**JNCI 17-1920R1**

**Article**

# Novel common genetic susceptibility loci for colorectal cancer

Stephanie L. Schmit[1,2*], Christopher K. Edlund[2*], Fredrick R. Schumacher[3*], Jian Gong[4*], Tabitha A. Harrison[4], Jeroen R. Huyghe[4], Chenxu Qu [2], David J. Van Den Berg[2], Hansong Wang[5], Stephanie Tring[2], Sarah J. Plummer[6], Demetrius Albanes[7], M. Henar Alonso[8-10], Christopher I. Amos[11], Kristen Anton[11], Aaron K. Aragaki[4], Volker Arndt[12], Elizabeth L. Barry[13], Sonja I. Berndt[7], Stéphane Bezieau[14,15], Stephanie Bien[4], Amanda Bloomer[1], Juergen Boehm[16], Marie-Christine Boutron-Ruault[17,18], Hermann Brenner[12,19,20], Stefanie Brezina[21], Daniel D. Buchanan[22-24], Katja Butterbach[12], Bette J. Caan[25], Peter T. Campbell[26], Christopher S. Carlson[4], Jose E. Castelao[27], Andrew T. Chan[28-30], Jenny Chang-Claude[31,32], Stephen J. Chanock[7], Iona Cheng[33], Ya-Wen Cheng[34], Lee Soo Chin[35], James M. Church[36], Timothy Church[37], Gerhard A. Coetzee[38], Michelle Cotterchio[39], Marcia Cruz Correa[40], Keith R. Curtis[4], David Duggan[41], Douglas F. Easton[42], Dallas English[22,43], Edith J.M. Feskens[44], Rocky Fischer[45], Liesel M. FitzGerald[43,46], Barbara K. Fortini[47], Lars G. Fritsche[48-50], Charles S. Fuchs[51,52], Manuela Gago-Dominguez[53,54], Manish Gala[28,30], Steven J. Gallinger[55], W. James Gauderman[2], Graham G. Giles[22,43], Edward L. Giovannucci[29,56], Stephanie M. Gogarten[57], Clicerio Gonzalez-Villalpando[58], Elena M. Gonzalez-Villalpando[59], William M. Grady[60], Joel K. Greenson[61], Andrea Gsur[21], Marc Gunter[62], Christopher A. Haiman[2], Jochen Hampe[63], Sophia Harlid[64], John F. Harju[45], Richard B. Hayes[65], Philipp Hofer[21], Michael Hoffmeister[12], John L. Hopper[66], Shu-Chen Huang[2], Jose Maria Huerta[9,67], Thomas J. Hudson[68,69], David J. Hunter[70], Gregory E. Idos[2],

Motoki Iwasaki[71,72], Rebecca D. Jackson[73], Eric J. Jacobs[26], Sun Ha Jee[74], Mark A. Jenkins[22], Wei-Hua Jia[75], Shuo Jiao[4], Amit D. Joshi[30,70], Laurence N. Kolonel[76], Suminori Kono[77], Charles Kooperberg[4], Vittorio Krogh[78], Tilman Kuehn[79], Sébastien Küry[15], Andrea LaCroix[4], Cecelia A. Laurie[57], Flavio Lejbkowicz[80,81], Mathieu Lemire[69], Heinz-Josef Lenz[2,82], David Levine[57], Christopher I. Li[83], Li Li[84], Wolfgang Lieb[85], Yi Lin[4], Noralane M. Lindor[86], Yun-Ru Liu[87], Fotios Loupakis[88], Yingchang Lu[89], Frank Luh[90,91], Jing Ma[92], Christoph Mancao[93], Frank J. Manion[45], Sanford D. Markowitz[94], Vicente Martin[9,95], Koichi Matsuda[96], Keitaro Matsuo[97,98], Kevin J. McDonnell[2], Caroline E. McNeil[2], Marilena Melas[2], Roger Milne[22,43], Antonio J. Molina[9,95], Bhramar Mukherjee[45], Neil Murphy[62], Polly A. Newcomb[4], Kenneth Offit[99], Hanane Omichessan[17,18], Domenico Palli[100], Jesus P. Paredes Cotoré[101], Julyann Pérez-Mayoral[102], Paul D. Pharoah[103], John D. Potter[4], Conghui Qu[4], Leon Raskin[89,104], Gad Rennert[80,81,105], Hedy S. Rennert[80,81], Bridget M. Riggs[1], Clemens Schafmayer[106], Robert E. Schoen[107], Thomas A. Sellers[1], Daniela Seminara[108], Gianluca Severi[17,109], Wei Shi[110], David Shibata[111], Xiao-Ou Shu [89,104], Erin M. Siegel[1], Martha L. Slattery[112], Melissa Southey[113], Zsofia K. Stadler[114,115], Mariana C. Stern[2], Sebastian Stintzing[116], Darin Taverna[117], Stephen N. Thibodeau[118], Duncan C. Thomas[2], Antonia Trichopoulou[119], Shoichiro Tsugane[71,120], Cornelia M. Ulrich[16], Franzel J.B. van Duijnhoven[44], Bethany van Guelpan[64], Joseph Vijai[114], Jarmo Virtamo[121], Stephanie J. Weinstein[7], Emily White[4], Aung Ko Win[22], Alicja Wolk[122], Michael Woods[123], Anna H. Wu[2], Kana Wu[124], Yong-Bing Xiang[125], Yun Yen[34,126], Brent W. Zanke[127,128], Yi-Xin Zeng[75], Ben Zhang[129], Niha Zubair[4], Sun-Seog Kweon[130,131], Jane C. Figueiredo[132,133], Wei Zheng[89,104], Loic Le Marchand[5], Annika Lindblom[134,135], Victor Moreno[8-10], Ulrike Peters[4,136†], Graham Casey[6†], Li Hsu[4†], David V. Conti[2†], Stephen B. Gruber[2,82†]

1. Department of Cancer Epidemiology, H. Lee Moffitt Cancer Center and Research

Institute, Tampa, Florida 33612, USA.

2. Department of Preventive Medicine, USC Norris Comprehensive Cancer Center, Keck School of Medicine, University of Southern California, Los Angeles, California 90033, USA.

3. Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, Ohio 44106, USA.

4. Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington 98109, USA.

5. Epidemiology Program, University of Hawaii Cancer Center, Honolulu, Hawaii 96822, USA.

6. Center for Public Health Genomics, University of Virginia, Charlottesville, Virginia 22908, USA.

7. Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland 20892, USA.

8. Catalan Institute of Oncology, Bellvitge Biomedical Research Institute (IDIBELL), Barcelona 08028, Spain.

9. CIBER Epidemiología y Salud Pública (CIBERESP), Madrid 28029, Spain.

10. University of Barcelona, Barcelona 08007, Spain.

11. Department of Biomedical Data Science, Geisel School of Medicine, Dartmouth College, Hanover, New Hampshire 03755, USA.

12. Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), Heidelberg 69120, Germany.

13. Department of Epidemiology, Geisel School of Medicine, Dartmouth College, Hanover,

New Hampshire 03755, USA.

14. Centre Hospitalier Universitaire Hotel-Dieu, Nantes 44093, France.

15. Service de Génétique Médicale, Centre Hospitalier Universitaire (CHU), Nantes 44093, France.

16. Huntsman Cancer Institute and Department of Population Health Sciences, University of Utah, Salt Lake City, Utah 84112, USA.

17. CESP (U1018 INSERM), Facultés de médecine Université Paris-Sud, UVSQ, Université Paris-Saclay, 94805, Villejuif, France.

18. Gustave Roussy, F-94805, Villejuif, France.

19. Division of Preventive Oncology, German Cancer Research Center (DKFZ) and National Center for Tumor Diseases (NCT), Heidelberg 69120, Germany.

20. German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg 69120, Germany.

21. Medical University of Vienna, Department of Medicine I, Institute of Cancer Research, Vienna A-1090, Austria.

22. Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Parkville, Victoria 3010, Australia.

23. Colorectal Oncogenomics Group, Department of Pathology, University of Melbourne, Melbourne, Victoria 3010, Australia.

24. Genetic Medicine and Familial Cancer Centre, The Royal Melbourne Hospital, Parkville, Victoria 3010, Australia.

25. Division of Research, Kaiser Permanente Medical Care Program of Northern California, Oakland, California 94612, USA.

26. Epidemiology Research Program, American Cancer Society, Atlanta, Georgia 30329, USA.

27. Genetic Oncology Unit, Instituto de Investigación Sanitaria Galicia Sur (IISGS), Complejo Hospitalario Universitario de Vigo (CHUVI), SERGAS, 36312 Vigo (Pontevedra) Spain.

28. Division of Gastroenterology, Massachusetts General Hospital, Boston, Massachusetts 02114, USA.

29. Harvard Medical School, Boston, Massachusetts 02115, USA.

30. Clinical and Translational Epidemiology Unit, Massachusetts General Hospital, Boston, Massachusetts 02114, USA.

31. Unit of Genetic Epidemiology, Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg 69121, Germany.

32. University Cancer Center Hamburg, University Medical Center Hamburg-Eppendorf, Hamburg 20251, Germany.

33. Cancer Prevention Institute of California, Fremont, California 94538, USA.

34. Ph.D. Program of Cancer Research and Drug Discovery, Taipei Medical University, Taipei 11031, Taiwan.

35. Cancer Science Institute of Singapore, National University of Singapore, Singapore 117599.

36. Department of Colorectal Surgery, Cleveland Clinic, Cleveland, Ohio 44195, USA.

37. Division of Environmental Health Sciences, University of Minnesota, Minneapolis, Minnesota 55455, USA.

38. Van Andel Research Institute, Grand Rapids, Michigan 49502, USA.

39. Prevention and Cancer Control, Cancer Care Ontario, Toronto, Ontario M5G 2L7, Canada.

40. Puerto Rico Cancer Center, University of Puerto Rico, San Juan 70344, Puerto Rico.

41. Genetic Basis of Human Disease Division, Translational Genomics Research Institute, Phoenix, Arizona 85004, USA.

42. Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care / Department of Oncology, University of Cambridge, Cambridge CB1 8RN, United Kingdom.

43. Cancer Epidemiology Centre, Cancer Council Victoria, Melbourne, Victoria 3004, Australia.

44. Division of Human Nutrition, Wageningen University and Research, 6708 PB Wageningen, The Netherlands.

45. University of Michigan Comprehensive Cancer Center, Ann Arbor, Michigan 48105, USA.

46. Menzies Institute for Medical Research, University of Tasmania, Hobart, Tasmania 7001, Australia.

47. W.M. Keck Science Department, Claremont Colleges, Claremont, California 91711, USA.

48. Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, MI 48109, USA.

49. Center for Statistical Genetics, University of Michigan School of Public Health, Ann Arbor, MI 48109, USA.

50. K.G. Jebsen Center for Genetic Epidemiology, Department of Public Health and Nursing,

Norwegian University of Science and Technology, 7491 Trondheim, Sør-Trøndelag, Norway.

51. Department of Medical Oncology, Dana-Farber Cancer Institute, Brookline, Massachusetts 02115, USA.

52. Department of Medicine, Brigham and Women's Institute, Brookline, Massachusetts 02115, USA.

53. Genomic Medicine Group, Galician Foundation of Genomic Medicine, Complejo Hospitalario Universitario de Santiago, Servicio Galego de Saude (SERGAS), Instituto de Investigación Sanitaria de Santiago de Compostela (IDIS), 15706 Santiago De Compostela, Spain.

54. Moores Cancer Center, University of California San Diego, La Jolla, California 92037 USA.

55. Zane Cohen Centre for Digestive Diseases, Mount Sinai Hospital, Toronto, Ontario M5T 3L9, Canada.

56. Channing Division of Network Medicine, Brigham and Women's Institute, Brookline, Massachusetts 02115, USA.

57. Department of Biostatistics, University of Washington, Seattle, Washington 98195, USA.

58. Unidad de Investigacion en Diabetes y Riesgo Cardiovascular, Centro de Investigacion en Salud Poblacional, Instituto Nacional de Salud Publica, 62100, Cuernavaca, Morelos, Mexico.

59. Centro de Estudios en Diabetes AC, 11800, Mexico City, Mexico.

60. Department of Medicine, Division of Gastroenterology, University of Washington School of Medicine, Seattle, Washington 98195, USA.

61. Department of Pathology, University of Michigan, Ann Arbor, Michigan 48104, USA.

62. Nutrition and Metabolism Section, IARC, Lyon 69372 CEDEX 08, France

63. Medical Department 1, University Hospital Dresden, TU Dresden, Dresden 01307, Germany.

64. Department of Radiation Sciences, Oncology, Umea University, Umea 901 87, Sweden.

65. Division of Epidemiology, Department of Population Health, New York University School of Medicine, New York, New York 10016, USA.

66. Centre for MEGA Epidemiology, The University of Melbourne, Carlton, Victoria 3010, Australia.

67. Department of Epidemiology, Murcia Regional Health Council, IMIB-Arrixaca, Murcia 30001, Spain.

68. AbbVie, Redwood City, California 94063, USA.

69. Ontario Institute for Cancer Research, Toronto, Ontario, M5G 0A3, Canada.

70. Program in Genetic Epidemiology and Statistical Genetics, Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts 02115, USA.

71. National Cancer Center, Tokyo 104-0045, Japan.

72. Division of Epidemiology, Center for Public Health Sciences, National Cancer Center, Tokyo 103-0045, Japan.

73. Department of Medicine, Ohio State University, Columbus, Ohio 43210, USA.

74. Department of Epidemiology and Health Promotion, Graduate School of Public Health, Yonsei University, Seoul 120-749, South Korea.

75. State Key Laboratory of Oncology in South China, Cancer Center, Sun Yatsen University, Guangzhou 510060, China.

76. Office of Public Health Studies, University of Hawaii Manoa, Honolulu, Hawaii 96822, USA.

77. Department of Preventive Medicine, Kyushu University, Fukuoka 812-8582, Japan.

78. Epidemiology and Prevention Unit, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan 20133, Italy.

79. Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg 69120, Germany.

80. Clalit Health Services National Israeli Cancer Control Center, Haifa 34361, Israel.

81. Department of Community Medicine and Epidemiology, Carmel Medical Center, Haifa 34361, Israel.

82. Department of Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California 90033, USA.

83. Translational Research Program, Fred Hutchinson Cancer Research Center, Seattle, Washington, 98109, USA.

84. Department of Family Medicine and Community Health, Mary Ann Swetland Center for Environmental Health, Case Comprehensive Cancer Center, Case Western Reserve University, Cleveland, Ohio 44106, USA.

85. Institute of Epidemiology, PopGen Biobank, Christian-Albrechts-University Kiel, Kiel 24105, Germany.

86. Department of Health Science Research, Mayo Clinic, Scottsdale, Arizona 85259, USA.

87. Joint Biobank, Office of Human Research, Taipei Medical University, Taipei 11031, Taiwan.

88. Unit of Oncology, Department of Clinical and Experimental Oncology, Instituto

Oncologico Veneto, IRCCS Padua 35128, Italy.

89. Division of Epidemiology, Vanderbilt Epidemiology Center, Vanderbilt University School of Medicine, Nashville, Tennessee 37203, USA.

90. School of Medicine, Taipei Medical University, Taipei 11031, Taiwan.

91. Sino-American Cancer Foundation, Temple City, California 91780, USA.

92. Harvard School of Public Health, Boston, Massachusetts 02114, USA.

93. Genentech, Inc., CH-4070 Basel, Switzerland.

94. Departments of Medicine and Genetics, Case Comprehensive Cancer Center, Case Western Reserve University, and University Hospitals of Cleveland, Cleveland, Ohio 44106, USA.

95. Research Group on Gene-Environment Interactions and Health, University of León, León 24004, Spain.

96. Laboratory of Genome Technology, Human Genome Center, Institute of Medical Science, The University of Tokyo, Tokyo 108-8639, Japan.

97. Department of Epidemiology, Nagoya University Graduate School of Medicine, Showa-ku, Nagoya 466-8550, Japan.

98. Division of Molecular and Clinical Epidemiology, Aichi Cancer Center Research Institute, Chikusa-Ku Nagoya 464-8681, Japan.

99. Department of Medicine, Clinical Genetics Service, Memorial Sloan Kettering Cancer Center, New York, New York 10065, USA.

100. Cancer Risk Factors and Life-Style Epidemiology Unit, Cancer Research and Prevention Institute-ISPO, Florence 50141, Italy.

101. Department of Surgery, Complejo Hospitalario Universitario de Santiago (CHUS),

Servicio Galego de Saúde (SERGAS), 15782 Santiago De Compostela, Spain.

102. Division of Cancer Biology, University of Puerto Rico, Comprehensive Cancer Center, San Juan 00921, Puerto Rico.

103. Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge CB2 1TN, United Kingdom.

104. Vanderbilt-Ingram Cancer Center, Vanderbilt University, Nashville, Tennessee 37232, USA.

105. Bruce Rappaport Faculty of Medicine, Technion-Israel Institute of Technology, Haifa 32003, Israel.

106. Department of Visceral and Thoracic Surgery, University Hospital Schleswig-Holstein, Kiel Campus, 24105 Kiel, Germany.

107. Department of Medicine and Epidemiology, University of Pittsburgh Medical Center, Pittsburgh, Pennsylvania 15213, USA.

108. Epidemiology and Genomics Research Program, Division of Cancer Control and Population Sciences, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA.

109. Human Genetics Foundation (HuGeF), Torino 10126, Italy.

110. Department of Surgery, Children's Hospital Los Angeles, Los Angeles, California 90027, USA.

111. Department of Surgery, University of Tennessee Health Science Center, Memphis, Tennessee 38163, USA.

112. Department of Internal Medicine, University of Utah Health Sciences Center, Salt Lake City, Utah 84132, USA.

113. Genetic Epidemiology Laboratory, Department of Pathology, University of Melbourne, Melbourne, Victoria 3010, Australia.

114. Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, New York 10065, USA.

115. Department of Medicine, Weill Cornell Medical Center, New York, New York 10065, USA.

116. Department of Hematology and Oncology University of Munich (LMU), Munich 81377, Germany.

117. Phoenix College, Phoenix, Arizona 85103, USA.

118. Mayo Clinic, Rochester, Minnesota 55902, USA.

119. Hellenic Health Foundation, Athens GR-115 27, Greece.

120. Center for Public Health Sciences, National Cancer Center, Tokyo 104-0045, Japan.

121. Department of Chronic Disease Prevention, National Institute for Health and Welfare, FI-00271 Helsinki, Finland.

122. Institute of Environmental Medicine, Karolinska Institutet Solna, SE-171 77 Stockholm, Sweden.

123. Discipline of Genetics, Faculty of Medicine, Memorial University of Newfoundland, St. John's, Newfoundland A1B 3V6, Canada.

124. Department of Nutrition, Harvard School of Public Health, Boston, Massachusetts 02115, USA.

125. State Key Laboratory of Oncogene and Related Genes & Department of Epidemiology, Shanghai Cancer Institute, Shanghai 200032, China.

126. Department of Medical Oncology and Therapeutic Research, City of Hope National

Medical Center, Duarte, California 91010, USA.

127. Clinical Epidemiology Program, Ottawa Hospital Research Institute, Ottawa, Ontario K1Y 4E9, Canada.

128. The University of Ottawa, Ottawa, Ontario K1N 6N5, Canada.

129. Division of Noncommunicable Disease Epidemiology and Southwest Hospital Clinical Research Center, Third Military Medical University, Chongqing 400038, China.

130. Department of Preventive Medicine, Chonnam National University Medical School, Gwangju 61469, South Korea.

131. South Korea Jeonnam Regional Cancer Center, Chonnam National University Hwasun Hospital, Hwasun 58128, South Korea.

132. Department of Medicine, Samuel Oschin Comprehensive Cancer Institute, Cedars-Sinai Medical Center, Los Angeles, California 90048, USA.

133. Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California 90033, USA.

134. Department of Clinical Genetics, Karolinska University Hospital Solna, SE-171 77 Stockholm, Sweden.

135. Department of Molecular Medicine and Surgery, Karolinska Institutet Solna, SE-171 77 Stockholm, Sweden.

136. Department of Epidemiology, School of Public Health, University of Washington, Seattle, Washington 98195, USA.

* Contributed equally to this work

† Jointly supervised this research

**Corresponding Author**

Stephen B. Gruber, MD, PhD, MPH

1441 Eastlake Avenue

Norman Topping Tower (NTT) Room 4436

Los Angeles, CA 90033

Phone: 323-865-0816

Fax: 323-865-0102

Email: sgruber@usc.edu

**Abbreviations**

| | |
|---|---|
| 1KGP | 1000 Genomes Project |
| ACCC | Asia Colorectal Cancer Consortium |
| ChIP-Seq | Chromatin immunoprecipitation and sequencing |
| CORECT | Colorectal Transdisciplinary Study |
| CRC | Colorectal cancer |
| ENCODE | ENCyclopedia Of DNA Elements |
| eQTL | Expression quantitative trait locus |
| GECCO | Genetics and Epidemiology of Colorectal Cancer Consortium |
| GTEx | Genotype-tissue expression |
| GWAS | Genome-wide association study |
| LD | Linkage disequilibrium |
| MAF | Minor allele frequency |
| MEC | Multiethnic cohort |
| OR | Odds ratio |
| PC | Principal component |
| PCA | Principal components analysis |
| QC | Quality control |
| SIGMA | Slim Initiative in Genomic Medicine for the Americas |

**Keywords**

Colorectal cancer; colon cancer; rectal cancer; epidemiology; GWAS; genetics; genome-wide association study; risk factor; genetic epidemiology; variant; susceptibility

**Abstract**

**Background:** Previous genome-wide association studies (GWAS) have identified 42 loci ($P<5x10^{-8}$) associated with the risk of colorectal cancer (CRC). Expanded consortium efforts facilitating the discovery of additional susceptibility loci may capture unexplained familial risk.

**Methods:** We conducted a GWAS in European-descent CRC cases and controls using a discovery-replication design, followed by examination of novel findings in a multiethnic sample (cumulative N=163,315). In the discovery stage (36,948 cases/30,864 controls), we identified genetic variants with minor allele frequency ≥1% associated with risk of CRC using logistic regression followed by a fixed effects inverse variance weighted meta-analysis. All novel independent variants reaching genome-wide statistical significance (two-sided $P<5x10^{-8}$) were tested for replication in separate European-ancestry samples (12,952 cases/48,383 controls). Next, we examined the generalizability of discovered variants in East Asians, African-Americans, and Hispanics (12,085 cases/22,083 controls). Finally, we examined the contributions of novel risk variants to familial relative risk and examined the prediction capabilities of a polygenic risk score. All statistical tests were two-sided.

**Results:** The discovery GWAS identified eleven variants associated with CRC at $P<5x10^{-8}$, of which nine (at 4q22.2/5p15.33/5p13.1/ 6p21.31/6p12.1/10q11.23/12q24.21/16q24.1/ 20q13.13) independently replicated at $P<0.05$. Multiethnic follow-up supported the generalizability of discovery findings. These results provide a 14.7% increase in familial relative risk explained by common risk alleles from 10.3% (95% CI 7.9 – 13.7%; known variants) to 11.85% (95% CI 9.2 – 15.5%; known and novel variants)**.** A polygenic risk score identifies 4.3% of the population at an odds ratio of at least 2.0.

**Conclusions:** This study provides insight into the architecture of common genetic variation

contributing to CRC etiology and improves risk prediction for individualized screening.

## Background

Colorectal cancer (CRC) is a complex polygenetic disease, and heritability accounts for up to 35% of the variation in risk of developing CRC [1, 2]. Some of this heritability is attributable to rare high-penetrance alleles associated with cancer syndromes, now routinely incorporated into clinical care. In addition, genome-wide association studies (GWAS) have identified variation in numerous regulatory regions and other genomic loci that contribute quantifiable risks for CRC development. Specifically, GWAS have identified approximately 70 common genetic variants across 42 regions ($P<5x10^{-8}$) associated with risk of CRC, as larger study populations have been amassed and racial/ethnic representation has increased [3-11]. Expanded consortium efforts facilitating the discovery of additional risk loci may capture unexplained familial risk.

Our prior collaborative work identified six novel CRC susceptibility loci based on a discovery sample of 18,299 cases and 19,656 controls of European ancestral heritage [12]. Results from this GWAS contributed to the development of the Illumina Infinium® OncoArray-500K BeadChip (OncoArray; San Diego, CA), a genotyping array designed to interrogate genomic variation associated with predisposition to five of the most common cancers (prostate, breast, colorectal, lung, and ovarian) [13]. Here, we describe results from a new discovery-replication GWAS, including for the first time findings from the OncoArray Project. Then, we present a follow-up evaluation of genome-wide statistically significant ($P<5x10^{-8}$) risk alleles in individuals from diverse ethnic groups (East Asian, Hispanic, and African American) to investigate if the findings generalize to other populations. Our goal was to discover and replicate new CRC susceptibility loci by assembling the largest international study population to date (N=163,315).

## Methods

*Study Overview*

This investigation included genetic data from 53 observational studies and clinical trials (**Supplementary Figure 1**, **Supplementary Table 1**). In the discovery stage, we combined genotype and epidemiologic data from individuals with European ancestry from all of our consortium efforts to date (CORECT, CCFR, and GECCO), including the new OncoArray Project (36,948 cases and 30,864 controls; **Supplementary Table 2, Supplementary Figures 2 and 3**). In the replication stage, we leveraged data from an independent set of European-descent participants (12,952 cases and 48,383 controls; **Supplementary Table 3**). In the follow-up stage to assess generalizability of findings, we examined data from a multiethnic sample set (12,085 cases and 22,083 controls) that included East Asians from the OncoArray Project (**Supplementary Table 4, Supplementary Figure 4**) and prior studies [14, 15], African Americans [15, 16], and Hispanics/Latinos [17]. Details of the study populations, genotyping, quality control (QC), and imputation for each stage of this GWAS are described in the **Supplementary Methods**. Participants provided written informed consent and the Institutional Review Boards at each center approved the study. For more specific information on consent and study approvals at each institution, see the Supplementary Methods.

*Statistical Analysis*

Detailed descriptions of the statistical analysis for each study stage are described in the **Supplementary Methods**. Briefly, we examined the association between allelic dosage for all autosomal variants with MAF ≥0.01 that passed stringent imputation quality control procedures and CRC status using logistic regression adjusted for appropriate study-specific covariates and principal components that capture global ancestry. Summary statistics from European-descent

samples included in our prior consortium efforts (Discovery Part 1)[18] and the OncoArray Project (Discovery Part 2) were combined in a fixed-effect inverse variance-weighted meta-analysis. Consistency of odds ratios (ORs) across studies were assessed using Cochran's *Q* test of heterogeneity. The most statistically significantly associated variant in each novel genome-wide statistically significant (2-sided $P < 5x10^{-8}$) locus from this discovery analysis was then examined for association with risk of CRC in the independent replication stage of European-ancestry participants (**Supplementary Methods**). Criteria for independent replication included a consistent direction of association and $P<0.05$ based on a meta-analysis of study-specific logistic regression models. Finally, all variants reaching genome-wide statistical significance ($P<5x10^{-8}$) in the discovery stage and $P<0.05$ in the replication stage were assessed for generalizability in the multiethnic follow-up stage of East Asians, African Americans, and Hispanics. All statistical tests were two-sided.

*Polygenic Risk Scores and Familial Relative Risk Explained*

Polygenic risk scores (PRS) in European-descent replication phase participants were calculated using previously known susceptibility variants and novel independently-replicated variants identified by this effort. PRS were categorized into percentile categories based on a weighted sum of risk allele counts among controls (<1%, 1-10%, 10-25%, 25-75%, 75-90%, 90-99%, >99%, with 25-75% serving as the reference). Weights were applied based on bias-corrected logORs from our European-descent discovery analysis. Logistic regression was used to examine CRC risk across PRS categories (after adjusting for age, sex, PCs, and PC*study) for known and known+novel variants, respectively. We also stratified the PRS at a clinically actionable threshold of OR≥2.0. To consider the applicability of our European-derived PRS to

East Asian populations, we also examined the performance of this score in the East Asian cases and controls genotyped on the OncoArray. Next, the contributions to familial risk of the known+novel and the known only variants were investigated. Sample inclusions and methods for bias correction, PRS, and family relative risk explained analyses are described in more detail in the **Supplementary Methods**.

*In Silico Functional Follow-up*

We conducted eQTL analysis in colonic mucosa from healthy controls (N=50) and normal mucosa adjacent to colon cancer (N=100) in the Colonomics study[19] as well as transverse colon tissues (N=169) from the Genotype-Tissue Expression (GTEx) project (**Supplementary Methods**) [20]. Briefly, in Colonomics, for each variant, Pearson partial correlation adjusted for tissue type (healthy or adjacent to tumor) was used to explore the association of dosage SNP/indel data with gene expression for genes located within 2MB of the SNP of interest. For GTEx, the laboratory and analytic methods have previously been described in detail [20].

Additionally, candidate functional variants were identified using published methods [21]. Briefly, index variants and SNPs (CEU, 1KGP, June 2014 release) in LD with each risk variant (we report $r^2 \geq 0.6$ except where noted as $r^2 \geq 0.2$) were aligned with chromatin immunoprecipitation and sequencing (ChIP-seq) tracks for histone methylation and acetylation marks associated with enhancers H3K4me1 and H3K27ac. For this study, we referenced Sigmoid Colon H3K27 acetylation from the Roadmap Epigenomics Consortium[22] as well as CRC cell lines SW480 and HCT-116 H3K4 monomethylation generated in our laboratory (G. Casey) and from the ENCODE project, respectively [23, 24]. To further characterize the novel

CRC genetic risk loci, we performed *in silico* bioinformatic functional annotation of each region.

**Results** *Discovery GWAS (European-descent)*

The discovery GWAS identified 11 common risk variants at 4q22.2, 5q15.33, 5p13.1, 6p21.31, 6p12.1, 10q11.23, 12q24.21, 13q13.2, 16q24.1, 20q11.22, and 20q13.13, all of which were independent of known risk loci (>500kb away or $r^2$>0.2 with a previously known variant) and reached the accepted genome-wide statistical significance threshold ($P<5x10^{-8}$) (**Table 1**). Association results from the discovery stage also indicated that 62 (92.5%) of the 67 known autosomal risk variants (three out of 70 known risk variants were excluded due to MAF<0.01, low quality imputation, or location on chromosome X) replicated at a nominal level of statistical significance ($P<0.05$; **Supplementary Table 5**). A quantile-quantile plot illustrates appropriate control for population stratification with a $\lambda=1.05$ (sample size adjusted $\lambda_{1000}=1.002$; **Supplementary Fig. 5**). A Manhattan plot illustrates the genomic location of novel loci in relation to previously published risk regions (**Figure 1**). Regional association plots in **Supplementary Figure 6** depict the 11 risk variants in the context of their surrounding linkage disequilibrium (LD) structures and nearby genes. The MAFs of these 11 variants in 1KGP Europeans ranged from 0.097 to 0.495, and the ORs for association ranged from 0.90 to 1.08 (**Table 1**). Effect sizes adjusted for potential bias in estimation due to the winner's curse are summarized in **Supplementary Table 6** and **Supplementary Figure 7**.

*Replication (European-descent)*

The association between each of the 11 candidate susceptibility variants identified in the

discovery stage and risk of CRC in an independent sample revealed consistent directions of association and consistent effect sizes for all variants (**Table 1**). Also, ORs for association were statistically significant for 9 of 11 variants. The remaining two loci that were identified in the discovery stage (rs10161980 and rs2295444) demonstrated supportive but not statistically significant evidence of replication, and thus require further validation in future studies. Notably, the two variants with statistical evidence of heterogeneity in the discovery stage meta-analysis replicated in this independent sample set (rs58791712 and rs2696839).

*Multiethnic Follow-up*

Subsequently, we examined the 9 novel, replicated risk variants across three diverse ethnic populations. We examined the association between each variant and risk of CRC in East Asians (N=21,630; **Supplementary Fig. 4**), African Americans (N=6,597), and Hispanics (N=5,941). All 9 variants demonstrated a consistent direction of association in follow-up studies except for rs62404968 and rs10994860 in Hispanics (**Table 2**). Eight out of the 9 variants (all but rs10994860) were associated with the risk of CRC in at least 1 population at a nominal level of statistical significance ($P<0.05$).

*Polygenic Risk Score Analysis and Familial Relative Risk Explained*

PRS analysis conducted in a subset of European-descent replication phase participants revealed that the estimated odds of developing CRC for individuals with scores in the top 1% as compared the 25-75% reference category was 2.18 (**Supplementary Table 7**). Based on the 76 known and novel variants, 4.3% of the study population could be identified for targeted screening based on a clinically actionable threshold of an OR≥2.0 (**Supplementary Table 7**)

[25, 26]. This is in comparison to 1.4% of the study population that is identifiable based on previously known variants only (data not shown). The known + novel PRS performed similarly in East Asians, and the cutpoint to reach a clinically actionable OR of at least 2.0 in this population was 99.1% (**Supplementary Table 7**).

Overall, 76 variants explained 11.9% (95% CI 9.2 – 15.5%) of the known familial relative risk as compared to 10.3% (95% CI 7.9 – 13.7%) for the previously known variants only. This represents a 14.7% increase in familial relative risk explained. Estimation of the proportion of explained familial risk incorporated uncertainty in risk estimation for each variant and uncertainty in the specification of the familial relative risk.

*eQTL Analysis*

Analysis of *cis* gene expression data for the 9 novel susceptibility variants revealed several noteworthy eQTLs in Colonomics and GTEx transverse colon samples (**Supplementary Table 8**). For example, rs10994860 is a statistically significant eQTL for *ASAH2* (effect size=-0.61; $P$=5.7E x $10^{-5}$ ). Further, in the Colonomics dataset, rs6906359 is a statistically significant eQTL for several genes including *BRPF3*, showing over-expression for C/C as compared to T/T genotypes (partial $r^2$=0.09, $P$=2.6 x $10^{-4}$). The most statistically significant eQTLs in each region with at least one variant associated at the $P$<0.05 level in the Colonomics dataset are summarized in **Supplementary Figure 8**.

**Discussion**

This collaborative study included over 163,000 individuals for the identification and

further evaluation of 9 replicable novel CRC genetic susceptibility loci. Nine low-penetrance risk loci represent approximately a 21% increase from those previously discovered to date (N=42). Nine risk variants replicated in an independent sample of European-ancestry participants and 8 of those generalized to at least one of three other racial/ethnic populations. Our findings contribute substantially to the known familial relative risk explained by low penetrance susceptibility alleles, with a 14.7% increase from 10.3% (previously known only) to 11.9% (known + novel reported here) explained. Further, PRS analysis underscores the impact of common CRC risk alleles, particularly among individuals with the highest counts of risk variants. Our findings suggest that 4.3% of the population could be targeted for earlier and more frequent screening based on germline genetic profiling of all known common CRC susceptibility variants. This supports our previous findings that GWAS have the potential to inform appropriate tailoring of screening guidelines to population subgroups [27].

The consistent direction of association for all 9 novel risk variants in East Asians and African Americans (all but 2 in Hispanics) underscores the generalizability of our findings from European-ancestry individuals. However, the statistically significant association of some but not all variants with CRC risk across the additional ethnic subgroups supports the importance of expanded sample sizes in certain populations as well as ongoing multiethnic fine-mapping studies to identify the strongest signals and most likely putative functional variant(s) at particular loci in other ancestral populations.

Two of the 9 risk alleles map to intragenic or coding regions. First, rs62404968 maps to 6p12.1 and lies within an intron of *BMP5*. *BMP5* encodes bone morphogenetic protein 5 that is part of the transforming growth factor-beta (TGF-β) superfamily. Members of the BMP and TGF-β family have been implicated as risk genes for CRC in previous GWAS, including *BMP2*

and *BMP4* on chromosomes 20 and 14, respectively [28]. The associated SNP, rs62404968 or any of the 20 SNPs in LD, do not map to any predicted regulatory/enhancer regions based on histone marks suggesting that further functional follow up is needed to understand the functional mechanism likely acting on the strong candidate gene *BMP5*. Second, rs10994860 maps to 10q11.23 and lies within exon 1 of *A1CF*, representing a putative candidate functional SNP. *A1CF* (APOBEC1 Complementation Factor) is a critical component of the apolipoprotein B mRNA editing enzyme complex. There are two SNPs (rs71457593 and rs10994720) in LD with rs10994860 that both map to histone peaks also suggesting potential functionality.

The remaining seven risk alleles map to intergenic regions of the genome. SNP rs1370821 maps to 4q22.2, with the two nearest genes being *ATOH1* and *SMARCAD1* (approximately 85kb away). *ATOH1* encodes atonal homolog BHLH transcription factor 1 that belongs to the basic helix-loop-helix family of transcription factors. *SMARCAD1* encodes Matrix-Associated Actin-Dependent Regulator Of Chromatin, a member of the SNF subfamily of helicase proteins that play an important role in heterochromatin reorganization following DNA replication. While the associated SNP, rs1370821, does not map to any candidate regulatory regions, two SNPs (rs2510787; rs2433324) in LD with rs1370821 lie within an intron of *PDLIM5* (encoding PDZ and LIM domain protein 5), and both map to histone marks. Also, rs1370821 warrants further functional characterization because of its proximity to *BMPR1B*, a gene where there is statistical evidence of an eQTL relationship by genotype in the Colonomics dataset and where the gene family is related to polyposis and CRC susceptibility [17].

The indel rs58791712 (G/GT) maps to 5p13.1. The nearest genes, *PTGER4* and *LINC00603*, lie approximately 400kb from the index variant. *PTGER4* encodes PGE2 Receptor EP4 Subtype and is one of four receptors identified for prostaglandin E2. This indel does not

map to any histone marks making it unlikely to be a functional variant. However, there are three SNPs (rs72748452, rs755989 and rs4957261) in LD with rs58791712 that overlap histone peaks.

The SNP rs2735940 maps to 5p15.33 and lies adjacent to the *TERT* gene. *TERT* encodes the Telomerase Catalytic Subunit protein that helps to maintain telomere ends by addition of the telomere repeat TTAGGG. *TERT* has been identified previously as a candidate risk gene in several cancers including CRC [29-34]. SNP rs2735940 does not map to any histone marks. However, this SNP is in LD with three SNPs (rs380145, rs246995 and rs246994) that map to histone marks and lie within an intron of *CLPTM1L* (rs380145) or the predicted gene *BC034612* (rs246995 and rs246994).

The SNP rs6906359 maps to 6p21.31, and the closest gene is *FKBP5* approximately 12kb away. *FKBP5* encodes FK506 Binding Protein 5, a member of the immunophilin protein family that plays a role in immunoregulation, protein folding, and trafficking. However, rs6906359 does not overlap any histone marks. Of the SNPs in LD with rs6906359 that overlap histone peaks, two SNPs (rs72894781 and rs72894784) map within an intron of *TEAD3*, one SNP (rs16878812) maps within an intron of *FKBP5*, and one SNP is intergenic (rs45493300).

The indel rs72013726 (CACAA/C) maps to 12q24.21. The nearest gene, *MED13L*, lies approximately 500kb from rs72013726. *MED13L* encodes Thyroid Hormone Receptor-Associated Protein 2 and is one of many proteins that function as a transcriptional coactivator for RNA polymerase II-transcribed genes. SNP rs72013726 maps to a histone peak, making it a potential functional SNP.

SNP rs2696839 maps to 16q24.1 and lies 15kb from the predicted gene *LOC146513.* While this SNP does not map to any histone marks, all four SNPs (rs12932862, rs12149163, rs12149501, and rs2665316) in LD with rs2696839 do. Of note, there are several lncRNAs in

this region.

SNP rs1810502 maps to 20q13.13 near the gene *PTPN1*, approximately 70kb away. *PTPN1* encodes Protein-Tyrosine Phosphatase 1B a member of the protein tyrosine phosphatase family. This SNP and 14 other SNPs in LD with rs1810502 map to histone marks, implying the possibility that any one of these 15 SNPs could be functionally relevant to CRC etiology.

Our study design has strengths and limitations. We conducted a rigorous two-stage study with discovery and independent replication in European-descent participants. Further, a major strength is that we utilized data from the independent replication phase to conduct PRS and familial relative risk explained analyses. Of note, despite a 14.7% increase beyond prior knowledge, still <12% of familial relative risk is explained by GWAS-identified alleles including our new 9 loci. Thus, additional efforts are needed to fully explain the genetic architecture of this complex disease, potentially with gene-environment interactions. Space limitations preclude detailed descriptions of eQTL analyses for each SNP. However, we found little or no evidence of the 9 novel index SNPs in relation to gene expression for our speculatively implicated genes. Additional eQTL analyses in expanded normal colon tissue sample sets that examine the full landscape of SNPs in LD with the index SNP may help to elucidate the impact of germline susceptibility loci on gene expression. Future studies will be advantageous to identify rare and intermediate frequency susceptibility alleles through expanded sample size as well as increased racial/ethnic minority inclusion. Multiethnic samples will be useful for fine-mapping known and novel risk regions as well as for identifying population-specific variation. In summary, this GWAS provides insight into the etiologies of CRC and provides a basis for future fine-mapping, functional characterization, and risk modeling research.

**Funding**

CORECT: The CORECT Study was supported by the National Cancer Institute, National Institutes of Health (NCI/NIH), U.S. Department of Health and Human Services (grant numbers U19 CA148107, R01 CA81488, P30 CA014089, R01 CA197350,; P01 CA196569; R01 CA201407) and National Institutes of Environmental Health Sciences, National Institutes of Health (grant number T32 ES013678). The ATBC Study was supported by the US Public Health Service contracts (N01-CN-45165, N01-RC-45035, N01-RC-37004, and HHSN261201000006C) from the National Cancer Institute. The Cancer Prevention Study-II Nutrition Cohort is funded by the American Cancer Society. ColoCare: This work was supported by the National Institutes of Health (grant numbers R01 CA189184, U01 CA206110, 2P30CA015704-40 (Gilliland)), the Matthias Lackas-Foundation, the German Consortium for Translational Cancer Research, and the EU TRANSCAN initiative. GECCO: Funding for GECCO was provided by the National Cancer Institute, National Institutes of Health, U.S. Department of Health and Human Services (grant numbers U01 CA137088, R01 CA059045, U01 CA164930). The Colon Cancer Family Registry (CFR) Illumina GWAS was supported by funding from the National Cancer Institute, National Institutes of Health (grant numbers U01 CA122839, R01 CA143247). The Colon CFR/CORECT Affymetrix Axiom GWAS and OncoArray GWAS were supported by funding from National Cancer Institute, National Institutes of Health (grant number U19 CA148107 to S Gruber). The Colon CFR participant recruitment and collection of data and biospecimens used in this study were supported by the National Cancer Institute, National Institutes of Health (grant number UM1 CA167551) and through cooperative agreements with the following Colon CFR centers: Australasian Colorectal Cancer Family Registry (NCI/NIH grant numbers U01 CA074778 and U01/U24 CA097735),

USC Consortium Colorectal Cancer Family Registry (NCI/NIH grant numbers U01/U24 CA074799), Mayo Clinic Cooperative Family Registry for Colon Cancer Studies (NCI/NIH grant number U01/U24 CA074800), Ontario Familial Colorectal Cancer Registry (NCI/NIH grant number U01/U24 CA074783), Seattle Colorectal Cancer Family Registry (NCI/NIH grant number U01/U24 CA074794), and University of Hawaii Colorectal Cancer Family Registry (NCI/NIH grant number U01/U24 CA074806), Additional support for case ascertainment was provided from the Surveillance, Epidemiology and End Results (SEER) Program of the National Cancer Institute to Fred Hutchinson Cancer Research Center (Control Nos. N01-CN-67009 and N01-PC-35142, and Contract No. HHSN2612013000121), the Hawai‘i Department of Health (Control Nos. N01-PC-67001 and N01-PC-35137, and Contract No. HHSN26120100037C, and the California Department of Public Health (contracts HHSN261201000035C awarded to the University of Southern California, and the following state cancer registries: AZ, CO, MN, NC, NH, and by the Victoria Cancer Registry and Ontario Cancer Registry. ESTHER/VERDI was supported by grants from the Baden-Württemberg Ministry of Science, Research and Arts and the German Cancer Aid. MCCS cohort recruitment was funded by VicHealth and Cancer Council Victoria. GALEON: FIS Intrasalud (PI13/01136). The MCCS was further supported by Australian NHMRC grants 509348, 209057, 251553 and 504711 and by infrastructure provided by Cancer Council Victoria. Cases and their vital status were ascertained through the Victorian Cancer Registry (VCR) and the Australian Institute of Health and Welfare (AIHW), including the National Death Index and the Australian Cancer Database. MSKCC: The work at Sloan Kettering in New York was supported by the Robert and Kate Niehaus Center for Inherited Cancer Genomics and the Romeo Milio Foundation. Moffitt: This work was supported by funding from the National Institutes of Health (grant numbers R01 CA189184, P30 CA076292),

Florida Department of Health Bankhead-Coley Grant 09BN-13, and the University of South Florida Oehler Foundation. Moffitt contributions were supported in part by the Total Cancer Care Initiative, Collaborative Data Services Core, and Tissue Core at the H. Lee Moffitt Cancer Center & Research Institute, a National Cancer Institute-designated Comprehensive Cancer Center (grant number P30 CA076292). SEARCH: Cancer Research UK (C490/A16561). The Spanish study was supported by Instituto de Salud Carlos III, co-funded by FEDER funds –a way to build Europe– (grants PI14-613 and PI09-1286), Catalan Government DURSI (grant 2014SGR647), and Junta de Castilla y León (grant LE22A10-2). The Swedish Low-risk Colorectal Cancer Study: The study was supported by grants from the Swedish research council; K2015-55X-22674-01-4, K2008-55X-20157-03-3, K2006-72X-20157-01-2 and the Stockholm County Council (ALF project). Research and contributions in Taiwan and Taipei Medical University were funded by Taiwan Ministry of Health and Wealth (MOHW105). CIDR genotyping for the Oncoarray was conducted under contract 268201200008I (to K. Doheny), through grant 101HG007491-01 (to C.I. Amos). The Norris Cotton Cancer Center - P30CA023108, The Quantitative Biology Research Institute - P20GM103534, and the Coordinating Center for Screen Detected Lesions - U01CA196386 also supported efforts of C.I. Amos. This work was also supported by the National Cancer Institute (grant numbers U01 CA1817700, R01 CA144040). ASTERISK: a Hospital Clinical Research Program (PHRC) and supported by the Regional Council of Pays de la Loire, the Groupement des Entreprises Françaises dans la Lutte contre le Cancer (GEFLUC), the Association Anne de Bretagne Génétique and the Ligue Régionale Contre le Cancer (LRCC). COLO2&3: National Institutes of Health (grant number R01 CA060987). DACHS: German Research Council (Deutsche Forschungsgemeinschaft, BR 1704/6-1, BR 1704/6-3, BR 1704/6-4 and CH 117/1-1), and the

German Federal Ministry of Education and Research (01KH0404 and 01ER0814). DALS: National Institutes of Health (grant number R01 CA048998 to M.L.S); HPFS is supported by National Institutes of Health (grant numbers P01 CA055075, UM1 CA167552, R01 137178, and P50 CA127003), NHS by the National Institutes of Health (grant numbers UM1 CA186107, R01 CA137178, P01 CA087969, and P50 CA127003), NHS II by the National Institutes of Health (grant numbers R01 050385CA and UM1 CA176726), and PHS by the National Institutes of Health (grant number R01 CA042182). MEC: National Institutes of Health (grant numbers R37 CA054281, P01 CA033619, and R01 CA063464). OFCCR: National Institutes of Health, through funding allocated to the Ontario Registry for Studies of Familial Colorectal Cancer (grant number U01 CA074783); see Colon CFR section above. As subset of ARCTIC, OFCCR is supported by a GL2 grant from the Ontario Research Fund, the Canadian Institutes of Health Research, and the Cancer Risk Evaluation (CaRE) Program grant from the Canadian Cancer Society Research Institute. Thomas J. Hudson and Brent W. Zanke are recipients of Senior Investigator Awards from the Ontario Institute for Cancer Research, through generous support from the Ontario Ministry of Research and Innovation. PLCO: Intramural Research Program of the Division of Cancer Epidemiology and Genetics and supported by contracts from the Division of Cancer Prevention, National Cancer Institute, NIH, DHHS. Additionally, a subset of control samples was genotyped as part of the Cancer Genetic Markers of Susceptibility (CGEMS) Prostate Cancer GWAS[35], Colon CGEMS pancreatic cancer scan (PanScan)[36, 37], and the Lung Cancer and Smoking study [38]. The prostate and PanScan study datasets were accessed with appropriate approval through the dbGaP online resource (http://cgems.cancer.gov/data/) accession numbers phs000207.v1.p1 and phs000206.v3.p2, respectively, and the lung datasets were accessed from the dbGaP website (http://www.ncbi.nlm.nih.gov/gap) through accession

number phs000093.v2.p2. Funding for the Lung Cancer and Smoking study was provided by National Institutes of Health (NIH), Genes, Environment and Health Initiative (GEI) Z01 CP 010200, NIH U01 HG004446, and NIH GEI U01 HG 004438. For the lung study, the GENEVA Coordinating Center provided assistance with genotype cleaning and general study coordination, 23 and the Johns Hopkins University Center for Inherited Disease Research conducted genotyping. PMH: National Institutes of Health (grant number R01 CA076366). VITAL: National Institutes of Health (grant number K05 CA154337). WHI: The WHI program is funded by the National Heart, Lung, and Blood Institute, National Institutes of Health, U.S. Department of Health and Human Services through contracts HHSN268201600018C, HHSN268201600001C, HHSN268201600002C, HHSN268201600003C, and HHSN268201600004C. COLON: The COLON study is sponsored by Wereld Kanker Onderzoek Fonds, including funds from grant 2014/1179 as part of the World Cancer Research Fund International Regular Grant Programme, by Alpe d'Huzes and the Dutch Cancer Society (UM 2012–5653, UW 2013-5927, UW2015-7946), and by TRANSCAN (JTC2012-MetaboCCC, JTC2013-FOCUS). NQplus: The NQplus study is sponsored by a ZonMW investment grant (98-10030); by PREVIEW, the project PREVention of diabetes through lifestyle intervention and population studies in Europe and around the World (PREVIEW) project which received funding from the European Union Seventh Framework Programme (FP7/2007–2013) under grant no. 312057; by funds from TI Food and Nutrition (cardiovascular health theme), a public–private partnership on pre-competitive research in food and nutrition; and by FOODBALL, the Food Biomarker Alliance, a project from JPI Healthy Diet for a Healthy Life. ACCC: The work at Vanderbilt University School of Medicine was supported partly by U.S. National Institutes of Health (grant numbers R37 CA070867, R01 CA182910, R01 CA148667), as well as Anne Potter

Wilson funds from Vanderbilt University School of Medicine. Participating studies (grant support) in ACCC are: Shanghai Women's Health Study (grant numbers R37 CA070867 and UM1CA182910), Shanghai Men's Health Study (grant number UM1CA173640), Shanghai Colorectal Cancer Study 3 (US NIH, grant numbers R37CA070867, R01CA188214 and Ingram Professorship funds), Guangzhou Colorectal Cancer Study (National Key Scientific and Technological Project – 2011ZX09307-001-04), the Japan BioBank Colorectal Cancer Study (grant from the Ministry of Education, Culture, Sports, Science and Technology of the Japanese government), the Hwasun Cancer Epidemiology Study–Colon and Rectum Cancer (HCES-CRC; grants from Chonnam National University Hwasun Hospital, HCRI15011-1), Aichi Colorectal Cancer Study (Grant-in-aid for Cancer Research, Grant for the Third Term Comprehensive Control Research for Cancer and Grants-in-Aid for Scientific Research from the Japanese Ministry of Education, Culture, Sports, Science and Technology, Nos. 17015018 and 221S0001), and KCPS-II Colorectal Cancer Study (National R&D Program for cancer control, 1220180; Seoul R&D Program, 10526). US-Japan Colorectal Cancer GWAS: The colorectal cancer GWAS among Japanese was funded through US National Institutes of Health (grant numbers R01 CA126895, R01 CA104132, U24 CA074806). Genotyping of the additional MEC controls was funded through the National Institutes of Health (grant numbers R01 CA132839, U01 HG004726). MEC was funded through the National Institutes of Health (grant numbers R37 CA054281, P01 CA033619, R01 CA063464). JPHC was supported by the National Cancer Center Research and Development Fund (since 2011) and a Grant-in-Aid for Cancer Research (from 1989 to 2010) from the Ministry of Health, Labor and Welfare of Japan. The Fukuoka Colorectal Cancer Study was funded by the Ministry of Education, Culture, Sports, Science and Technology, Japan. Hispanic Colorectal Cancer GWAS: This work was supported by the

National Institutes of Health (grant numbers R01 CA155101, U01 HG004726, R01 CA140561, T32 ES013678, U19 CA148107, P30 CA014089). The US-Japan Colorectal Cancer GWAS and the African American Colorectal Cancer GWAS were funded through the US National Institutes of Health (grant numbers 1R01-CA126895, 1R01-CA126895-S1, 1R01-CA104132, 2U24-CA074806). The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health. Additional funds were provided by the NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. Donors were enrolled at Biospecimen Source Sites funded by NCI\SAIC-Frederick, Inc. (SAIC-F) subcontracts to the National Disease Research Interchange (10XS170), Roswell Park Cancer Institute (10XS171), and Science Care, Inc. (X10S172). The Laboratory, Data Analysis, and Coordinating Center (LDACC) was funded through a contract (HHSN268201000029C) to The Broad Institute, Inc. Biorepository operations were funded through an SAIC-F subcontract to Van Andel Institute (10ST1035). Additional data repository and project management were provided by SAIC-F (HHSN261200800001E). The Brain Bank was supported by a supplement to University of Miami grants DA006227 & DA033684 and to contract N01MH000028. Statistical Methods development grants were made to the University of Geneva (MH090941 & MH101814), the University of Chicago (MH090951, MH090937, MH101820, MH101825), the University of North Carolina - Chapel Hill (MH090936 & MH101819), Harvard University (MH090948), Stanford University (MH101782), Washington University St Louis (MH101810), and the University of Pennsylvania (MH101822). The data used for the analyses described in this manuscript were obtained from the GTEx Portal on 10/19/2016.

**Notes**

The funders had no role in the design of the study; the collection, analysis, and interpretation of the data; the writing of the manuscript; and the decision to submit the manuscript for publication. Colorectal Transdisciplinary Study (CORECT): The content of this manuscript does not necessarily reflect the views or policies of the National Cancer Institute or any of the collaborating centers in the CORECT Consortium, nor does mention of trade names, commercial products or organizations imply endorsement by the US Government or the CORECT Consortium.

We thank Alina Hoehn for her valuable contributions to table/figure generation and organization of this manuscript. We are incredibly grateful for the contributions of Dr. Brian Henderson and Dr. Roger Green over the course of this study and acknowledge them in memoriam. We are also grateful for support from Daniel and Maryann Fong. ColoCare: We thank the many investigators and staff who made this research possible in ColoCare Seattle and ColoCare Heidelberg. ColoCare was initiated and developed at the Fred Hutchinson Cancer Research Center by Drs. Ulrich and Grady. CCFR: The Colon CFR graciously thanks the generous contributions of their study participants, dedication of study staff, and financial support from the U.S. National Cancer Institute, without which this important registry would not exist. Galeon: GALEON wishes to thank the Department of Surgery of University Hospital of Santiago (CHUS), Sara Miranda Ponte, Carmen M Redondo, and the staff of the Department of Pathology and Biobank of CHUS, Instituto de Investigación Sanitaria de Santiago (IDIS), Instituto de Investigación Sanitaria Galicia Sur (IISGS), SERGAS, Vigo, Spain, and Programa Grupos Emergentes, Cancer Genetics Unit, CHUVI Vigo Hospital, Instituto de Salud Carlos III, Spain. MCCS: This study was made possible by the contribution of many people, including the original

investigators and the diligent team who recruited participants and continue to work on follow-up. We would also like to express our gratitude to the many thousands of Melbourne residents who took part in the study and provided blood samples.

SEARCH: We acknowledge the contributions of Mitul Shah, Val Rhenius, Sue Irvine, Craig Luccarini, Patricia Harrington, Don Conroy, Rebecca Mayes, and Caroline Baynes.

The Swedish low-risk colorectal cancer study: we thank Berith Wejderot and the Swedish low-risk colorectal cancer study group.

Genetics & Epidemiology of Colorectal Cancer Consortium (GECCO): we thank all those at the GECCO Coordinating Center for helping bring together the data and people that made this project possible.

ASTERISK: we are very grateful to Dr. Bruno Buecher without whom this project would not have existed. We also thank all those who agreed to participate in this study, including the patients and the healthy control persons, as well as all the physicians, technicians and students.

DACHS: we thank all participants and cooperating clinicians, and Ute Handte-Daub, Renate Hettler-Jensen, Utz Benscheid, Muhabbet Celik and Ursula Eilber for excellent technical assistance.

HPFS, NHS and PHS: we acknowledge Patrice Soule and Hardeep Ranu of the Dana-Farber Harvard Cancer Center High-Throughput Polymorphism Core who assisted in the genotyping for NHS, HPFS and PHS under the supervision of Dr. Immaculata Devivo and Dr. David Hunter, Qin (Carolyn) Guo and Lixue Zhu who assisted in programming for NHS and HPFS and Haiyan Zhang who assisted in programming for the PHS. We thank the participants and staff of the Nurses' Health Study and the Health Professionals Follow-Up Study, for their valuable contributions as well as the following state cancer registries for their help: AL, AZ, AR, CA, CO,

CT, DE, FL, GA, ID, IL, IN, IA, KY, LA, ME, MD, MA, MI, NE, NH, NJ, NY, NC, ND, OH, OK, OR, PA, RI, SC, TN, TX, VA, WA, WY. In addition, this study was approved by the Connecticut Department of Public Health (DPH) Human Investigations Committee. Certain data used in this publication were obtained from the DPH. We assume full responsibility for analyses and interpretation of these data. PLCO: we thank Drs. Christine Berg and Philip Prorok, Division of Cancer Prevention, National Cancer Institute, the Screening Center investigators and staff or the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial, Mr. Tom Riley and staff, Information Management Services Inc., Ms. Barbara O'Brien and staff, Westat Inc. and Drs. Bill Kopp, Wen Shao and staff, SAIC-Frederick. Most importantly, we acknowledge the study participants for their contributions for making this study possible. The statements contained herein are solely those of the authors and do not represent or imply concurrence or endorsement by NCI.

PMH: We thank the study participants and staff of the Hormones and Colon Cancer study.

WHI: we thank the WHI investigators and staff for their dedication, and the study participants for making the program possible. A full listing of WHI investigators can be found at https://cleo.whi.org/researchers/Documents%20%20Write%20a%20Paper/WHI%20Investigator%20Short 20List.pdf.

COLON and NQplus: the authors would like to thank the COLON and NQplus investigators at Wageningen University & Research and the involved clinicians in the participating hospitals.

EPIC: We thank all participants and health care personnel in the Västerbotten Intervention Programme, as well as the Department of biobank research, Umeå University, and Biobanken norr, Västerbotten County Council.

UK Biobank: This research has been conducted using the UK Biobank Resource under

Application Number 8614. ACCC: We thank all study participants and research staff of all studies for their contributions and commitment to this project, Regina Courtney for DNA preparation and Jing He for data processing. Aichi Colorectal Cancer Study appreciates the support of Cancer Bio Bank Aichi for this project.

Hispanic Colorectal Cancer Study: we are indebted to the individuals who participated in this study. Without their assistance, we could not have conducted any of our research. We would like to thank Nathalie Nguyen, Julissa Ramirez, Yaquelin Perez, Daniel Collin, Alicia Rivera, Lauren Gerstmann, and the student intern staff for their assistance in logistical support and management, interviewing patients, and data entry. Finally, we would like to especially acknowledge Dr. Brian E. Henderson, who passed away before this paper was submitted. Without his mentorship and tremendous efforts in co-founding the Multiethnic Cohort, this work would not have been possible.

SIGMA: we would like to acknowledge all participants and investigators in this study including: Teresa Tusié-Luna, Carlos A. Aguilar-Salinas, Hortensia Moreno-Macías, Alicia Huerta-Chagoya, María Luisa Ordóñez-Sánchez, Rosario Rodríguez-Guillén, Ivette Cruz-Bautista, Maribel Rodríguez-Torres, Linda Liliana Muñóz-Hernández, Olimpia Arellano-Campos, Donají Gómez, Ulices Alvirde.

Competing Financial Interest

Christoph Mancao is an employee of Genentech and holds shares/stocks from Roche/Genentech. The other authors have no competing interests to declare.

**References**

1. Lichtenstein P, Holm NV, Verkasalo PK, *et al.* Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. N Engl J Med 2000;343(2):78-85.

2. Burt R. Inheritance of Colorectal Cancer. Drug Discov Today Dis Mech 2007;4(4):293-300.

3. Peters U, Bien S, Zubair N. Genetic architecture of colorectal cancer. Gut 2015;64(10):1623-36.

4. Lemire M, Qu C, Loo LW, *et al.* A genome-wide association study for colorectal cancer identifies a risk locus in 14q23.1. Hum Genet 2015;134(11-12):1249-62.

5. Orlando G, Law PJ, Palin K, *et al.* Variation at 2q35 (PNKD and TMBIM1) influences colorectal cancer risk and identifies a pleiotropic effect with inflammatory bowel disease. Hum Mol Genet 2016; 10.1093/hmg/ddw087.

6. Zeng C, Matsuda K, Jia WH, *et al.* Identification of Susceptibility Loci and Genes for Colorectal Cancer Risk. Gastroenterology 2016;150(7):1633-45.

7. Gruber SB, Moreno V, Rozek LS, *et al.* Genetic variation in 8q24 associated with risk of colorectal cancer. Cancer Biol Ther 2007;6(7):1143-7.

8. Broderick P, Carvajal-Carmona L, Pittman AM, *et al.* A genome-wide association study shows that common alleles of SMAD7 influence colorectal cancer risk. Nat Genet 2007;39(11):1315-7.

9. Carvajal-Carmona LG, Cazier JB, Jones AM, *et al.* Fine-mapping of colorectal cancer susceptibility loci at 8q23.3, 16q22.1 and 19q13.11: refinement of association signals and use of in silico analysis to suggest functional variation and unexpected candidate target genes. Hum Mol Genet 2011;20(14):2879-88.

10. Houlston RS, Cheadle J, Dobbins SE, *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 2010;42(11):973-7.

11. Houlston RS, Webb E, Broderick P, *et al.* Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. Nat Genet 2008;40(12):1426-35.

12. Schumacher FR, Schmit SL, Jiao S, *et al.* Genome-wide association study of colorectal cancer identifies six new susceptibility loci. Nat Commun 2015;6:7138.

13. Amos CI, Dennis J, Wang Z, *et al.* The OncoArray Consortium: a Network for Understanding the Genetic Architecture of Common Cancers. Cancer Epidemiol Biomarkers Prev 2016; 10.1158/1055-9965.EPI-16-0106.

14. Zhang B, Jia WH, Matsuda K, *et al.* Large-scale genetic study in East Asians identifies six new loci associated with colorectal cancer risk. Nat Genet 2014;46(6):533-42.

15. Wang H, Burnett T, Kono S, *et al.* Trans-ethnic genome-wide association study of colorectal cancer identifies a new susceptibility locus in VTI1A. Nat Commun 2014;5:4613.

16. Wang H, Haiman CA, Burnett T, *et al.* Fine-mapping of genome-wide association study-identified risk loci for colorectal cancer in African Americans. Hum Mol Genet 2013;22(24):5048-55.

17. Schmit SL, Schumacher FR, Edlund CK, *et al.* Genome-wide association study of colorectal cancer in Hispanics. Carcinogenesis 2016;37(6):547-56.

18. Schumacher FR, Schmit SL, Jiao S, *et al.* Genome-wide association study of colorectal cancer identifies six new susceptibility loci. Nat Commun 2015;6:7138.

19. Closa A, Cordero D, Sanz-Pamplona R, *et al.* Identification of candidate susceptibility genes for colorectal cancer through eQTL analysis. Carcinogenesis 2014;35(9):2039-46.

20. Consortium GT. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science 2015;348(6235):648-60.

21. Fortini BK, Tring S, Plummer SJ, *et al.* Multiple functional risk variants in a SMAD7 enhancer implicate a colorectal cancer risk haplotype. PLoS One 2014;9(11):e111914.

22. Bernstein BE, Stamatoyannopoulos JA, Costello JF, *et al.* The NIH Roadmap Epigenomics Mapping Consortium. Nat Biotechnol 2010;28(10):1045-8.

23. O'Geen H, Echipare L, Farnham PJ. Using ChIP-seq technology to generate high-resolution profiles of histone modifications. Methods Mol Biol 2011;791:265-86.

24. Consortium EP. The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 2004;306(5696):636-40.

25. Tung N, Domchek SM, Stadler Z, *et al.* Counselling framework for moderate-penetrance cancer-susceptibility mutations. Nat Rev Clin Oncol 2016;13(9):581-8.

26. Network NCC. *Genetic/Familial High-Risk Assessment: Colorectal (Version 2.2016)*. https://www.nccn.org/professionals/physician_gls/f_guidelines.asp.

27. Hsu L, Jeon J, Brenner H, *et al.* A model to determine colorectal cancer risk using common genetic susceptibility loci. Gastroenterology 2015;148(7):1330-9 e14.

28. Houlston RS, Webb E, Broderick P, *et al.* Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. Nat Genet 2008;40(12):1426-35.

29. Karami S, Han Y, Pande M, *et al.* Telomere structure and maintenance gene variants and risk of five cancer types. Int J Cancer 2016; 10.1002/ijc.30288.

30. Wang Z, Zhu B, Zhang M, *et al.* Imputation and subset-based association analysis across different cancer types identifies multiple independent risk loci in the TERT-CLPTM1L region on chromosome 5p15.33. Hum Mol Genet 2014;23(24):6616-33.

31. Walsh KM, Codd V, Smirnov IV, *et al.* Variants near TERT and TERC influencing telomere length are associated with high-grade glioma risk. Nat Genet 2014;46(7):731-5.

32. Bojesen SE, Pooley KA, Johnatty SE, *et al.* Multiple independent variants at the TERT locus are associated with telomere length and risks of breast and ovarian cancer. Nat Genet 2013;45(4):371-84, 384e1-2.

33. Haiman CA, Chen GK, Vachon CM, *et al.* A common variant at the TERT-CLPTM1L locus is associated with estrogen receptor-negative breast cancer. Nat Genet 2011;43(12):1210-4.

34. Kinnersley B, Migliorini G, Broderick P, *et al.* The TERT variant rs2736100 is associated with colorectal cancer risk. Br J Cancer 2012;107(6):1001-8.

35. Yeager M, Orr N, Hayes RB, *et al.* Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. Nat Genet 2007;39(5):645-9.

36. Amundadottir L, Kraft P, Stolzenberg-Solomon RZ, *et al.* Genome-wide association study identifies variants in the ABO locus associated with susceptibility to pancreatic cancer. Nat Genet 2009;41(9):986-90.

37. Petersen GM, Amundadottir L, Fuchs CS, *et al.* A genome-wide association study identifies pancreatic cancer susceptibility loci on chromosomes 13q22.1, 1q32.1 and 5p15.33. Nat Genet 2010;42(3):224-8.

38. Landi MT, Chatterjee N, Yu K, *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 2009;85(5):679-91.

**Table 1.** Eleven novel low-penetrance risk variants identified from the discovery GWAS (European-descent) with $P<5x10^{-8}$ and their results in an independent replication set.

| Locus | EFF/REF allele | rsID:CHR:BP | FRQ_EFF (1KGP EUR) | Discovery ($N_{case}$ = 36,948; $N_{control}$ = 30,864) | | | | Replication ($N_{case}$=12,952; $N_{control}$=48,383) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | OR (95%CI) | P* | $I^2$, % | $P_{heterogeneity}$† | OR (95%CI) | P* | $I^2$, % | $P_{heterogeneity}$† |
| 4q22.2 | T/C | rs1370821:4:94943383 | 0.401 | 1.07 (1.04 - 1.09) | $4.0 \times 10^{-8}$ | 0 | 0.58 | 1.05 (1.02 – 1.08) | 0.003 | 42.1 | 0.14 |
| 5p15.33 | A/G | rs2735940:5:1296486 | 0.511 | 0.92 (0.89 - 0.94) | $3.1 \times 10^{-13}$ | 0 | 0.59 | 0.93 (0.90 – 0.96) | $3.0 \times 10^{-6}$ | 0 | 0.75 |
| 5p13.1 | G/GT | rs58791712:5:40281797‡ | 0.745 | 0.91 (0.89 - 0.93) | $7.3 \times 10^{-14}$ | 56.7 | 0.13 | 0.90 (0.87 – 0.93) | $1.1 \times 10^{-9}$ | 28.4 | 0.24 |
| 6p21.31 | T/C | rs6906359:6:35528378‡ | 0.097 | 0.90 (0.86 - 0.93) | $3.4 \times 10^{-8}$ | 0 | 0.65 | 0.93 (0.89 – 0.98) | 0.005 | 0 | 0.55 |
| 6p12.1 | T/C | rs62404968:6:55714314 | 0.248 | 0.92 (0.89 - 0.94) | $8.6 \times 10^{-10}$ | 0 | 0.32 | 0.94 (0.90 – 0.97) | $3.8 \times 10^{-4}$ | 0 | 0.96 |
| 10q11.23 | T/C | rs10994860:10:52645424 | 0.202 | 0.92 (0.89 - 0.95) | $3.5 \times 10^{-8}$ | 0 | 0.35 | 0.96 (0.92 – 1.00) | 0.04 | 0 | 0.43 |
| 12q24.21 | CACAA/C | rs72013726:12:115890835‡ | 0.505 | 0.93 (0.90 - 0.95) | $5.0 \times 10^{-11}$ | 0 | 0.84 | 0.95 (0.92 – 0.98) | $9.1 \times 10^{-4}$ | 0 | 0.83 |
| 13q13.2 | C/G | rs10161980:13:34093518 | 0.620 | 1.08 (1.05 - 1.10) | $4.7 \times 10^{-9}$ | 0 | 0.81 | 1.03 (0.99 – 1.06) | 0.13 | 21.6 | 0.28 |
| 16q24.1 | C/G | rs2696839:16:86340448 | 0.495 | 0.94 (0.92 - 0.96) | $2.0 \times 10^{-8}$ | 75.6 | 0.04 | 0.96 (0.93 – 0.99) | 0.009 | 25.5 | 0.25 |
| 20q11.22 | T/C | rs2295444:20:33173883 | 0.495 | 0.93 (0.91 - 0.95) | $3.3 \times 10^{-9}$ | 0 | 0.97 | 0.97 (0.94 – 1.00) | 0.08 | 0 | 0.59 |
| 20q13.13 | T/C | rs1810502:20:49057488 | 0.449 | 0.93 (0.91 - 0.96) | $1.02 \times 10^{-8}$ | 0 | 0.98 | 0.94 (0.91 – 0.97) | $5.9 \times 10^{-5}$ | 11.8 | 0.34 |

*P values were derived from a fixed-effects inverse variance weighted meta-analysis. All tests were two-sided. Abbreviations: EFF = effect allele; REF = reference allele (reference category for the odds ratios); CHR = chromosome; BP = position; FRQ = frequency; 1KGP EUR = 1000 Genomes Europeans; OR = odds ratio; CI= confidence interval.† P values were derived from Cochran's *Q* test of heterogeneity. All tests were two-sided.
‡ Proxies were used in the independent replication stage ($r^2$ values from 1KGP Phase 3 Release 5): rs12520534 (chr5:40281761), $r^2$=1.0; rs144037597 (chr6: 35528204), $r^2$=1.0; rs12822984 (chr12:115888504), $r^2$=0.81.

**Table 2.** Multiethnic follow-up of 9 novel, independently-replicated low-penetrance risk variants

| Locus | EFF/REF allele | rsID:CHR:BP | FRQ_EFF (1KGP EAS) | FRQ_EFF (1KGP AMR) | FRQ_EFF (1KGP AFR) | East Asians (OncoArray, ACCC, US-Japan GWAS) | | | | Hispanic/Latinos (HCCS, MEC, SIGMA) | | African Americans (AA CRC GWAS) | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | OR (95%CI) | P* | $I^2$, % | $P_{heterogeneity}$ † | OR (95%CI) | P* | OR (95%CI) | P* |
| 4q22.2 | T/C | rs1370821:4:94943383 | 0.331 | 0.274 | 0.065 | 1.03 (0.99 - 1.08) | 0.13 | 49.6 | 0.14 | 1.17 (1.06 - 1.29) | 0.001 | 1.04 (0.92 - 1.17) | 0.54 |
| 5p15.33 | A/G | rs2735940:5:12964486 | 0.478 | 0.432 | 0.521 | 0.93 (0.87 - 1.00) | 0.03 | 61.2 | 0.11 | 0.99 (0.91 - 1.08) | 0.84 | 0.90 (0.83 - 0.98) | 0.01 |
| 5p13.1 | G/GT | rs58791712:5:40281797 | 0.956 | 0.765 | 0.924 | 0.87 (0.75 - 1.02) | 0.09 | 0 | 0.57 | 0.85 (0.77 - 0.94) | 0.001 | NA | NA |
| 6p21.31 | T/C | rs6906359:6:35528378 | 0.069 | 0.138 | 0.141 | 0.99 (0.91 - 1.07) | 0.73 | 0 | 0.45 | 0.82 (0.73 - 0.93) | 0.001 | 0.96 (0.84 - 1.08) | 0.47 |
| 6p12.1 | T/C | rs62404968:6:55714314 | 0.061 | 0.133 | 0.072 | 0.97 (0.88 - 1.05) | 0.44 | 60 | 0.08 | 1.03 (0.90 - 1.17) | 0.69 | 0.85 (0.74 - 0.97) | 0.02 |
| 10q11.23 | T/C | rs10994860:10:52645424 | 0.047 | 0.110 | 0.222 | 0.97 (0.89 - 1.06) | 0.47 | 51.2 | 0.13 | 1.00 (0.87 - 1.16) | 0.97 | 0.99 (0.90 - 1.09) | 0.87 |
| 12q24.21 | CACAA/C | rs72013726:12:115890835 | 0.643 | 0.633 | 0.360 | 0.92 (0.87 - 0.98) | 0.007 | 53.9 | 0.14 | 0.97 (0.89 - 1.06) | 0.53 | NA | NA |
| 16q24.1 | C/G | rs2696839:16:86340448 | 0.253 | 0.334 | 0.293 | 0.93 (0.89 - 0.98) | 0.004 | 45 | 0.18 | 0.90 (0.82 - 0.98) | 0.02 | 0.92 (0.84 - 1.00) | 0.06 |
| 20q13.13 | T/C | rs1810502: 20: 49057488 | 0.612 | 0.507 | 0.545 | 0.94 (0.90 - 0.98) | 0.007 | 49.2 | 0.16 | 0.92 (0.84 - 1.00) | 0.05 | 0.95 (0.88 - 1.03) | 0.24 |

*P values were derived from a fixed-effects inverse variance weighted meta-analysis. All tests were two-sided. Abbreviations: EFF = effect allele; REF = reference allele (reference category for the odds ratios); CHR = chromosome; BP = position; FRQ = frequency; 1KGP = 1000 Genomes; EAS = East Asian; AMR = Ad Mixed American; AFR = African; ACCC = Asia Colorectal Cancer Consortium; GWAS = genome-wide association study; OR = odds ratio; CI= confidence interval; HCCS = Hispanic Colorectal Cancer Study; MEC = Multiethnic Cohort; SIGMA = Slim Initiative in Genomic Medicine for the Americas; AA = African American; CRC = colorectal cancer.

† P values were derived from Cochran's *Q* test of heterogeneity. All tests were two-sided.

## Figure Legends

**Figure 1.** Manhattan plot summarizing the discovery GWAS association results. ($N_{case}$=36,948; $N_{control}$=30,864). Green = known risk loci (within 500 kilobases (kb) or $r^2$>0.2 with an index variant); red = novel risk loci (outside 500kb or $r^2$>0.2 with an index variant).