymetrix Axiom |
| ACS/CPSII | | 548 | 538 | 1,086 | Affymetrix Axiom |
| Melbourne | | 539 | 469 | 1,008 | Affymetrix Axiom |
| Newfoundland | | 195 | 477 | 672 | Affymetrix Axiom |
| | Sum | 7,561 | 6,328 | 13,889 | |
| | | | | | |
| **GECCO (European Ancestry)** | | | | | |
| ASTERISK | | 948 | 947 | 1,895 | Illumina 300K |
| CR2&3 | | 87 | 125 | 212 | Illumina 300K |
| DACHS 1 | | 1,710 | 1,707 | 3,417 | Illumina 300K |
| DACHS 2 | | 675 | 498 | 1,173 | Illumina OmniExpress |
| DALS 1 | | 706 | 710 | 1,416 | Illumina 550K/610K |
| DALS 2 | | 410 | 464 | 874 | Illumina 300K |
| HPFS 1 | | 227 | 230 | 457 | Illumina OmniExpress |
| HPFS 2 | | 176 | 172 | 348 | Illumina OmniExpress |
| MEC | | 328 | 346 | 674 | Illumina 300K |
| NHS 1 | | 394 | 774 | 1,168 | Illumina OmniExpress |
| NHS 2 | | 159 | 181 | 340 | Illumina OmniExpress |
| PHS | | 382 | 389 | 771 | Illumina OmniExpress |
| PMH | | 280 | 122 | 402 | Illumina 300K |
| PLCO 1 | | 533 | 1,976 | 2,509 | Illumina 550K/610K |
| PLCO 2 | | 486 | 415 | 901 | Illumina 300K |
| VITAL | | 285 | 288 | 573 | Illumina 300K |
| WHI 1 | | 470 | 1,529 | 1,999 | Illumina 550K/610K |
| WHI 2 | | 1,006 | 1,010 | 2,016 | Illumina 300K |
| | Sum | 9,262 | 11,883 | 21,145 | |
| OFCCR | | 650 | 522 | 1,172 | Affymetrix 100K/500K |
| | Sum | 9,912 | 12,405 | 22,317 | |

**Supplementary Table 4.** Associations for 2 SNPs with $P < 5\times10^{-8}$ in the combined Japanese (JPN) and African American (AA) GWAS and 9 SNPs with $P < 5\times10^{-8}$ in the combined analysis of all subjects (including replication studies) on chromosome 10.

| | | | | | Risk Allele Freq. | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **SNP** | **BP** | **Risk** | **Other** | **Study** | **Case** | **Control** | **OR (95% CI)** | **P-value** | **$I^2$** |
| **SNPs with $P < 5\times10^{-8}$ in AA+JPN** | | | | | | | | | |
| rs7894915 | 114277039 | G | C | AA | 0.22 | 0.19 | 1.19 (1.08-1.31) | 3.9E-04 | |
| | | | | JPN | 0.28 | 0.25 | 1.18 (1.09-1.28) | 3.4E-05 | |
| | | | | AA+JPN | | | 1.19 (1.12-1.26) | 4.8E-08 | 0 |
| | | | | CORECT | 0.10 | 0.09 | 1.09 (1.01-1.19) | 3.3E-02 | 39 |
| | | | | GECCO | 0.10 | 0.09 | 1.09 (1.02-1.18) | 1.6E-02 | 0 |
| | | | | Combined | | | 1.13 (1.09-1.18) | 1.6E-09 | 6 |
| rs10082356 | 114278181 | G | A | AA | 0.22 | 0.19 | 1.19 (1.08-1.31) | 3.9E-04 | |
| | | | | JPN | 0.28 | 0.25 | 1.18 (1.09-1.28) | 3.4E-05 | |
| | | | | AA+JPN | | | 1.19 (1.12-1.26) | 4.9E-08 | 0 |
| | | | | CORECT | 0.10 | 0.09 | 1.10 (1.01-1.19) | 3.2E-02 | 37 |
| | | | | GECCO | 0.10 | 0.09 | 1.09 (1.02-1.18) | 1.5E-02 | 0 |
| | | | | Combined | | | 1.13 (1.09-1.18) | 1.5E-09 | 5 |
| **SNPs with $P < 5\times10^{-8}$ in combined analysis of all subjects** | | | | | | | | | |
| rs10509964 | 114270474 | T | A | AA | 0.20 | 0.17 | 1.17 (1.06-1.29) | 2.1E-03 | |
| | | | | JPN | 0.28 | 0.24 | 1.18 (1.09-1.28) | 7.0E-05 | |
| | | | | AA+JPN | | | 1.18 (1.10-1.25) | 5.0E-07 | 0 |
| | | | | CORECT | 0.10 | 0.09 | 1.10 (1.01-1.20) | 2.3E-02 | 37 |
| | | | | GECCO | 0.10 | 0.09 | 1.10 (1.02-1.18) | 1.8E-02 | 0 |
| | | | | Combined | | | 1.13 (1.09-1.18) | 6.8E-09 | 0 |
| rs12263620 | 114269239 | G | A | AA | 0.20 | 0.17 | 1.17 (1.06-1.29) | 2.1E-03 | |
| | | | | JPN | 0.28 | 0.25 | 1.19 (1.09-1.28) | 3.4E-05 | |
| | | | | AA+JPN | | | 1.18 (1.11-1.26) | 2.5E-07 | 0 |
| | | | | CORECT | 0.10 | 0.09 | 1.10 (1.01-1.19) | 2.9E-02 | 38 |
| | | | | GECCO | 0.10 | 0.09 | 1.10 (1.02-1.18) | 1.8E-02 | 0 |
| | | | | Combined | | | 1.13 (1.09-1.18) | 5.1E-09 | 2 |
| rs17129834 | 114280318 | C | T | AA | 0.18 | 0.16 | 1.16 (1.05-1.30) | 5.6E-03 | |
| | | | | JPN | 0.28 | 0.25 | 1.19 (1.10-1.29) | 2.2E-05 | |
| | | | | AA+JPN | | | 1.18 (1.11-1.26) | 4.3E-07 | 0 |
| | | | | CORECT | 0.10 | 0.09 | 1.10 (1.01-1.19) | 3.2E-02 | 40 |
| | | | | GECCO | 0.09 | 0.09 | 1.09 (1.02-1.18) | 1.8E-02 | 0 |
| | | | | Combined | | | 1.13 (1.08-1.18) | 1.1E-08 | 3 |
| rs34139797 | 114274559 | G | A | AA | 0.23 | 0.19 | 1.19 (1.08-1.31) | 4.5E-04 | |
| | | | | JPN | 0.28 | 0.25 | 1.18 (1.09-1.28) | 4.1E-05 | |
| | | | | AA+JPN | | | 1.18 (1.11-1.26) | 6.8E-08 | 0 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | CORECT | 0.10 | 0.09 | 1.10 (1.01-1.19) | 3.0E-02 | 37 |
| | | | | GECCO | 0.10 | 0.09 | 1.09 (1.02-1.18) | 1.7E-02 | 0 |
| | | | | Combined | | | 1.13 (1.09-1.18) | 1.9E-09 | 4 |
| rs4414150 | 114278740 | A | T | AA | 0.20 | 0.17 | 1.17 (1.05-1.29) | 2.8E-03 | |
| | | | | JPN | 0.28 | 0.25 | 1.18 (1.09-1.28) | 3.3E-05 | |
| | | | | AA+JPN | | | 1.18 (1.11-1.25) | 3.1E-07 | 0 |
| | | | | CORECT | 0.10 | 0.09 | 1.09 (1.01-1.19) | 3.4E-02 | 38 |
| | | | | GECCO | 0.10 | 0.09 | 1.09 (1.02-1.18) | 1.6E-02 | 0 |
| | | | | Combined | | | 1.13 (1.08-1.18) | 7.3E-09 | 2 |
| rs4554811 | 114278734 | G | A | AA | 0.20 | 0.17 | 1.17 (1.05-1.29) | 2.8E-03 | |
| | | | | JPN | 0.28 | 0.25 | 1.19 (1.09-1.28) | 3.3E-05 | |
| | | | | AA+JPN | | | 1.18 (1.11-1.25) | 3.3E-07 | 0 |
| | | | | CORECT | 0.10 | 0.09 | 1.09 (1.01-1.19) | 3.4E-02 | 38 |
| | | | | GECCO | 0.10 | 0.09 | 1.09 (1.02-1.18) | 1.6E-02 | 0 |
| | | | | Combined | | | 1.13 (1.08-1.18) | 7.5E-09 | 2 |
| rs4554812 | 114278871 | T | A | AA | 0.22 | 0.19 | 1.19 (1.08-1.31) | 4.5E-04 | |
| | | | | JPN | 0.28 | 0.25 | 1.19 (1.09-1.28) | 3.2E-05 | |
| | | | | AA+JPN | | | 1.19 (1.12-1.26) | 5.2E-08 | 0 |
| | | | | CORECT | 0.10 | 0.09 | 1.09 (1.01-1.19) | 3.4E-02 | 39 |
| | | | | GECCO | 0.10 | 0.09 | 1.09 (1.02-1.17) | 1.7E-02 | 0 |
| | | | | Combined | | | 1.13 (1.09-1.18) | 2.0E-09 | 6 |
| rs78142102 | 114274269 | C | T | AA | 0.23 | 0.19 | 1.19 (1.08-1.31) | 3.9E-04 | |
| | | | | JPN | 0.28 | 0.25 | 1.18 (1.09-1.28) | 4.3E-05 | |
| | | | | AA+JPN | | | 1.19 (1.11-1.26) | 6.3E-08 | 0 |
| | | | | CORECT | 0.10 | 0.09 | 1.10 (1.01-1.19) | 3.1E-02 | 38 |
| | | | | GECCO | | | NA | | |
| | | | | Combined | | | 1.15 (1.10-1.21) | 1.8E-08 | 33 |
| rs7895362 | 114277173 | A | G | AA | 0.20 | 0.17 | 1.17 (1.06-1.30) | 2.6E-03 | |
| | | | | JPN | 0.28 | 0.25 | 1.18 (1.09-1.28) | 3.3E-05 | |
| | | | | AA+JPN | | | 1.18 (1.11-1.26) | 3.0E-07 | 0 |
| | | | | CORECT | 0.10 | 0.09 | 1.10 (1.01-1.19) | 3.0E-02 | 37 |
| | | | | GECCO | 0.10 | 0.09 | 1.10 (1.02-1.18) | 1.5E-02 | 0 |
| | | | | Combined | | | 1.13 (1.09-1.18) | 5.4E-09 | 1 |

OFCCR in GECCO was not included due to QC issues in this dataset.

**Supplementary Table 5.** Stratified analysis of rs12241008 by sex and site, stage and age at diagnosis**.** The risk allele was C.

| | In African Americans | | | | In Japanese | | | |
|---|---|---|---|---|---|---|---|---|
| | Cases | Controls | OR (95% CI) | P | Cases | Controls | OR (95% CI) | P |
| Females | 939 | 1658 | 1.06 (0.92-1.22) | 0.44 | 1106 | 1707 | 1.17 (1.03-1.33) | 0.01 |
| Males | 954 | 3044 | 1.32 (1.16-1.50) | 0.00003 | 1521 | 2090 | 1.20 (1.08-1.33) | 0.0006 |
| | | | | $P_{het}$ = 0.04 | | | | $P_{het}$ = 0.80 |
| Colon Cancer[1] | 1462 | 4702 | 1.18 (1.07-1.31) | 0.002 | 1709 | 3797 | 1.19 (1.09-1.30) | 0.0002 |
| Rectal Cancer[1] | 399 | 4702 | 1.19 (0.99-1.42) | 0.06 | 859 | 3797 | 1.20 (1.07-1.35) | 0.002 |
| | | | | $P_{het}$ [2]= 0.74 | | | | $P_{het}$ [2]= 0.93 |
| In situ/local | 309 | 4702 | 1.28 (1.05-1.55) | 0.03 | 925 | 3797 | 1.16 (1.04-1.29) | 0.006 |
| Regional/distant | 518 | 4702 | 1.19 (1.01-1.39) | 0.01 | 1091 | 3797 | 1.19 (1.07-1.34) | 0.002 |
| | | | | $P_{het}$ [2] = 0.73 | | | | $P_{het}$ [2] = 0.66 |
| Age at diagnosis (≤55) | 555 | 4702 | 1.05 (0.86-1.28) | 0.65 | 491 | 3797 | 1.22 (1.03-1.45) | 0.02 |
| Age at diagnosis (>55) | 1244 | 4702 | 1.21 (1.09-1.35) | 0.0005 | 2124 | 3797 | 1.19 (1.09-1.30) | $7.2\times10^{-5}$ |
| | | | | $P_{het}$ [2] = 0.87 | | | | $P_{het}$ [2] = 0.77 |

All analyses adjusted for age at blood draw, sex (where appropriate), the first 4 principal components and BMI.

[1] Cases with both colon and rectal diagnosis were excluded.

[2] From case-only analysis.

**Supplementary Table 6.** Association results for rs79453636 on Chromosome 7 in African Americans and in Japanese

| BP | Risk/ Other | study | Risk allele frequency | | OR (95% CI) | P | $R^2$ | $I^2$ (%) [a] |
|---|---|---|---|---|---|---|---|---|
| | | | Cases | Controls | | | | |
| 118250670 | C/T | AA | 0.097 | 0.065 | 1.48 (1.29-1.71) | $2.9\times10^{-8}$ | 0.99 | |
| | | JPN | 0.162 | 0.160 | 1.02 (0.92-1.12) | 0.73 | 1.00 | |
| | | AA+JPN | | | 1.15 (1.06-1.24) | $6.7\times10^{-4}$ | | 95 |
| | | CORECT+GECCO | NA | NA | 1.08 (0.96-1.21) | 0.18 | NA | NA |

[a] $I^2$ of heterogeneity

**Supplementary Table 7.** Association results 30 known colorectal cancer risk SNPs in the Japanese, African American and the combined GWAS

| | | | | Original GWAS | | Japanese GWAS | | | | | AA GWAS | | | | Japanese+AA GWAS | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Ref.** | **SNP** | **Locus** | **BP** | **Alleles[a]** | **OR** | **Alleles** | **OR (95% CI)** | **P** | **$R^2$** | **DC** | **Alleles[a]** | **OR** | **P** | **DC[e]** | **Alleles[a]** | **OR** | **P** | **$I^2$** | **DC[e]** |
| [1] | rs6691170 | 1q41 | 222045446 | T/G | 1.06 | na | | | | | T/G | 1.02 | 0.72 | | na | | | | |
| [1] | rs6687758 | 1q41 | 222164948 | G/A | 1.09 | G/A | 1.03 (0.96-1.12) | 0.42 | | | G/A | 0.99 | 0.86 | 1 | A/G | 0.98 | 0.59 | 0 | |
| [2] | rs11903757 | 2q32.3 | 192587204 | C/T | 1.16 | C/T | 1.09 (0.59-1.99) | 0.31 | 0.99 | | C/T | 0.99 | 0.84 | 1 | T/C | 0.98 | 0.67 | 0 | |
| [1] | rs10936599 | 3q26.2 | 169492101 | T/C | 0.93 | C/T | 0.96 (0.89-1.04) | 0.31 | | 1 | T/C | 0.96 | 0.62 | | T/C | 1.02 | 0.48 | 0 | 1 |
| [3 b] | rs647161 | 5q31.1 | 134499092 | A/C | 1.17 | A/C | 1.03 (0.96-1.11) | 0.41 | 0.97 | | A/C | 1.14 | **0.002** | | A/C | 1.08 | 0.007 | 69 | |
| [4] | rs1321311 | 6p21 | 36622900 | A/C | 1.10 | A/C | 1.04 (0.93-1.16) | 0.49 | | | A/C | 0.96 | 0.34 | 1 | A/C | 0.99 | 0.72 | 20 | 1 |
| [5 b,c] | rs7758229 | 6q26-q27 | 160840252 | T/G | 1.28 | T/G | 0.95 (0.84-1.07) | 0.38 | 1.00 | 1 | T/G | 1.23 | 0.056 | | T/G | 1.01 | 0.83 | 77 | |
| [6] | rs16892766 | 8q23.3 | 117630683 | C/A | 1.25 | na | | | | | C/A | 1.17 | **0.0058** | | na | | | | |
| [7] | rs10505477 | 8q24.21 | 128407443 | A/G | 1.18 | A/G | 1.13 (1.05-1.21) | **1.7E-03** | | | G/A | 0.91 | 0.09 | | A/G | 1.12 | 4.0E-04 | 0 | |
| [8] | rs6983267 | 8q24.21 | 128413305 | G/T | 1.21 | T/G | 0.88 (0.82-0.94) | **5.0E-04** | 1.00 | | T/G | 0.87 | **0.029** | | T/G | 0.87 | 4.1E-05 | 0 | |
| [9] | rs7014346 | 8q24.21 | 128424792 | A/G | 1.19 | A/G | 1.15 (1.06-1.25) | **8.3E-04** | | | A/G | 1.05 | 0.27 | | A/G | 1.10 | 1.9E-03 | 62 | |
| [6] | rs10795668 | 10p14 | 8701219 | A/G | 0.89 | A/G | 0.90 (0.84-0.97) | **4.1E-03** | 1.00 | | A/G | 0.98 | 0.77 | | A/G | 0.91 | 0.006 | 0 | |
| [4] | rs3824999 | 11q13.4 | 74345550 | C/A | 1.08 | G/T | 1.02 (0.95-1.09) | 0.60 | | | G/T | 1.15 | **0.009** | | T/G | 0.95 | 0.057 | 71 | |
| [9] | rs3802842 | 11q23.1 | 111171709 | C/A | 1.11 | C/A | 1.08 (1.00-1.16) | **0.039** | | | C/A | 1.04 | 0.40 | | A/C | 0.94 | 0.035 | 0 | |
| [3 b] | rs10774214 | 12p13.32 | 4368352 | T/C | 1.17 | C/T | 0.91 (0.85-0.98) | **0.011** | 0.97 | | C/T | 0.95 | 0.19 | | T/C | 1.08 | 0.005 | 0 | |
| [1] | rs7136702 | 12q13.13 | 50880216 | T/C | 1.06 | C/T | 1.05 (0.98-1.13) | 0.19 | | 1 | C/T | 0.97 | 0.41 | | T/C | 0.99 | 0.65 | 56 | 1 |
| [1] | rs11169552 | 12q13.13 | 51155663 | T/C | 0.92 | T/C | 1.00 (0.92-1.08) | 0.98 | | | T/C | 1.04 | 0.58 | 1 | T/C | 1.01 | 0.80 | 0 | 1 |
| [10] | rs4444235 | 14q22.2 | 54410919 | C/T | 1.11 | T/C | 0.96 (0.89-1.03) | 0.24 | | | C/T | 1.03 | 0.45 | | T/C | 0.96 | 0.16 | 0 | |
| [11] | rs1957636 | 14q22.2 | 54560018 | A/G | 1.08 | C/T | 0.97 (0.91-1.05) | 0.47 | | | C/T | 0.97 | 0.54 | | T/C | 1.03 | 0.34 | 0 | |
| [11] | rs16969681 | 15q13.3 | 32993111 | T/C | 1.18 | T/C | 1.04 (0.97-1.11) | 0.31 | 1.00 | | T/C | 1.16 | **0.010** | | T/C | 1.07 | 0.028 | 64 | |
| [12] | rs4779584 | 15q13.3 | 32994756 | T/C | 1.35 | C/T | 0.90 (0.82-0.98) | **0.020** | 1.00 | | C/T | 0.98 | 0.62 | | T/C | 1.06 | 0.057 | 51 | |
| [11] | rs11632715 | 15q13.3 | 33004247 | A/G | 1.12 | G/A | 1.03 (0.94-1.13) | 0.52 | | 1 | A/G | 1.04 | 0.34 | | A/G | 1.01 | 0.76 | 20 | |
| [10] | rs9929218 | 16q22.1 | 68820946 | A/G | 0.92 | A/G | 0.93 (0.85-1.02) | 0.13 | | | A/G | 0.93 | 0.12 | | A/G | 0.93 | 0.030 | 0 | |
| [13] | rs4939827 | 18q21.1 | 46453463 | C/T | 0.85 | T/C | 1.11 (1.02-1.21) | 0.017 | | | T/C | 1.08 | 0.096 | | T/C | 1.09 | 0.004 | 0 | |
| [10] | rs10411210 | 19q13.1 | 33532300 | T/C | 0.87 | T/C | 0.98 (0.89-1.08) | 0.66 | | | T/C | 0.94 | 0.11 | | T/C | 0.95 | 0.13 | 0 | |
| [10] | rs961253 | 20p12.3 | 6404281 | A/C | 1.12 | A/C | 1.10 (0.98-1.24) | 0.09 | 1.00 | | A/C | 1.09 | 0.054 | | A/C | 1.09 | 0.011 | 0 | |
| [11] | rs4813802 | 20p12.3 | 6699595 | G/T | 1.09 | G/T | 1.09 (1.00-1.19) | 0.053 | | | G/T | 1.10 | 0.11 | | T/G | 0.92 | 0.012 | 0 | |
| [3 b] | rs2423279 | 20p12.3 | 7812350 | C/T | 1.14 | na | | | | | C/T | 0.94 | 0.19 | 1 | na | | | | |
| [1] | rs4925386 | 20q13.33 | 60921044 | T/C | 0.93 | T/C | 0.98 (0.90-1.08) | 0.73 | | | C/T | 1.05 | 0.27 | | T/C | 0.97 | 0.31 | 0 | |
| [4] | rs5934683 | 23p22.2 | 9751474 | T/C | 1.07 | C/T | 0.93 (0.81-1.07) | 0.32 | | | C/T | 1.00 | 0.93 | | T/C | 1.03 | 0.51 | 0 | |

AA: African Americans.
Highlighted are P-values < 0.05.

$R^2$ is shown for imputed markers only
[a] Risk/other allele
[b] In East Asians. Other previous associations were reported in Europeans
[c] Left colon only
[d] Risk allele frequency
[e] DC: Direction change indicator. "1" indicates change in OR direction between the corresponding analysis and the original published GWAS. Blank means no change.

**Supplementary Table 8.** Functional annotation of rs12241008 and SNPs in high LD ($r^2$>0.8) in Europeans, Africans, and Asians. Tables were generated with HaploReg (http://www.broadinstitute.org/mammals/haploreg/haploreg.php) [14].

**Query SNP: rs12241008 and variants with $r^2$ >= 0.8 (European)**

| | | LD | | | | | Frequency | | | SiPhy | Histone Marks | | DNAse | Proteins | Predicted Regulatory | Ref Seq | dbSNP |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Chr | Pos (hg19) | ($r^2$) | (D') | Variant | Ref | Alt | AFR | ASN | EUR | Conserv | Promoter | Enhancer | Hypersensitivity | Bound | Motifs Changed | genes | func annot |
| 10 | 114270474 | 0.91 | 0.95 | rs10509964 | A | T | 0.13 | 0.3 | 0.09 | | | | | | CEBPA,SIX5 | VTI1A | intronic |
| 10 | 114274269 | 0.91 | 0.95 | rs78142102 | T | C | 0.16 | 0.3 | 0.09 | | | | | | 7 altered motifs | VTI1A | intronic |
| 10 | 114274559 | 0.91 | 0.95 | rs34139797 | A | G | 0.17 | 0.3 | 0.09 | | | | | | DMRT1,Mef2 | VTI1A | intronic |
| 10 | 114277039 | 0.94 | 0.97 | rs7894915 | C | G | 0.16 | 0.3 | 0.09 | | NHEK | Huvec | | | BCL | VTI1A | intronic |
| 10 | 114277173 | 0.94 | 0.97 | rs7895362 | G | A | 0.13 | 0.3 | 0.09 | | NHEK | Huvec | MCF-7 | | 8 altered motifs | VTI1A | intronic |
| 10 | 114278181 | 0.92 | 0.97 | rs10082356 | A | G | 0.16 | 0.3 | 0.09 | | | NHEK, H1, GM12878 | HCT-116 | | | VTI1A | intronic |
| 10 | 114278734 | 0.98 | 1 | rs4554811 | A | G | 0.14 | 0.3 | 0.09 | | | GM12878, | | | 14 altered motifs | VTI1A | intronic |
| 10 | 114278740 | 0.98 | 1 | rs4414150 | T | A | 0.13 | 0.3 | 0.09 | | | GM12878, NHEK | | | 14 altered motifs | VTI1A | intronic |
| 10 | 114278871 | 0.98 | 1 | rs4554812 | A | T | 0.16 | 0.3 | 0.09 | | | GM12878, NHEK | | | DMRT5, Irf, Pou1f1, RFX5 | VTI1A | intronic |
| 10 | 114280318 | 1 | 1 | rs17129834 | T | C | 0.13 | 0.3 | 0.09 | | | | SK-N-MC | | | VTI1A | intronic |
| **10** | **114280702** | **1** | **1** | **rs12241008** | **T** | **C** | **0.16** | **0.3** | **0.09** | | | | | | **Esr2,Myc,RFX5** | **VTI1A** | **intronic** |
| 10 | 114285498 | 0.92 | 1 | rs113979074 | T | 11-mer | 0.16 | 0.28 | 0.09 | | | NHEK, GM12878, HMEC, HSMM | | | 14 altered motifs | VTI1A | intronic |
| 10 | 114286188 | 1 | 1 | rs12240720 | T | C | 0.16 | 0.29 | 0.09 | | | NHEK, GM12878, HMEC, HSMM, H1 | 18 cell types | BATF, PAX5C20, POL2H8, CFOS, FOSL2, FOXA1, STAT3 | Egr-1,Pax-5,Pax-6 | VTI1A | intronic |
| 10 | 114288619 | 0.97 | 1 | rs12246635 | T | C | 0.16 | 0.29 | 0.1 | conserved | | | | | Foxp3,GATA,Mrg | VTI1A | intronic |
| 10 | 114289017 | 0.94 | 0.97 | rs11195986 | G | T | 0.16 | 0.28 | 0.09 | | | | | | GR,HNF4,RXRA | VTI1A | intronic |
| 10 | 114291787 | 0.94 | 0.97 | rs17129851 | G | A | 0.13 | 0.26 | 0.09 | | NHLF | GM12878 | | | 11 altered motifs | VTI1A | intronic |
| 10 | 114292303 | 0.94 | 0.97 | rs4439449 | G | A | 0.13 | 0.28 | 0.09 | | NHLF | GM12878 | Th1, AoSMC, GM18507, GM19240, GM12864 | BATF, MEF2A, MEF2C, OCT2, PAX5N19, POU2F2, SRF | 5 altered motifs | VTI1A | intronic |
| 10 | 114294892 | 0.93 | 0.97 | rs12255141 | A | G | 0.16 | 0.29 | 0.1 | | | | | | HDAC2,Nanog | VTI1A | intronic |
| 10 | 114296308 | 0.93 | 0.97 | rs73365049 | C | T | 0.16 | 0.29 | 0.1 | | | Huvec | | | Nkx6-1,Pdx1,Pou4f3 | VTI1A | intronic |
| 10 | 114299884 | 0.93 | 0.98 | rs58575880 | T | C | 0.32 | 0.29 | 0.1 | | | | | | 4 altered motifs | VTI1A | intronic |

**Query SNP: rs12241008 and variants with $r^2$ >= 0.8 (African)**

| | | LD | | | | | Frequency | | | SiPhy | Histone Marks | | DNAse | Proteins | Predicted Regulatory | Ref Seq | dbSNP |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Chr | Pos (hg19) | ($r^2$) | (D') | Variant | Ref | Alt | AFR | ASN | EUR | Conserv | Promoter | Enhancer | Hypersensitivity | Bound | Motifs Changed | genes | func annot |
| 10 | 114274269 | 0.93 | 0.97 | rs78142102 | T | C | 0.16 | 0.3 | 0.09 | | | | | | 7 altered motifs | VTI1A | intronic |
| 10 | 114274559 | 0.89 | 0.97 | rs34139797 | A | G | 0.17 | 0.3 | 0.09 | | | | | | DMRT1,Mef2 | VTI1A | intronic |
| 10 | 114277039 | 0.94 | 0.97 | rs7894915 | C | G | 0.16 | 0.3 | 0.09 | | NHEK | Huvec | | | BCL | VTI1A | intronic |
| 10 | 114278181 | 0.94 | 0.97 | rs10082356 | A | G | 0.16 | 0.3 | 0.09 | | | NHEK, H1, GM12878 | HCT-116 | | | VTI1A | intronic |
| 10 | 114278871 | 0.99 | 1 | rs4554812 | A | T | 0.16 | 0.3 | 0.09 | | | GM12878, NHEK | | | DMRT5, Irf, Pou1f1, RFX5 | VTI1A | intronic |
| **10** | **114280702** | **1** | **1** | **rs12241008** | **T** | **C** | **0.16** | **0.3** | **0.09** | | | | | | **Esr2,Myc,RFX5** | **VTI1A** | **intronic** |
| 10 | 114285498 | 0.9 | 0.97 | rs113979074 | T | 11-mer | 0.16 | 0.28 | 0.09 | | | NHEK, GM12878, HMEC, HSMM | | | 14 altered motifs | VTI1A | intronic |
| 10 | 114286188 | 0.97 | 0.99 | rs12240720 | T | C | 0.16 | 0.29 | 0.09 | | | NHEK, GM12878, HMEC, HSMM, H1 | 18 cell types | BATF, PAX5C20, POL2H8, CFOS, FOSL2, | Egr-1,Pax-5,Pax-6 | VTI1A | intronic |
| 10 | 114288619 | 0.99 | 1 | rs12246635 | T | C | 0.16 | 0.29 | 0.1 | conserved | | | | | Foxp3,GATA,Mrg | VTI1A | intronic |
| 10 | 114289017 | 0.99 | 1 | rs11195986 | G | T | 0.16 | 0.28 | 0.09 | | | | | | GR,HNF4,RXRA | VTI1A | intronic |
| 10 | 114294892 | 0.93 | 0.97 | rs12255141 | A | G | 0.16 | 0.29 | 0.1 | | | | | | HDAC2,Nanog | VTI1A | intronic |
| 10 | 114296308 | 0.91 | 0.96 | rs73365049 | C | T | 0.16 | 0.29 | 0.1 | | | Huvec | | | Nkx6-1,Pdx1,Pou4f3 | VTI1A | intronic |

**Query SNP: rs12241008 and variants with $r^2$ >= 0.8 (Asian)**

| | | LD | | | | | Frequency | | | SiPhy | Histone Marks | | DNAse | Proteins | Predicted Regulatory | Ref Seq | dbSNP |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Chr | Pos (hg19) | ($r^2$) | (D') | Variant | Ref | Alt | AFR | ASN | EUR | Conserv | Promoter | Enhancer | Hypersensitivity | Bound | Motifs Changed | genes | func annot |
| 10 | 114259714 | 0.8 | 0.95 | rs77705687 | A | G | 0.02 | 0.33 | 0.06 | | | K562 | | | GATA | VTI1A | intronic |
| 10 | 114260084 | 0.8 | 0.95 | rs58936340 | G | C | 0.06 | 0.33 | 0.06 | | | | | TCF4 | 4 altered motifs | VTI1A | intronic |
| 10 | 114262773 | 0.9 | 0.95 | rs111936498 | A | T | 0.02 | 0.3 | 0.07 | | | | | | Irx | VTI1A | intronic |
| 10 | 114265071 | 0.9 | 0.95 | rs77802795 | T | C | 0.02 | 0.3 | 0.07 | | | | | | Foxf1,Foxi1,Foxq1 | VTI1A | intronic |
| 10 | 114269239 | 0.93 | 0.97 | rs12263620 | A | G | 0.13 | 0.3 | 0.09 | | | | | | 5 altered motifs | VTI1A | intronic |
| 10 | 114270474 | 0.94 | 0.98 | rs10509964 | A | T | 0.13 | 0.3 | 0.09 | | | | | | CEBPA,SIX5 | VTI1A | intronic |
| 10 | 114274269 | 0.93 | 0.97 | rs78142102 | T | C | 0.16 | 0.3 | 0.09 | | | | | | 7 altered motifs | VTI1A | intronic |
| 10 | 114274559 | 0.94 | 0.97 | rs34139797 | A | G | 0.17 | 0.3 | 0.09 | | | | | | DMRT1,Mef2 | VTI1A | intronic |
| 10 | 114276661 | 0.96 | 0.98 | rs76880073 | A | G | 0.01 | 0.3 | 0.07 | | | NHEK | | | GATA,HNF1 | VTI1A | intronic |
| 10 | 114277039 | 0.96 | 0.98 | rs7894915 | C | G | 0.16 | 0.3 | 0.09 | | NHEK | Huvec | | | BCL | VTI1A | intronic |
| 10 | 114277173 | 0.96 | 0.98 | rs7895362 | G | A | 0.13 | 0.3 | 0.09 | | NHEK | Huvec | MCF-7 | | 8 altered motifs | VTI1A | intronic |
| 10 | 114278181 | 0.96 | 0.98 | rs10082356 | A | G | 0.16 | 0.3 | 0.09 | | | NHEK, H1, GM12878 | HCT-116 | | | VTI1A | intronic |
| 10 | 114278734 | 0.98 | 1 | rs4554811 | A | G | 0.14 | 0.3 | 0.09 | | | GM12878, NHEK | | | 14 altered motifs | VTI1A | intronic |
| 10 | 114278740 | 0.98 | 1 | rs4414150 | T | A | 0.13 | 0.3 | 0.09 | | | GM12878, NHEK | | | 14 altered motifs | VTI1A | intronic |
| 10 | 114278871 | 0.98 | 1 | rs4554812 | A | T | 0.16 | 0.3 | 0.09 | | | GM12878, NHEK | | | 4 altered motifs | VTI1A | intronic |
| 10 | 114280318 | 1 | 1 | rs17129834 | T | C | 0.13 | 0.3 | 0.09 | | | | SK-N-MC | | | VTI1A | intronic |
| **10** | **114280702** | **1** | **1** | **rs12241008** | **T** | **C** | **0.16** | **0.3** | **0.09** | | | | | | **Esr2,Myc,RFX5** | **VTI1A** | **intronic** |
| 10 | 114283116 | 0.99 | 1 | rs17129837 | C | T | 0.02 | 0.3 | 0.06 | | | | | | 4 altered motifs | VTI1A | intronic |
| 10 | 114284643 | 0.85 | 0.94 | rs7081965 | A | T | 0.47 | 0.31 | 0.24 | conserved | | NHEK, GM12878, HMEC, HSMM, H1, NHLF | BJ, HIPEpiC, NHLF | | Cdx2,Hoxa10,Hoxd10 | VTI1A | intronic |
| 10 | 114285498 | 0.86 | 0.98 | rs113979074 | T | 11-mer | 0.16 | 0.28 | 0.09 | | | NHEK, GM12878, HMEC, HSMM | | | 14 altered motifs | VTI1A | intronic |
| 10 | 114286188 | 0.94 | 1 | rs12240720 | T | C | 0.16 | 0.29 | 0.09 | | | NHEK, GM12878, HMEC, HSMM, H1 | 18 cell types | BATF, PAX5C20, POL2H8, CFOS, FOSL2, FOXA1, STAT3 | Egr-1,Pax-5,Pax-6 | VTI1A | intronic |
| 10 | 114288619 | 0.93 | 1 | rs12246635 | T | C | 0.16 | 0.29 | 0.1 | conserved | | | | | Foxp3,GATA,Mrg | VTI1A | intronic |
| 10 | 114289017 | 0.92 | 1 | rs11195986 | G | T | 0.16 | 0.28 | 0.09 | | | | | | GR,HNF4,RXRA | VTI1A | intronic |
| 10 | 114291787 | 0.8 | 1 | rs17129851 | G | A | 0.13 | 0.26 | 0.09 | | NHLF | GM12878 | | | 11 altered motifs | VTI1A | intronic |
| 10 | 114292303 | 0.92 | 1 | rs4439449 | G | A | 0.13 | 0.28 | 0.09 | | NHLF | GM12878 | Th1, AoSMC, GM18507, GM19240, GM12864 | BATF, MEF2A, MEF2C, OCT2, PAX5N19, POU2F2, SRF | 5 altered motifs | VTI1A | intronic |
| 10 | 114294892 | 0.92 | 0.99 | rs12255141 | A | G | 0.16 | 0.29 | 0.1 | | | | | | HDAC2,Nanog | VTI1A | intronic |
| 10 | 114296308 | 0.92 | 0.99 | rs73365049 | C | T | 0.16 | 0.29 | 0.1 | | | Huvec | | | Nkx6-1,Pdx1,Pou4f3 | VTI1A | intronic |
| 10 | 114297803 | 0.9 | 0.98 | rs142420348 | AG | A | 0.02 | 0.29 | 0.06 | conserved | | | | | AP-1 | VTI1A | intronic |
| 10 | 114299884 | 0.89 | 0.97 | rs58575880 | T | C | 0.32 | 0.29 | 0.1 | | | | | | 4 altered motifs | VTI1A | intronic |
| 10 | 114306441 | 0.81 | 0.95 | rs80207046 | A | T | 0.03 | 0.28 | 0.07 | | | GM12878, Huvec | 9 cell types | 4 bound proteins | 5 altered motifs | VTI1A | intronic |
| 10 | 114307546 | 0.86 | 0.96 | rs74236144 | T | C | 0.01 | 0.29 | 0.07 | | | | | | 7 altered motifs | VTI1A | intronic |

## Supplementary Methods

### Japanese subjects, genotyping and Quality Control (QC)

The Multiethnic Cohort (MEC). The MEC includes >215,000 men and women aged 45-75 years at recruitment from five different racial/ethnic groups (African Americans, Japanese Americans, Native Hawaiians, Latinos and European Americans) in Hawaii and California [15]. The cohort was assembled in 1993-1996 by mailing a self-administered, 26-page questionnaire to obtain extensive information on demographics, medical and reproductive histories, medication use, family history of various cancers, physical activity and diet. Identification of incident cancer cases is by regular linkage with the Hawaii and Los Angeles County and California SEER registries. MEC subjects re-contacted mostly from 1995 to 2001 for blood collection included incident cases with breast, prostate, or colorectal cancers, as well as a random sample of cohort participants to serve as controls in nested genetic case-control studies (participation rate 72% and 63%, respectively). From 2001 to 2006, blood was also collected prospectively, without regard for cancer diagnosis, from willing cohort participants (participation rate 43%).

In the first batch (MEC1), 1,703 Japanese MEC CRC cases and controls were genotyped on the Illumina 1M-duo arrays by the Broad Genotyping Center (Boston, Massachusetts). Data on 1,602 subjects (803 cases, 799 controls) with call rates >97% (after zeroing out intensity-only probes or monomorphic loci) for 1,007,656 assays were released to us. Twenty-four ethnicity outliers and 58 relatives were removed and 1,520 subjects (761 cases, 759 controls) remained. Markers were excluded if call rates were <95%, more than one Mendelian error (7 HapMap trios) or >3 discordant genotype pairs between duplicate scans (>400 duplicate pairs) were observed, P for Hardy-Weinberg Equilibrium (HWE) was $< 10^{-5}$ in controls or if genotyping clustering was of poor quality (these were checked randomly). 993,804 markers passed the above filters. The average concordance rate was >99.99% (median 100%) among duplicate pairs.

Five plates with relatively lower average call rates from the first batch were re-clustered by the Broad using a customized clustering definition file. After ~440k SNPs with zero calls were removed, 42 samples with call rates >97% were recovered – we name these data as MEC2. We restricted the analysis to the 993,804 markers retained in MEC1 and excluded SNPs with call rates < 95% within this subset and SNPs with one or more discordance among 9 duplicate pairs and kept 703,817 markers. The mean concordance rate among duplicates was 99.97% (median 100%). In addition, 1 subject was removed due to being a sibling to another subject. This left 41 subjects (23 cases, 18 controls).

To provide a larger control pool, 1,033 prostate cancer-free men with data on 527,110 markers and 808 breast cancer-free women with data on 488,282 markers were drawn from the MEC prostate cancer (MEC-PrCa) [16]and breast cancer (MEC-BrCa) [17]studies. These subjects were genotyped using the Illumina 660W-quad arrays and pre-filtered using similar criteria as above. Allele frequencies were compared between MEC-PrCa and MEC-BrCa and 568 SNPs (among 487,373 overlapping variants) with differences > 0.1 were excluded.

Then all MEC data were combined. Among the 116 duplicate pairs and based on the overlapping 504,262 SNPs, the mean concordance rate was 99.94% (all >99.4%). These 116 duplicated records and another 140 subjects from MEC-PrCa or MEC-BrCa were removed to eliminate close relatives (≥ 2nd degree), resulting in 861 men (17 cases, 844 controls) from MEC-PrCa and 724 women (5 cases, 719 controls) from MEC-BrCa. We again restricted to the 993,804 markers in MEC1 and removed 2 SNPs with three allele mismatches in the combined data, 205,644 SNPs with MAFs <0.01, 1,247 variants dropped in Illumina's new manifest file (version _H for 1M-duo) and 10 SNPs for which allele frequency differences between MEC1 and Fukuoka (see below) were > 0.45 (9 of them had non-complimentary alleles). The number of variants contributed by each batch was: 786,901 from MEC1, 558,151 from MEC2, 450,114 from MEC-PrCa and 446,586 from MEC-BrCa. There were 3,146 MEC subjects (806 cases, 2,340 controls) remaining at this point.

Colorectal Cancer Family Registry (CCFR). The CCFR is a consortium initiated in 1997 that is dedicated to the establishment of a collaborative infrastructure for interdisciplinary studies in the genetic epidemiology of CRC [18]. Participating sites use standardized instruments and protocols to collect family history information, epidemiological and clinical data, screening behavior, and related biological specimens (blood samples and tumor blocks). Cases were ascertained from both population-based cancer registries and family clinics with a diagnosis of invasive adenocarcinoma of the colon or rectum. Since 2004, the minority component of the CCFR has been expanded to recruit additional African American and Japanese American families in Hawaii, Northern and Southern California and in North Carolina. The numbers of CCFR Japanese American subjects included in this study were 520 from Hawaii (240 cases, 280 controls), 40 from Seattle (40 cases), 78 from University of Southern California (USC) (59 cases, 19 controls) and 59 from Cancer Prevention Institute of California (CPIC, formerly Northern California Cancer Center) (45 cases, 14 controls). Altogether 697 CCFR subjects (384 cases, 313 controls) were genotyped as part of the Japanese GWAS.

Colorectal cancer study on Oahu, Hawaii (CR2&3). Study details have been described in [19]. Incident adenocarcinoma cases of Japanese, European or Native Hawaiian ancestry, residing on Oahu, Hawaii and diagnosed during 1994–99 were identified through the Hawaii Surveillance, Epidemiology and End Results (SEER) cancer registry. We only included 155 cases of Japanese ancestries for the genotyping.

The Fukuoka Colorectal Cancer Study is a community-based case-control study of incident colorectal adenocarcinomas [20,21]. Cases were recruited from two university hospitals and six affiliated hospitals in the study area (Fukuoka City and three adjacent areas), and controls were randomly selected in the study area by frequency-matching to the expected distribution of incident cases with respect to sex and 10-year age class. Cases were patients aged 20-74 years (at diagnosis) who were admitted to a participating hospital for surgical treatment during the period from September 2000 to December 2003. The patients were excluded in the study if they lived outside the study area or if they had a prior history of removal of the colorectum or a morbid condition of familial adenomatous polyposis or inflammatory bowel disease. Of 1,053 eligible cases, a total of 840 cases (80%) participated in the interview, and 685 out of them gave an informed consent to genotyping. A total of 1,500 persons aged 20-74 years were selected as control candidates by two-stage random sampling based on resident registry. Of 1,382 eligible controls, 833 (60%) participated in the interview from January 2001 to December 2002, and 778 gave an informed consent to genotyping. The cases and controls were interviewed in person regarding lifestyle factors including alcohol use, smoking, and diet. DNA was extracted from the buffy coat by

using a commercial kit (QIAGEN GmbH, Hilden, Germany). 1,463 subjects (685 cases, 778 controls) were genotyped.

Nagano Colorectal Cancer Study. A hospital-based case-control study [22] was conducted between October 1998 and March 2002 at four hospitals in Nagano Prefecture, Japan. Eligible cases were patients aged 20–74 years with newly diagnosed, histologically confirmed colorectal cancer during the survey at those hospitals. We enrolled 121 colorectal cancer cases. No patient refused to participate. Healthy controls who were confirmed as not having any cancer were selected from medical checkup examinees in the four hospitals. Two controls were matched to each case on sex, age (within three years), hospital and residential area and 245 controls participated (participation rate = 99%). Participants completed a self-administered questionnaire, which included questions on demographic characteristics, anthropometric factors, smoking habits, family history of cancer, and medical history. Dietary habits were investigated using a 141-item semi-quantitative food-frequency questionnaire, which was developed and validated in the Japanese population. Blood samples were collected at the same time as the questionnaire, and a buffy coat was preserved at –80°C until analysis. This study was approved by the Institutional Review Board of the National Cancer Center, Tokyo, Japan. In this study, we excluded subjects whose DNA samples were not available and selected one control per case. 212 subjects (106 cases, 106 controls) were genotyped.

Genotyping for the CCFR, CR2&3, Fukuoka and Nagano studies. 2,527 subjects (1,330 cases, 1,197 controls) were genotyped by the USC Genomics Center (Los Angeles, CA) for 1,039,513 markers. We removed 381 subjects due to call rates <97% (n=33) or gender mismatch (n=30), 279 relatives, 31 ethnicity outliers, 4 subjects missing key covariates and 4 subjects who appeared as MZ twins but covariate information did not support the relationship. 2,146 subjects (1,235 cases, 911 controls) remained. Concordance rates among all 36 duplicate pairs were > 99.99% (median 100%). Markers were excluded if the following was met: 1) call rate <95%, 2) any discordance between duplicate pairs, 3) p-value for HWE test among Nagano or Fukuoka controls $< 10^{-4}$, 5) alleles did not match those in MEC data, 6) CNVs, 7) MAF < 0.01 in this dataset (n=244,750), 8) 10 SNPs for which allele frequency differences between MEC1 and Fukuoka were large (> 0.45) and 9) variants dropped in Illumina's new manifest file (version _H for 1M-duo). These resulted in 790,064 available variants.

Japan Public Health Center-based prospective study (JPHC). The JPHC study [23,24] was initiated in 1990 and includes 140,420 subjects aged 40-69 years living in municipalities supervised by 11 public health centers. Approximately 80% of the public health center participants returned a self-administered questionnaire on demographic characteristics, personal and familial medical histories, anthropometric factors, physical activity, smoking and drinking habits, as well as diet. About 35% provided a blood sample at the baseline survey during 1990-95. This study was approved by the institutional review board of the National Cancer Center, Tokyo, Japan. In this study, we excluded ineligible subjects and subjects in two public health centers because one center did not collect information on cancer incidence and the other did not have DNA samples available. The 32,989 participants who did not report any diagnosis of cancer in the baseline questionnaire and provided a blood sample were followed until the end of 2009. During follow-up, 675 cases of colorectal cancer were identified. For each case, one control was selected, using incidence density sampling, from subjects who had no prior history of colorectal cancer at the time when the case was diagnosed. Controls were matched to cases on sex, age (within 3 years), date of blood draw (within 3 months), time since last meal (within 4 hours), and study location.

1,332 JPHC subjects (670 cases, 662 controls) were genotyped by the USC Epigenomics Center on the Illumina 1M-duo arrays for 1,199,064 markers. Thirty-two subjects were excluded due to call rates <90% (n=2), gender mismatch (n=13) and relatedness checking (n=17). 1,300 subjects (657 cases, 643 controls) remained. The mean concordance rate among 15 duplicate pairs was > 99.99% (all > 99.96%). Markers were dropped based on the following: 1) call rate < 90%, 2) any discordance among duplicate pairs, 3) p-value for HWE test among controls $< 10^{-5}$, and 4) MAF < 0.005 in this dataset and not present in any of the above studies (n=231,949). 782,650 variants were retained for analysis.

Combining data. In combining all 6,592 subjects, we removed 154 closer-than-second-degree relatives and 14 ethnicity outliers using a stricter criterion based on PCs. 6,424 subjects (2,627 cases, 3,797 controls) were used in data analysis. A summary of the number of subjects and markers contributed by each study/batch is shown in Supplementary Table 1. 323,968 genotyped markers overlapped across all batches.

**African American subjects, genotyping and QC**

A total of 5,062 MEC (442 CRC cases, 4,620 controls) African American subjects gentoyped on the Illumina 1M-duo array were included. Among them, 4,567 were genotyped for GWAS of prostate cancer [25] and breast cancer [26] (about half of them are prostate or breast cancer cases, the other half controls). 132 of these 4,567 had a CRC diagnosis and were used as cases for the CRC GWAS. The remaining subjects were used as controls. Also, 1,289 CCFR subjects (999 cases, 290 controls) and subjects from the 4 studies below, genotyped on the Illumina 1M-duo array, were included.

The Southern Community Cohort Study (SCCS). The SCCS is a prospective cohort investigation initiated in 2001 enrolling residents aged 40-79 years across 12 southern states [27]. The large majority (86%) of participants were enrolled at community health centers (CHCs), institutions providing basic and preventative health care in underserved communities, so the cohort includes low income segments of society typically not included in large numbers in other cohorts. Both African and non-African Americans are included, with more than twice as many African Americans enrolled to help address under-representation of blacks in previous epidemiologic studies of cancer. Study participants completed a detailed baseline questionnaire, via in-person computer-assisted interviews at CHCs, and nearly 90% provided a biologic specimen. Follow-up of the cohort and identification of cancer cases was conducted with national mortality registers and with linkage to state cancer registries. We included 164 AA CRC cases and a stratified matched (on age and sex) random sample of 160 controls with available DNA in this analysis.

MD Anderson data. The cases were from a series of histopathologically confirmed colorectal cancer cases of all race/ethnicities enrolled in TexGen at MD Anderson Cancer Center from 2002 to 2010. Under the TexGen protocol, patients were consented to provide blood for future medical research at the Texas Medical Center. Participants also completed a baseline questionnaire that included demographic and core risk factor questions. More details can be found in [28]. There were 189 AA patients with available DNA and questionnaire data that were included for genotyping.

University of North Carolina CanCORS study (UNC-CanCORS). UNC_CanCORS assembled a prospective population-based cohort of colorectal cancer patients from a 33-county area in North Carolina diagnosed between April 2003 and January 2005. Cases were identified using the North Carolina Central Cancer Registry. 84 AA cases were included for genotyping.

The North Carolina Rectal Cancer Study (UNC-Rectal). The study included population-based cases from a 33-county area in central and eastern North Carolina. Individuals with a first diagnosis of histologically confirmed sigmoid colon, rectosigmoid, or rectal adenocarcinoma between May 2001 and September 2006 were identified from the North Carolina Central Cancer Registry. Cases were between the ages of 40 and 80. Controls were selected from two sources: North Carolina Department of Motor Vehicles records (for controls under the age of 65) and Center for Medicare and Medicaid Services records (for controls age 65 and older). The controls were frequency matched to cases on age, sex and race. 112 cases and 108 controls were included.

Among the 7,168 study participants described above, subjects were excluded based on the following criteria: 1) call rates < 95% ($n$ = 167), 2) no age information ($n$ = 2), 3) gender mismatch ($n$ = 39), i.e. when the reported gender is different from that estimated based on X chromosome inbreeding coefficient $F$ (calculated by PLINK), 4) ancestry outliers ($n$ = 98) based on principal components (discussed below), and 5) closer than 2nd-degree relatives ($n$ = 435), where relationships were derived from estimated probabilities of sharing 0, 1, or 2 allele based on genomic data (calculated by PLINK). Relatives were removed in the following order or priority: 1) subjects with many relatives, 2) controls, 3) samples with lower call rates. Sample replicates (2%) were included and the average concordance rates were > 99.9%. Starting from 1,192,666 markers, we excluded markers of poor clustering property or with call rates < 95%, MAFs < 0.005, more than one discordant pair among sample replicates, and $p$-values < $10^{-10}$ from HWE test in MEC controls and of poor clustering quality. These resulted in 6,427 subjects (4,609 controls, 1,818 cases) on 1,049,327 markers.

Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial (PLCO). In addition, we included genotypes on 160 PLCO subjects (76 cases, 94 controls). The PLCO is a multi-institution study sponsored and coordinated by the National Cancer Institute (NCI) with more than 154,000 men and women (http://dcp.cancer.gov/plco). Participants aged 55-74 joined the trial between 1992 and 2001 at 10 centers across the country. Cases were ascertained through annual questionnaires mailed to the participants, and positive reports were followed-up through abstraction of medical records or death certificates. Half of these participants were selected by chance to receive specific screening tests and half to receive routine care by their personal health care providers. All participants provide regular updates on a variety of health measurements in an annual questionnaire, as well as blood samples for use in studies of biologic markers of cancer risk. PLCO subjects (n = 171) were genotyped by the NCI genotyping center on the Illumina Omni 2.5M array and were pre-filtered by the PLCO data coordinating center using similar criteria as described above. One subject was excluded due to relatedness. For this analysis, 527,383 markers overlapped with those from the other batch were retained. Allele frequencies matched well between PLCO and MEC controls (only 77 differences were > 0.1 and all were ≤ 0.13).

**Replication in European-descent populations**

GWAS IN CORECT

The CORECT study meta-analysis was conducted using germline DNA from 6 observational studies, namely, the Molecular Epidemiology of Colorectal Cancer study (MECC) [29], CCFR (described above), Kentucky case-control study (Kentucky) [30], Newfoundland case-control study (Newfoundland), American Cancer Society CPS II nested case-control study (ACS/CPSII) [31] and the Melbourne nested

case-control study (Melbourne)[32]. All subjects were self-reported whites. Summaries of the 6 studies and the platforms used to generate genotype data are in Supplementary Table 3.

*The Molecular Epidemiology of Colorectal Cancer Study* (MECC) [29] is a population-based, case-control study of pathologically-confirmed, incident cases of CRC recruited from a geographically-defined region of northern Israel. Participant recruitment began in 1998 and remains on-going. Individually-matched controls with no prior history of CRC are selected from the same source population that gave rise to cases using the Clalit Health Services database. Matching factors include age, sex, Jewish ethnicity (Jew versus non-Jew), and primary clinic site. Subjects are interviewed to obtain demographic and clinical information, family history, and dietary habits. Biospecimens including blood, paraffin blocks, and snap frozen tumors are also collected. MECC1 leverages data on 484 cases and 498 controls. Case selection excluded microsatellite instable (MSI-H) tumors. MECC2 utilizes genotypes from 1,120cases and 820 controls. We examined the correspondence between self-reported ancestry and genotypic classification and only included cases and controls that clustered with Ashkenazi Jews based on PCs.

*The Kentucky study* was initiated in July 2003 through the University of Kentucky Cancer Center [30]. A web-based reporting system implemented by the Kentucky Cancer Registry in 2003 has facilitated rapid report of cases state- wide, with approximately 76.8% of all cases reported to the registry within 6 months of diagnosis. Cases (>21 years) diagnosed with histologically confirmed colon cancer and entered into the registry within 6 months of their diagnoses are invited to join the study. Unrelated controls are recruited through random digit dialing and are frequency matched to the cases by age (±5 years), gender, and race. Excluded from the study are those individuals who have been diagnosed with colon cancer because of known hereditary forms of colon cancer or polyposis such as familial adenomatous polyposis (FAP), hereditary non-polyposis colorectal cancer (HNPCC), Peutz-Jeghers, and Cowden disease. Currently there are more than 1,040 incident population-based cases of colorectal cancer and 1,750 population-based controls fully recruited, with comprehensive epidemiologic data, pathology data, and DNA from cases and controls.

*American Cancer Society Cancer Prevention Study II (ACS/CPSII)* is a cohort study [31]started by the American Cancer Society in 1982 to investigate the relationship between dietary, lifestyle and other etiologic factors and cancer mortality. Approximately 1.2 million men and women were enrolled in the study from 50 states in the U.S. In 1992, a subset of these (~184,000) participants were enrolled in the CPS-II Nutrition Cohort to examine the relationship between dietary and other exposures and cancer incidence. Blood samples were drawn from approximately 39,376 members of the Nutrition Cohort from 1998 to 2001, and buccal cells were collected from 69,467 members from 2001 to 2002. Cancer cases are identified by self-report through follow-up questionnaires followed by verification through medical records and/or linkage to state cancer registries as well as death certificates. CRC cases and controls were frequency matched on sex, ethnicity, date of birth, sample collection date and DNA type.

The *Newfoundland case-control study* (*Newfoundland*) [33,34]includes pathology confirmed CRC cases less than 75 years of age diagnosed between January 1999 and December 2003, as identified from the Newfoundland Colorectal Cancer Registry, which is modeled after CCFR. The Newfoundland Cancer Registry registers all cases of invasive cancer diagnosed among residents of the province of Newfoundland and Labrador. Consenting patients received a family history questionnaire and were asked to provide a blood sample and to permit access to tumor tissue and medical records. If a patient was deceased, we sought the participation of a close relative for the purposes of obtaining the family history and for permission to access tissue blocks and medical records. Use of proxies in this way

removes the bias of excluding advanced stage patients who die before they can give consent. Controls were identified by random digit dialing from the residents of the province, and matched to the cases on sex and five-year age groups. Controls provided a blood sample and filled out a risk factor questionnaire.

The *Melbourne case-control study* (*Melbourne*)[32] is nested in the Melbourne Collaborative Cohort Study, a prospective cohort study of 41,514 people (17,045 men, 24,469 women) ages 27 to 80 years at baseline. Recruitment occurred between 1990 and 1994. Southern European migrants to Australia were oversampled to extend the range of lifestyle exposures and to increase genetic variation. Subjects were recruited via the Electoral Rolls (registration to vote is compulsory for adults in Australia), advertisements, and community announcements in local media. Extensive information was collected at baseline in face-to-face interviews that included questionnaires (diet, physical activity etc.) and physical measurements, including lean and fat mass by bioelectric impedance, and blood pressure. Passive follow-up has been conducted by record linkage to Electoral Rolls, electronic phonebooks, and the Victorian Cancer Registry and death records. Cases were identified from notifications to the Victorian Cancer Registry and to the National Cancer Statistics Clearing House, of diagnoses of invasive or metastatic colorectal adenocarcinoma. Controls were randomly selected among the cancer-free cohort members and frequency-matched on sex, country of birth (Australia/UK, Italy and Greece), and year of baseline attendance.

Genotype data were cleaned based on quality control (QC) metrics at the individual subject and SNP levels. Samples with <95% call rate, sex mismatches (between self-reported and genotypic predicted sex), low concordance with previous genotype data, duplicate samples, unanticipated genotype concordance, identity-by-descent (IBD) with another sample, or ethnic outliers as identified by visual inspection of PCA cluster plots were removed. Prior to imputation, SNPs with <95% call rate, concordance <95% with 1000 Genomes in samples genotyped for quality control, or HWE $p<10^{-4}$ in controls were excluded. All SNPs overlapping 1000 Genomes were matched to the forward strand.

To analyze genotype data generated from three different platforms that measure different genetic markers and to increase the coverage of variation that is measurable across the genome, imputation of genotypes was performed for both autosomal and X chromosome markers. IMPUTE2 was used to impute missing genotypes for study samples based on the cosmopolitan panel of reference haplotypes from Phase I of the 1000 Genomes Project (March 2012 release; n = 1092). In order to enter subsequent statistical analysis steps, genetic markers resulting from the imputation had to pass stringent imputation quality and accuracy filters (info>=0.7, certainty>=0.9, concordance>=0.9 between directly measured and imputed genotypes after masking input genotypes (for genotyped markers only). Version differences in Illumina 1M and Omni1 platforms led us to exclude SNPs where allele frequency differences in CCFR cases were identified ($p<10^{-6}$). Further, we restricted our SNP list to those with study-specific MAF ≥1%.

Each contributing dataset was first analyzed in a study-specific fashion, allowing for adjustment for appropriate covariates, including age, sex, study center, genotyping batch, and 2-4 principal components. Then, study-specific results were analyzed using an inverse-variance-weighted, fixed-effects meta-analysis which assumed homogeneity of effects across all studies using METAL software. To examine the association between each variant and CRC risk, we specified a log-additive genetic model, where each additional copy of the minor allele was assumed to confer the same magnitude of risk or protection. Each SNP was coded as a dosage, the expected number of effect alleles.

Quantile-quantile (Q-Q) plots were generated for each study as well as the overall meta-analysis to examine the distribution of p-values compared to the distribution under null expectations. The genomic control lambda (GC λ) associated with the observed p-value distribution was examined for each study and the summary meta-analysis, with little evidence of unadjusted population stratification.

Statistical analysis and plotting were conducted using a combination of PLINK v1.07 [35], R v2.15.2, and METAL [36].

## GWAS IN GECCO

The GECCO GWAS consisted of European-descent participants within the French Association Study Evaluating RISK for sporadic colorectal cancer (ASTERISK); Hawaiian Colorectal Cancer Studies 2 and 3 (CR2&3); Darmkrebs: Chancen der Verhütung durch Screening (DACHS); Diet, Activity, and Lifestyle Study (DALS); Health Professionals Follow-up Study (HPFS); Multiethnic Cohort (MEC); Nurses' Health Study (NHS); Ontario Familial Colorectal Cancer Registry (OFCCR); Physician's Health Study (PHS); Postmenopausal Hormone study (PMH); Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial (PLCO); VITamins And Lifestyle (VITAL); and the Women's Health Initiative (WHI). Phase one genotyping on samples from PLCO, WHI, and DALS (PLCO Set 1, WHI Set 1, and DALS Set 1) was done using Illumina HumanHap 550K, 610K, or combined Illumina 300K and 240K, and has been described previously[37]. Samples from OFCCR are included in GECCO from previous genotyping using Affymetrix platforms[7]. Samples from ASTERISK, CR2&3, DACHS Set 1, DALS Set 2, MEC, PMH, PLCO Set 2, VITAL, and WHI Set 2 were genotyped using Illumina HumanCytoSNP. Samples from HPFS, NHS, PHS, DACHS Set 2 were successfully genotyped using Illumina HumanOmniExpress.

DNA was extracted from blood samples or, for a subset of DACHS, HPFS, MEC, NHS, and PLCO samples, and for all VITAL samples, from buccal cells, using conventional methods. All studies included 1 to 6% blinded duplicates to monitor quality of the genotyping. All individual-level genotype data were managed and underwent quality assurance and quality control (QA/QC) at the Ontario Institute for Cancer Research (OFCCR), the University of Washington (HPFS, NHS, PHS, and DACHS Set 2), or at the Fred Hutchinson Cancer Research Center (all other studies). Details on the QA/QC have been described before[2]. In brief, samples were excluded based on call rate, heterozygosity, unexpected duplicates, gender discrepancy, and unexpectedly high identity-by-descent or unexpected genotype concordance (> 65%) with another individual. All analyses were restricted to samples clustering with the Utah residents with Northern and Western European ancestry from the CEPH collection (CEU) population in principal component analysis, including the HapMap II populations as reference. SNPs were excluded if they were triallelic, not assigned an rs number, or were reported or observed as not performing consistently across platforms. Additionally, genotyped SNPs were excluded based on call rate (< 98%), lack of Hardy Weinberg Equilibrium in controls (HWE, $p < 1 \times 10^{-4}$), and minor allele frequency (MAF < 5% in Set 1 for PLCO, WHI, DALS, and OFCCR; minor allele count < 10 for remaining studies).

The genotype data were imputed to increase the density of genetic variants. The haplotypes of 1,092 samples (all populations) from release version 2 of the 1000 Genomes Project Phase I were used as the reference panel. Combining reference data from all populations helps improve imputation accuracy of low-frequency variants[38]. The target panel comprised genome-wide genotype data obtained using the

methods described above. The target panel was phased using Beagle[39], and the phased target panel was imputed to the 1000 Genomes reference panel using Minimac[40]. Rsq was used as the imputation quality measure for imputed SNPs[41]. Imputed SNPs were excluded such that variants with low MAFs required higher imputation quality: For SNPs with MAF>0.01, those with Rsq≤0.3 were excluded; for MAFs of 0.005-0.01, Rsq<0.5 were excluded; and for MAF<0.005, Rsq<0.99 were excluded.

For each study, the association between SNPs and risk for colorectal cancer were estimated by calculating betas, odds ratios (ORs), standard errors, 95% confidence intervals (CIs), and p-values using logistic regression models with log-additive genetic effects. Each directly genotyped SNP was coded as 0, 1 or 2 copies of the risk allele. For imputed SNPs, the expected number of copies of the risk allele (the "dosage") was used, which has been shown to give unbiased estimates in the association test for imputed SNPs[42]. Adjusted covariates include age, sex (when appropriate), center (when appropriate), smoking status (PHS only), batch effects (ASTERISK only), and the first three principal components from EIGENSTRAT to account for population substructure. Quantile-quantile (Q-Q) plots were assessed to determine whether the distribution of the p-values in each study population was consistent with the null distribution (except for the extreme tail). The genomic inflation factor ($\lambda$) was computed to measure the over-dispersion of the test-statistics from the association tests by dividing the median of the squared Z statistics by 0.455, the median of a chi-squared distribution with 1 degree of freedom. The inflation factor $\lambda$ was between 0.997 and 1.040 for individual studies, indicating there is little evidence of residual population substructure, cryptic relatedness, or differential genotyping between cases and controls. This result was consistent with the visual inspection of the study-specific Q-Q plots.

Inverse-variance weighted, fixed-effects meta-analysis was conducted to combine beta estimates and standard errors across individual studies. In this approach, the beta estimate of each study was weighted by its inverse variance, and a combined estimate was calculated by summing the weighted betas and dividing by the summed weights. For imputed SNPs, it has been shown that the inverse variance is approximately proportional to the imputation quality[42]. Thus, the inverse variance weighting scheme automatically incorporates imputation quality in the meta-analysis for imputed SNPs. The heterogeneity p-values were calculated based on Cochran's Q statistic [43] and investigated sources for heterogeneity if $p < 0.05$ for the ten most significant SNPs. PLINK and R were used to conduct the statistical analysis.

## Description of Study Populations in the GECCO GWAS

*Ontario Familial Colorectal Cancer Registry (OFCCR).* In GECCO, a subset of the Assessment of Risk in Colorectal Tumours in Canada (ARCTIC) from the Ontario Registry for Studies of Familial Colorectal Cancer (OFCCR) was used. Both the case-control study and the OFCCR[44,45] have been described in detail previously, as have GWAS results[7]. In brief, cases were confirmed incident colorectal cancer (CRC) cases ages 20 to 74 years, residents of Ontario identified through comprehensive registry and diagnosed between July 1997 and June 2000. Population-based controls were randomly selected among Ontario residents (random-digit-dialing and listing of all Ontario residents), and matched by sex and 5-year age groups. A total of 1,236 CRC cases and 1,223 controls were successfully genotyped on at least one of the Illumina 1536 GoldenGate assay, the Affymetrix GeneChip® Human Mapping 100K and 500K Array Set, and a 10K non-synonymous SNP chip. Analysis was based on a set of unrelated subjects who were non-Hispanic, White by self-report or by investigation of genetic ancestry. Subjects were further excluded if there was a sample mix-up, if they

were missing epidemiologic questionnaire data, if they were appendix cases, or if they were overlapped with the Colon Cancer Family Registry. Additionally, only samples genotyped on the Affymetrix GeneChip® 500K Array were utilized in order to avoid coverage issues in imputation.

*French Association Study Evaluating RISK for sporadic colorectal cancer (ASTERISK)*[46]. Participants were recruited from the Pays de la Loire region in France between December 2002 and March 2006. Eligibility criteria for cases included being of Caucasian origin, being greater than or 40 years of age at diagnosis, and having no family history of colorectal cancer or polyps. Cases were patients with first primary colorectal cancer diagnosed in one of the six public hospitals and five clinics located in the Pays de la Loire region which participated in the study. Cases were confirmed based on medical and pathology reports. Controls were recruited at two Health Examination Centers of the Pays de la Loire region, and the recruitment of controls greater than or 70 years was completed in the departments of internal medicine and hepatogastroenterology of the University Hospital Center of Nantes, located in the same region. Controls were eligible to participate if they were Caucasian, aged greater than or 40 years, and had no family history of colorectal cancer or polyps. In the presence of the physician, each participant filled out a standardized questionnaire on family information, medical history, lifestyle, and dietary intake. Cases and controls provided a blood sample.

*Darmkrebs: Chancen der Verhütung durch Screening (DACHS)*[47,48]. This German study was initiated as a large population-based case-control study in 2003 in the Rhine-Neckar-Odenwald region (southwest region of Germany) to assess the potential of endoscopic screening for reduction of colorectal cancer risk and to investigate etiologic determinants of disease, particularly lifestyle/environmental factors and genetic factors. Cases with a first diagnosis of invasive colorectal cancer (ICO-10 codes C18-C20) who were at least 30 years of age (no upper age limit), German speaking, a resident in the study region, and mentally and physically able to participate in a one-hour interview, were recruited by their treating physicians either in the hospital a few days after surgery, or by mail after discharge from the hospital. Cases were confirmed based on histologic reports and hospital discharge letters following diagnosis of colorectal cancer. All hospitals treating colorectal cancer patients in the study region participated. Based on estimates from population-based cancer registries, more than 50% of all potentially eligible patients with incident colorectal cancer in the study region were included. Community-based controls were randomly selected from population registries, employing frequency matching with respect to age (5-year groups), sex, and county of residence. Controls with a history of colorectal cancer were excluded. Controls were contacted by mail and follow-up calls. The participation rate was 51%. During an in-person interview, data were collected on demographics, medical history, family history of CRC, and various life-style factors, as were blood and mouthwash samples. The Set 1 scan consisted of a subset of participants recruited up to 2007, and samples were frequency matched on age and gender. The Set 2 scan consisted of additional subjects that were recruited up to 2010 as part of this ongoing study.

*Diet, Activity, and Lifestyle Study (DALS)*[49]. DALS is a population-based case-control study of colon cancer. Participants were recruited between 1991 and 1994 from three locations: the Kaiser Permanente Medical Care Program (KPMCP) of Northern California, an eight-county area in Utah, and the metropolitan Twin Cities area of Minnesota. Eligibility criteria for cases included age at diagnosis between 30 and 79 years, diagnosis with first primary colon cancer (ICD-O-2 codes 18.0 and 18.2-18.9) between October 1$^{st}$ 1991 and September 30$^{th}$ 1994, English speaking, and competency to complete the interview. Individuals with cancer of the rectosigmoid junction or rectum were excluded, as were those with a pathology report noting familial adenomatous polyposis, Crohn's disease, or ulcerative colitis. A rapid-reporting system was used to identify all incident cases of colon cancer resulting in the majority of

cases being interviewed within four months of diagnosis. Controls from KPMCP were randomly selected from membership lists. In Utah, controls under 65 years of age were randomly selected through random-digit dialing and driver license lists. Controls, 65 years of age and older, were randomly selected from Health Care Financing Administration lists. In Minnesota, controls were identified from Minnesota driver's license or state ID lists. Cases and controls were matched to cases by 5-year age groups and sex. The Set I scan consisted of a subset of the study designed above, from Utah, Minnesota, and KPMCP, and was restricted to subjects who self-reported as White non-Hispanic. The Set 2 scan consisted of subjects from Utah and Minnesota that were not genotyped in Set 1. Set 2 was restricted to subjects who self-reported as White non-Hispanic and those that had appropriate consent to post data to dbGaP.

*Hawai'i Colorectal Cancer Studies 2 & 3 (CR2&3)* [19]. Patients with colorectal cancer were identified through the rapid reporting system of the Hawaii SEER registry and consisted of all Japanese, Caucasian, and Native Hawaiian residents of Oahu who were newly diagnosed with an adenocarcinoma of the colon or rectum between January 1994 and August 1998. Control subjects were selected from participants in an on-going population-based health survey conducted by the Hawaii State Department of Health and from Health Care Financing Administration participants. Controls were matched to cases by sex, ethnicity, and age (within two years). Personal interviews were obtained from 768 matched pairs, resulting in a participation rate of 58.2% for cases and 53.2% for controls. A questionnaire, administered during an in-person interview, included questions about demographics, lifetime history of tobacco, alcohol use, aspirin use, physical activity, personal medical history, family history of colorectal cancer, height and weight, diet (FFQ), and postmenopausal hormone use. A blood sample was obtained from 548 (71%) of interviewed cases and 662 (86%) of interviewed controls. SEER staging information was extracted from the Hawaii Tumor Registry. In GECCO, self-reported Caucasian subjects with DNA, and clinical and epidemiologic data were selected for genotyping.

*Health Professionals Follow-up Study (HPFS)*[50]. The HPFS is a parallel prospective study to the Nurses' Health Study (NHS). The HPFS cohort comprises 51,529 men who, in 1986, responded to a mailed questionnaire. The participants are U.S. male dentists, optometrists, osteopaths, podiatrists, pharmacists, and veterinarians born between 1910 and 1946. Participants have provided information on health related exposures, including: current and past smoking history, age, weight, height, diet, physical activity, aspirin use, and family history of colorectal cancer. Colorectal cancer and other outcomes were reported by participants or next-of-kin and followed up through review of the medical and pathology record by physicians. Overall, more than 97% of self-reported colorectal cancers were confirmed by medical record review. Information was abstracted on histology and primary location. Incident cases are defined as those occurring after the subject provided the blood sample. Prevalent cases are defined as those occurring after enrollment in the study, but prior to the subject providing the blood sample. Follow-up has been excellent, with 94% of the men responding to date. Colorectal cancer cases were ascertained through January 1, 2008. In 1993-95, 18,825 men in HPFS mailed in blood samples by overnight courier which were aliquoted into buffy coat and stored in liquid nitrogen. In 2001-04, 13,956 men in HPFS who had not previously provided a blood sample mailed in a "swish-and-spit" sample of buccal cells. Incident cases are defined as those occurring after the subject provided a blood or buccal sample. Prevalent cases are defined as those occurring after enrollment in the study in 1986, but prior to the subject providing either a blood or buccal sample. After excluding participants with histories of cancer (except non-melanoma skin), ulcerative colitis, or familial polyposis, two case-control sets were constructed from which DNA was isolated from either buffy coat or buccal cells for genotyping: 1) a case-control set with cases of colorectal cancer matched to randomly selected controls who provided a blood sample and were free of colorectal cancer at the same time the colorectal cancer was diagnosed in

the cases; 2) a case-control set with cases of colorectal cancer matched to randomly selected controls who provided a buccal sample and were free of colorectal cancer at the same time the colorectal cancer was diagnosed in the case. For both case-control sets, matching criteria included year of birth (within 1 year) and month/year of blood or buccal cell sampling (within six months). Cases were pair matched 1:1, 1:2, or 1:3 with a control participant(s).

*Multiethnic Cohort Study (MEC)*. MEC[15] has been described above. In GECCO, self-reported white subjects from the nested case-control study described above with DNA, and clinical and epidemiologic data were selected for genotyping

*Nurses' Health Study (NHS)*. The NHS cohort [51]began in 1976 when 121,700 married female registered nurses aged 30 to 55 years returned the initial questionnaire that ascertained a variety of important health-related exposures. Since 1976, follow-up questionnaires have been mailed every two years. Colorectal cancer and other outcomes were reported by participants or next-of-kin and followed up through review of the medical and pathology record by physicians. Overall, more than 97% of self-reported colorectal cancers were confirmed by medical-record review. Information was abstracted on histology and primary location. Follow-up has been high: as a proportion of the total possible follow-up time, follow-up has been over 92%. Colorectal cancer cases were ascertained through June 1, 2008. In 1989-90, 32,826 women in NHS I, mailed in blood samples by overnight courier which were aliquoted into buffy coat and stored in liquid nitrogen. In 2001-04, 29,684 women in NHS I who did not previously provide a blood sample mailed in a "swish-and-spit" sample of buccal cells. Incident cases are defined as those occurring after the subject provided a blood or buccal sample. Prevalent cases are defined as those occurring after enrollment in the study in 1976, but prior to the subject providing either a blood or buccal sample. After excluding participants with histories of cancer (except non-melanoma skin), ulcerative colitis, or familial polyposis, two case-control sets were constructed from which DNA was isolated from either buffy coat or buccal cells for genotyping: 1) a case-control set with cases of colorectal cancer matched to randomly selected controls who provided a blood sample and were free of colorectal cancer at the same time the colorectal cancer was diagnosed in the case; 2) a case-control set with cases of colorectal cancer matched to randomly selected controls who provided a buccal sample and were free of colorectal cancer at the same time the colorectal cancer was diagnosed in the cases. For both case-control sets, matching criteria included year of birth (within one year) and month / year of blood or buccal cell sampling (within six months). Cases were pair matched 1:1, 1:2, or 1:3 with a control participant(s).

*Physician's Health Study (PHS)*. The PHS [52,53]was established as a randomized, double-blind, placebo-controlled trial of aspirin and ß-carotene among 22,071 healthy U.S. male physicians, between 40 and 84 years of age in 1982. Participants completed two mailed questionnaires before being randomly assigned, additional questionnaires at six and 12 months, and questionnaires annually thereafter. In addition, participants were sent postcards at six months to ascertain status. From August 1982 to December 1984, 14,916 baseline blood samples were collected from the physicians during the run-in phase before randomization. When participants report a diagnosis of cancer, medical records and pathology reports are reviewed by study physicians who are blinded to exposure data. Among those who provided baseline blood samples, colorectal cases were ascertained through March 31, 2008, and controls were matched on age (within one year for younger participants, up to five years for older participants) and smoking status (never, past, current). Cases were "pair" matched 1:1, 1:2 or 1:3 with a control participant(s). Due to DNA availability samples were genotyped in two batches on the same platform at the same genotyping center at different time points.

*Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial (PLCO)*. PLCO [54,55]has been described above and details are available online (http://dcp.cancer.gov/plco).

The Set 1 scan included a subset of 577 colon cancer cases self-reported as being non-Hispanic White with available DNA samples, questionnaire data, and appropriate consent for ancillary epidemiologic studies. Cases were excluded if they had a history of inflammatory bowel disease, polyps, polyposis syndrome or cancer (excluding basal or squamous cell skin cancer). Controls come from the Cancer Genetic Markers of Susceptibility (CGEMS) prostate cancer scan[56] (all male) and the GWAS of Lung Cancer and Smoking [57](enriched for smokers) along with an additional 92 non-Hispanic White female controls. For the Set 2 scan, cases were colorectal cancers from both arms of the trial, which were not already included in Set 1. Samples were excluded if participants did not sign appropriate consents, if DNA was unavailable, if baseline questionnaire data with follow-up were unavailable, if they had a history of colon cancer prior to the trial, if they were a rare cancer, and if they were already in colon GWAS, or if they were a control in the prostate or lung populations. Controls were frequency matched 1:1 to cases without replacement, and cases were not eligible to be controls. Matching criteria were age at enrollment (two year blocks), enrollment date (two year blocks), sex, race / ethnicity, trial arm, and study year of diagnosis (i.e. controls must be cancer free into the case's year of diagnosis).

*Postmenopausal Hormones Supplementary Study to the Colon Cancer Family Registry (PMH-CCFR*)[58]. Eligible case patients included all female residents, ages 50 to 74 years, residing in the 13 counties in Washington State reporting to the Cancer Surveillance SEER program, who were newly diagnosed with invasive colorectal adenocarcinoma (ICD-O C18.0, C18.2-.9, C19.9, C20.0-.9) between October 1998 and February 2002. Eligibility for all individuals was limited to those who were English-speaking with available telephone numbers, in which they could be contacted. On average, cases were identified within four months of diagnosis. The overall response proportion of eligible cases identified was 73%. Community-based controls were randomly selected according to age distribution (in 5-year age intervals) of the eligible cases by using lists of licensed drivers from the Washington State Department of Licensing for individuals, ages 50 to 64 years, and rosters from the Health Care Financing Administration (now the Centers for Medicare and Medicaid) for individuals older than 64 years. The overall response proportion of eligible controls was 66%. In GECCO, samples with sufficient DNA extracted from blood were genotyped. Only participants that were not part of the CCFR Seattle site were included in the sample set.

*VITamins And Lifestyle (VITAL).* The VITamins And Lifestyle (VITAL) cohort comprises 77,721 Washington State men and women aged 50 to 76 years, recruited from 2000 to 2002 to investigate the association of supplement use and lifestyle factors with cancer risk. Subjects were recruited by mail, from October 2000 to December 2002, using names purchased from a commercial mailing list. All subjects completed a 24 page questionnaire and buccal-cell specimens for DNA was self-collected by 70% of the participants. Subjects are followed for cancer by linkage to the western Washington SEER cancer registry and are censored when they move out of the area covered by the registry or at time of death. Details of this study have been previously described[59]. In GECCO, a nested case-control set was genotyped. Samples included colorectal cancer cases with DNA, excluding subject with colorectal cancer before baseline, in situ cases, (large cell) neuroendocrine carcinoma, squamous cell carcinoma, carcinoid tumor, Goblet cell carcinoid, any type of lymphoma, including non-Hodgkin, Mantle cell, large B-cell, or follicular lymphoma. Controls were matched on age at enrollment (within one year), enrollment date (within one year), sex, and race / ethnicity. One control was randomly selected per case

among all controls that matched on the four factors above and where the control follow-up time was greater than follow-up time of the case until diagnosis.

*Women's Health Initiative (WHI).* WHI is a long-term health study of 161,808 post-menopausal women aged 50 to 79 years at 40 clinical centers throughout the U.S. WHI comprises a Clinical Trial (CT) arm, an Observational Study (OS) arm, and several extension studies. The details of WHI have been previously described[60,61] and are available online (https://cleo.whi.org/SitePages/Home.aspx). In GECCO, Set 1 cases were selected from the September 12, 2005 database and were comprised of centrally adjudicated colon cancer cases from the Observational Study (OS) who self-reported as White. Controls were first selected among controls previously genotyped as part of a Hip Fracture GWAS conducted within the WHI OS and matched to cases on age (within three years), enrollment date (within 365 days), hysterectomy status, and prevalent conditions at baseline. For 37 cases, there was not a control match in the Hip Fracture GWAS. For these participants, a matched control was identified in the WHI OS based on same criteria. In the Set 2 scan, cases were selected from the August 2009 database and were comprised of centrally adjudicated colon and colorectal cancer cases from the OS and CT who were not genotyped in Set 1. In addition, case and control participants were subject to the following exclusion criteria: a prior history of colorectal cancer at baseline, IRB approval not available for data submission into dbGaP, and not sufficient DNA available. Matching criteria included age (within years), race/ethnicity, WHI date (within three years), WHI Calcium and Vitamin D study date (within three years), and randomization arms (OS flag, hormone therapy assignments, dietary modification assignments, calcium/vitamin D assignments). In addition, they were matched on the four regions of randomization centers. Each case was matched with one control (1:1) that exactly met the matching criteria. Control selection was done in a time-forward manner, selecting one control for each case first from the risk set at the time of the case's event. The matching algorithm was allowed to select the closest match based on a criterion to minimize an overall distance measure. Each matching factor was given the same weight. Additional available controls that were genotyped as part of the Hip Fracture GWAS were included to improve power.

## Supplementary References


1. Houlston, R.S. *et al.* Meta-analysis of three genome-wide association studies identifies susceptibility loci for colorectal cancer at 1q41, 3q26.2, 12q13.13 and 20q13.33. *Nat Genet* **42**, 973-977 (2010).
2. Peters, U. *et al.* Identification of Genetic Susceptibility Loci for Colorectal Tumors in a Genome-Wide Meta-analysis. *Gastroenterology* **144**, 799-807 e724 (2013).
3. Jia, W.-H. *et al.* Genome-wide association analyses in east Asians identify new susceptibility loci for colorectal cancer. *Nat Genet* **45**, 191-196 (2013).
4. Dunlop, M.G. *et al.* Common variation near CDKN1A, POLD3 and SHROOM2 influences colorectal cancer risk. *Nat Genet* **44**, 770-776 (2012).
5. Cui, R. *et al.* Common variant in 6q26-q27 is associated with distal colon cancer in an Asian population. *Gut* **60**, 799-805 (2011).
6. Tomlinson, I.P. *et al.* A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. *Nat Genet* **40**, 623-630 (2008).
7. Zanke, B.W. *et al.* Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. *Nat Genet* **39**, 989-994 (2007).
8. Tomlinson, I. *et al.* A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nat Genet* **39**, 984-988 (2007).
9. Tenesa, A. *et al.* Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. *Nat Genet* **40**, 631-637 (2008).
10. Houlston, R.S. *et al.* Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. *Nat Genet* **40**, 1426-1435 (2008).
11. Tomlinson, I.P. *et al.* Multiple common susceptibility variants near BMP pathway loci GREM1, BMP4, and BMP2 explain part of the missing heritability of colorectal cancer. *PLoS Genet* **7**, e1002105 (2011).
12. Jaeger, E. *et al.* Common genetic variants at the CRAC1 (HMPS) locus on chromosome 15q13.3 influence colorectal cancer risk. *Nat Genet* **40**, 26-28 (2008).
13. Broderick, P. *et al.* A genome-wide association study shows that common alleles of SMAD7 influence colorectal cancer risk. *Nat Genet* **39**, 1315-1317 (2007).
14. Ward, L.D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res* **40**, D930-934 (2012).
15. Kolonel, L.N. *et al.* A multiethnic cohort in Hawaii and Los Angeles: baseline characteristics. *Am J Epidemiol* **151**, 346-357 (2000).
16. Cheng, I. *et al.* Evaluating genetic risk for prostate cancer among Japanese and Latinos. *Cancer Epidemiol Biomarkers Prev* **21**, 2048-2058 (2012).
17. Siddiq, A. *et al.* A meta-analysis of genome-wide association studies of breast cancer identifies two novel susceptibility loci at 6q14 and 20q11. *Hum Mol Genet* **21**, 5373-5384 (2012).
18. Newcomb, P.A. *et al.* Colon Cancer Family Registry: an international resource for studies of the genetic epidemiology of colon cancer. *Cancer Epidemiol Biomarkers Prev* **16**, 2331-2343 (2007).
19. Le Marchand, L. *et al.* Combined effects of well-done red meat, smoking, and rapid N-acetyltransferase 2 and CYP1A2 phenotypes in increasing colorectalcancer risk. *Cancer Epidemiol Biomarkers Prev* **10**, 1259–1266 (2001).

20. Kono, S., Toyomura, K., Yin, G., Nagano, J. & Mizoue, T. A case-control study of colorectal cancer in relation to lifestyle factors and genetic polymorphisms: design and conduct of the Fukuoka Colorectal Cancer Study. *Asian Pac J Cancer Prev* **5**, 393-400 (2004).
21. Yin, G. *et al.* Methylenetetrahydrofolate reductase C677T and A1298C polymorphisms and colorectal cancer: The Fukuoka Colorectal Cancer Study. . *Cancer Sci* **95**, 908-913 (2004).
22. Otani, T. *et al.* Folate, vitamin B6, vitamin B12, and vitamin B2 intake, genetic polymorphisms of related enzymes, and risk of colorectal cancer in a hospital-based case-control study in Japan. *Nutr Cancer* **53**, 42-50 (2005).
23. Otani, T. *et al.* Plasma C-reactive protein and risk of colorectal cancer in a nested case-control study: Japan Public Health Center-based prospective study. *Cancer Epidemiol Biomarkers Prev* **15**, 690-695 (2006).
24. Tsugane, S. & Sobue, T. Baseline survey of JPHC study--design and participation rate. Japan Public Health Center-based Prospective Study on Cancer and Cardiovascular Diseases. *J Epidemiol* **11**, S24-29 (2001).
25. Haiman, C.A. *et al.* Genome-wide association study of prostate cancer in men of African ancestry identifies a susceptibility locus at 17q21. *Nat Genet* **43**, 570-573 (2011).
26. Chen, F. *et al.* A genome-wide association study of breast cancer in women of African ancestry. *Hum Genet* **132**, 39-48 (2013).
27. Signorello, L.B. *et al.* Southern community cohort study: establishing a cohort to investigate health disparities. *J Natl Med Assoc* **97**, 972–979 (2005).
28. Pande, M., Amos, C.I., Eng, C. & Frazier, M.L. Interactions between cigarette smoking and selected polymorphisms in xenobiotic metabolizing enzymes in risk for colorectal cancer: A case-only analysis. *Molecular Carcinogenesis* **49**, 974-980 (2010).
29. Gruber, S.B. *et al.* Genetic variation in 8q24 associated with risk of colorectal cancer. *Cancer Biol Ther* **6**, 1143-1147 (2007).
30. Li, L. *et al.* A common 8q24 variant and the risk of colon cancer: a population-based case-control study. *Cancer Epidemiol Biomarkers Prev* **17**, 339-342 (2008).
31. Calle, E.E. *et al.* The American Cancer Society Cancer Prevention Study II Nutrition Cohort: rationale, study design, and baseline characteristics. *Cancer* **94**, 2490-2501 (2002).
32. Bassett, J.K. *et al.* Body size, weight change, and risk of colon cancer. *Cancer Epidemiol Biomarkers Prev* **19**, 2978-2986 (2010).
33. Green, R.C. *et al.* Very high incidence of familial colorectal cancer in Newfoundland: a comparison with Ontario and 13 other population-based studies. *Fam Cancer* **6**, 53-62 (2007).
34. Woods, M.O. *et al.* The genetic basis of colorectal cancer in a population-based incident cohort with a high rate of familial disease. *Gut* **59**, 1369-1377 (2010).
35. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-575 (2007).
36. Willer, C.J., Li, Y. & Abecasis, G.R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
37. Peters, U. *et al.* Meta-analysis of new genome-wide association studies of colorectal cancer risk. *Hum Genet* **131**, 217-234 (2012).
38. Howie, B., Marchini, J. & Stephens, M. Genotype imputation with thousands of genomes. *G3 (Bethesda)* **1**, 457-470 (2011).
39. Browning, S.R. & Browning, B.L. Rapid and accurate haplotype phasing and missing data inference for whole genome association studies using localized haplotype clustering. *Am J Hum Genet* **81**, 1084-1097 (2007).

40. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G.R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet* **44**, 955-959 (2012).
41. Li, Y., Willer, C.J., Ding, J., Scheet, P. & Abecasis, G.R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol* **34**, 816-834 (2010).
42. Jiao, S., Hsu, L., Hutter, C.M. & Peters, U. The use of imputed values in the meta-analysis of genome-wide association studies. *Genet Epidemiol* **35**, 597-605 (2011).
43. Ioannidis, J.P., Patsopoulos, N.A. & Evangelou, E. Heterogeneity in meta-analyses of genome-wide association investigations. *PLoS ONE* **2**, e841 (2007).
44. Cotterchio, M., Manno, M., Klar, N., McLaughlin, J. & Gallinger, S. Colorectal screening is associated with reduced colorectal cancer risk: a case-control study within the population-based Ontario Familial Colorectal Cancer Registry. *Cancer Causes Control* **16**, 865-875 (2005).
45. Cotterchio, M. *et al.* Ontario familial colon cancer registry: methods and first-year response rates. *Chronic Dis Can* **21**, 81-86 (2000).
46. Kury, S. *et al.* Combinations of cytochrome P450 gene polymorphisms enhancing the risk for sporadic colorectal cancer related to red meat consumption. *Cancer Epidemiol Biomarkers Prev* **16**, 1460-1467 (2007).
47. Brenner, H., Chang-Claude, J., Seiler, C.M., Rickert, A. & Hoffmeister, M. Protection from colorectal cancer after colonoscopy: a population-based, case-control study. *Ann Intern Med* **154**, 22-30 (2011).
48. Lilla, C. *et al.* Effect of NAT1 and NAT2 genetic polymorphisms on colorectal cancer risk associated with exposure to tobacco smoke and meat consumption. *Cancer Epidemiol Biomarkers Prev* **15**, 99-107 (2006).
49. Slattery, M.L. *et al.* Energy balance and colon cancer--beyond physical activity. *Cancer Res* **57**, 75-80 (1997).
50. Rimm, E.B. *et al.* Validity of self-reported waist and hip circumferences in men and women. *Epidemiology* **1**, 466-473 (1990).
51. Belanger, C.F., Hennekens, C.H., Rosner, B. & Speizer, F.E. The nurses' health study. *Am J Nurs* **78**, 1039-1040 (1978).
52. Hennekens, C.H. & Eberlein, K. A randomized trial of aspirin and beta-carotene among U.S. physicians. *Prev Med* **14**, 165-168 (1985).
53. Christen, W.G., Gaziano, J.M. & Hennekens, C.H. Design of Physicians' Health Study II--a randomized trial of beta-carotene, vitamins E and C, and multivitamins, in prevention of cancer, cardiovascular disease, and eye disease, and review of results of completed trials. *Ann Epidemiol* **10**, 125-134 (2000).
54. Prorok, P.C. *et al.* Design of the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial. *Control Clin Trials* **21**, 273S-309S (2000).
55. Gohagan, J.K. *et al.* The Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial of the National Cancer Institute: history, organization, and status. *Control Clin Trials* **21**, 251S-272S (2000).
56. Yeager, M. *et al.* Identification of a new prostate cancer susceptibility locus on chromosome 8q24. *Nat Genet* **41**, 1055-1057 (2009).
57. Landi, M.T. *et al.* A genome-wide association study of lung cancer identifies a region of chromosome 5p15 associated with risk for adenocarcinoma. *Am J Hum Genet* **85**, 679-691 (2009).

58. Newcomb, P.A. *et al.* Estrogen plus progestin use, microsatellite instability, and the risk of colorectal cancer in women. *Cancer Res* **67**, 7534-7539 (2007).
59. White, E. *et al.* VITamins And Lifestyle cohort study: study design and characteristics of supplement users. *Am J Epidemiol* **159**, 83-93 (2004).
60. Hays, J. *et al.* The Women's Health Initiative recruitment methods and results. *Ann Epidemiol* **13**, S18-77 (2003).
61. Design of the Women's Health Initiative clinical trial and observational study. The Women's Health Initiative Study Group. *Control Clin Trials* **19**, 61-109 (1998).