Deciphering colorectal cancer genetics through multi-omic analysis of 100,204 cases and 154,587 controls of European and east Asian ancestries

Fernandez-Rozadilla, Ceres; Timofeeva, Maria; Chen, Zhishan; Law, Philip; Thomas, Minta; Schmit, Stephanie; Díez-Obrero, Virginia; Hsu, Li; Fernandez-Tajes, Juan; Palles, Claire; Sherwood, Kitty; Briggs, Sarah; Svinti, Victoria; Donnelly, Kevin; Farrington, Susan; Blackmur, James; Vaughan-Shaw, Peter; Shu, Xiao ou; Long, Jirong; Cai, Qiuyin; Guo, Xingyi; Lu, Yingchang; Broderick, Peter; Studd, James; Huyghe, Jeroen; Harrison, Tabitha; Conti, David; Dampier, Christopher; Devall, Mathew; Schumacher, Fredrick; Melas, Marilena; Rennert, Gad; Obón-Santacana, Mireia; Martín-Sánchez, Vicente; Moratalla-Navarro, Ferran; Oh, Jae Hwan; Kim, Jeongseon; Jee, Sun Ha; Jung, Keum Ji; Kweon, Sun Seog; Shin, Min Ho; Shin, Aesun; Ahn, Yoon Ok; Kim, Dong Hyun; Oze, Isao; Wen, Wanqing; Matsuo, Keitaro; Matsuda, Koichi; Tanikawa, Chizu; Ren, Zefang; Gao, Yu Tang; Jia, Wei Hua; Hopper, John; Jenkins, Mark; Win, Aung Ko; Pai, Rish; Figueiredo, Jane; Haile, Robert; Gallinger, Steven; Woods, Michael; Newcomb, Polly; Duggan, David; Cheadle, Jeremy; Kaplan, Richard; Maughan, Timothy; Kerr, Rachel; Kerr, David; Kirac, Iva; Böhm, Jan; Mecklin, Lukka Pekka; Jousilahti, Pekka; Knekt, Paul; Aaltonen, Lauri; Rissanen, Harri; Pukkala, Eero; Eriksson, Johan; Cajuso, Tatiana; Hänninen, Ulrika; Kondelin, Johanna; Palin, Kimmo; Tanskanen, Tomas; Renkonen-Sinisalo, Laura; Zanke, Brent; Männistö, Satu; Albanes, Demetrius; Weinstein, Stephanie; Ruiz-Narvaez, Edward; Palmer, Julie; Buchanan, Daniel; Platz, Elizabeth; Visvanathan, Kala; Ulrich, Cornelia; Siegel, Erin; Brezina, Stefanie; Gsur, Andrea; Campbell, Peter; Chang-Claude, Jenny; Hoffmeister, Michael; Brenner, Hermann; Slattery, Martha; Potter, John; Tsilidis, Konstantinos; Schulze, Matthias; Gunter, Marc; Murphy, Neil; Castells, Antoni; Castellví-Bel, Sergi; Moreira, Leticia; Arndt, Volker; Shcherbina, Anna; Stern, Mariana; Pardamean, Bens; Bishop, Timothy; Giles, Graham; Southey, Melissa; Idos, Gregory; McDonnell, Kevin; Abu-Ful, Zomoroda; Greenson, Joel; Shulman, Katerina; Lejbkowicz, Flavio; Offit, Kenneth; Su, Yu Ru; Steinfelder, Robert; Keku, Temitope; van Guelpen, Bethany; Hudson, Thomas; Hampel, Heather; Pearlman, Rachel; Berndt, Sonja; Hayes, Richard; Martinez, Marie Elena; Thomas, Sushma; Corley, Douglas; Pharoah, Paul; Larsson, Susanna; Yen, Yun; Lenz, Heinz Josef; White, Emily; Li, Li; Doheny, Kimberly; Pugh, Elizabeth; Shelford, Tameka; Chan, Andrew; Cruz-Correa, Marcia; Lindblom, Annika; Hunter, David; Joshi, Amit; Schafmayer, Clemens; Scacheri, Peter; Kundaje, Anshul; Nickerson, Deborah; Schoen, Robert; Hampe, Jochen; Stadler, Zsofia; Vodicka, Pavel; Vodickova, Ludmila; Vymetalkova, Veronika; Papadopoulos, Nickolas; Edlund, Chistopher; Gauderman, William; Thomas, Duncan; Shibata, David; Toland, Amanda; Markowitz, Sanford; Kim, Andre; Chanock, Stephen; van Duijnhoven, Franzel; Feskens, Edith; Sakoda, Lori; Gago-Dominguez, Manuela; Wolk, Alicja; Naccarati, Alessio; Pardini, Barbara; FitzGerald, Liesel; Lee, Soo Chin; Ogino, Shuji; Bien, Stephanie; Kooperberg, Charles; Li, Christopher; Lin, Yi; Prentice, Ross; Qu, Conghui; Bézieau, Stéphane; Tangen, Catherine; Mardis, Elaine; Yamaji, Taiki; Sawada, Norie; Iwasaki, Motoki; Haiman, Christopher; Le Marchand, Loic; Wu, Anna; Qu, Chenxu; McNeil, Caroline; Coetzee, Gerhard; Hayward, Caroline; Deary, Ian; Harris, Sarah; Theodoratou, Evropi; Reid, Stuart; Walker, Marion; Ooi, Li Yin; Moreno, Victor; Casey, Graham; Gruber, Stephen; Tomlinson, Ian; Zheng, Wei; Dunlop, Malcolm; Houlston, Richard; Peters, Ulrike

Published in:
Nature Genetics

1. Supplementary Information:

A. Flat Files

Item	Present?	Filename This should be the name the file is saved as when it is uploaded to our system, and should include the file extension. The extension must be .pdf	A brief, numerical description of file contents. i.e.: *Supplementary Figures 1-4, Supplementary Discussion, and Supplementary Tables 1-4.*
Supplementary Information	Yes	Supplementary_Information.pdf	Supplementary Figures 1-5; Supplementary Note with references
Reporting Summary	Yes	Reporting Summary - Fernandez-Rozadilla.pdf	

B. Additional Supplementary Files

Type	Number If there are multiple files of the same type this should be the numerical indicator. i.e. "1" for Video 1, "2" for Video 2, etc.	Filename This should be the name the file is saved as when it is uploaded to our system, and should include the file extension. i.e.: *Smith_ Supplementary_Video_1.mov*	Legend or Descriptive Caption Describe the contents of the file
Supplementary Table	1-21	Supplementary_Tables.pdf	Legends and data for supplementary Tables 1-21

Deciphering colorectal cancer genetics through multi-omic analysis of 100,204 cases and 154,587 controls of European and East Asian ancestries

Ceres Fernandez-Rozadilla[1,2^], Maria Timofeeva[3,4^], Zhishan Chen[5^], Philip Law[6^], Minta Thomas[7^], Stephanie Schmit[8,9^], Virginia Díez-Obrero[10,11,12,13^], Li Hsu[7,14^], Juan Fernandez-Tajes[1], Claire Palles[15], Kitty Sherwood[1], Sarah Briggs[16], Victoria Svinti[3], Kevin Donnelly[3], Susan Farrington[3], James Blackmur[3], Peter Vaughan-Shaw[3], Xiao-ou Shu[5], Jirong Long[5], Qiuyin Cai[5], Xingyi Guo[5,17], Yingchang Lu[5], Peter Broderick[6], James Studd[6], Jeroen Huyghe[7], Tabitha Harrison[7], David Conti[18], Christopher Dampier[19], Mathew Devall[19], Fredrick Schumacher[20,21], Marilena Melas[22], Gad Rennert[23,24,25], Mireia Obón-Santacana[11,10,26], Vicente Martín-Sánchez[12,27], Ferran Moratalla-

Navarro[10,11,12,13], Jae Hwan Oh[28], Jeongseon Kim[29], Sun Ha Jee[30], Keum Ji Jung[30], Sun-Seog Kweon[31], Min-Ho Shin[31], Aesun Shin[32,33], Yoon-Ok Ahn[32], Dong-Hyun Kim[34], Isao Oze[35], Wanqing Wen[5], Keitaro Matsuo[36,37], Koichi Matsuda[38], Chizu Tanikawa[39], Zefang Ren[40], Yu-Tang Gao[41], Wei-Hua Jia[42], John Hopper[43,44], Mark Jenkins[43], Aung Ko Win[43], Rish Pai[45], Jane Figueiredo[46,18], Robert Haile[47], Steven Gallinger[48], Michael Woods[49], Polly Newcomb[7,50], David Duggan[51], Jeremy Cheadle[52], Richard Kaplan[53], Timothy Maughan[54], Rachel Kerr[55], David Kerr[56], Iva Kirac[57], Jan Böhm[58], Lukka-Pekka Mecklin[59], Pekka Jousilahti[60], Paul Knekt[60], Lauri Aaltonen[61,62], Harri Rissanen[63], Eero Pukkala[64,65], Johan Eriksson[66,67,68], Tatiana Cajuso[62,61], Ulrika Hänninen[62,61], Johanna Kondelin[62,61], Kimmo Palin[62,61], Tomas Tanskanen[62,61], Laura Renkonen-Sinisalo[69], Brent Zanke[70], Satu Männistö[63], Demetrius Albanes[71], Stephanie Weinstein[71], Edward Ruiz-Narvaez[72], Julie Palmer[73,74], Daniel Buchanan[75,76,77], Elizabeth Platz[78], Kala Visvanathan[78], Cornelia Ulrich[79], Erin Siegel[80], Stefanie Brezina[81], Andrea Gsur[81], Peter Campbell[82], Jenny Chang-Claude[83,84], Michael Hoffmeister[85], Hermann Brenner[85,86,87], Martha Slattery[88], John Potter[89,7], Konstantinos Tsilidis[90,91], Matthias Schulze[92,93], Marc Gunter[94], Neil Murphy[94], Antoni Castells[95], Sergi Castellví-Bel[95], Leticia Moreira[95], Volker Arndt[85], Anna Shcherbina[96], Mariana Stern[97,98], Bens Pardamean[99], Timothy Bishop[100], Graham Giles[101,43,102], Melissa Southey[102,103,101], Gregory Idos[104], Kevin McDonnell[104,24,25], Zomoroda Abu-Ful[23], Joel Greenson[105,24,25], Katerina Shulman[23], Flavio Lejbkowicz[106,23,25], Kenneth Offit[107,108], Yu-Ru Su[109], Robert Steinfelder[7], Temitope Keku[110], Bethany van Guelpen[111,112], Thomas Hudson[113], Heather Hampel[114], Rachel Pearlman[114], Sonja Berndt[71], Richard Hayes[115], Marie Elena Martinez[116,117], Sushma Thomas[118], Douglas Corley[119,120], Paul Pharoah[121], Susanna Larsson[122], Yun Yen[123], Heinz-Josef Lenz[124], Emily White[7,125], Li Li[21], Kimberly Doheny[126], Elizabeth Pugh[126], Tameka Shelford[126], Andrew Chan[127,128,129,130,131,132], Marcia Cruz-Correa[133], Annika Lindblom[134,135], David Hunter[131,136], Amit Joshi[131,127], Clemens Schafmayer[137], Peter Scacheri[138], Anshul Kundaje[96,139], Deborah Nickerson[140], Robert Schoen[141], Jochen Hampe[142], Zsofia Stadler[143,108], Pavel Vodicka[144,145,146], Ludmila Vodickova[144,145,146], Veronika Vymetalkova[144,145,146], Nickolas Papadopoulos[147,148,149], Chistopher Edlund[18], William Gauderman[18], Duncan Thomas[18], David Shibata[150], Amanda Toland[151], Sanford Markowitz[152], Andre Kim[18], Stephen Chanock[71], Franzel van Duijnhoven[153], Edith Feskens[154], Lori Sakoda[119,7], Manuela Gago-Dominguez[155,156], Alicja Wolk[122], Alessio Naccarati[157,158], Barbara Pardini[157,158],

Liesel FitzGerald[159,101], Soo Chin Lee[160], Shuji Ogino[161,162,131,163], Stephanie Bien[7], Charles Kooperberg[7], Christopher Li[7], Yi Lin[7], Ross Prentice[7,164], Conghui Qu[7], Stéphane Bézieau[165], Catherine Tangen[166], Elaine Mardis[167], Taiki Yamaji[168], Norie Sawada[169], Motoki Iwasaki[168,169], Christopher Haiman[170], Loic Le Marchand[171], Anna Wu[172], Chenxu Qu[173], Caroline McNeil[173], Gerhard Coetzee[174], Caroline Hayward[175], Ian Deary[176], Sarah Harris[177], Evropi Theodoratou[178], Stuart Reid[3], Marion Walker[3], Li Yin Ooi[179,3], Victor Moreno[10,11,12,13*], Graham Casey[19*], Stephen Gruber[104*], Ian Tomlinson[1,15*], Wei Zheng[5*], Malcolm Dunlop[3*], Richard Houlston[6*], Ulrike Peters[7,180*]

[1]Edinburgh Cancer Research Centre, Institute of Genomics and Cancer, University of Edinburgh, Edinburgh, United Kingdom, [2]Genomic Medicine Group, Instituto de Investigacion Sanitaria de Santiago (IDIS), Santiago de Compostela, Spain, [3]Colon Cancer Genetics Group, Medical Research Council Human Genetics Unit, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, United Kingdom, [4]Danish Institute for Advanced Study (DIAS), Department of Public Health, University of Southern Denmark, Odense, Denmark, [5]Division of Epidemiology, Department of Medicine, Vanderbilt-Ingram Cancer Center, Vanderbilt Epidemiology Center, Vanderbilt University Medical Center, Nashville, USA, [6]Division of Genetics and Epidemiology, The Institute of Cancer Research, London, United Kingdom, [7]Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, USA, [8]Genomic Medicine Institute, Cleveland Clinic, Cleveland, USA, [9]Population and Cancer Prevention Program, Case Comprehensive Cancer Center, Cleveland, USA, [10]Colorectal Cancer Group, ONCOBELL Program, Bellvitge Biomedical Research Institute (IDIBELL), Barcelona, Spain, [11]Oncology Data Analytics Program, Catalan Institute of Oncology (ICO), Barcelona, Spain, [12]Consortium for Biomedical Research in Epidemiology and Public Health (CIBERESP). Madrid, Madrid, Spain, [13]Department of Clinical Sciences, Faculty of Medicine, University of Barcelona, Barcelona, Spain, [14]Department of Biostatistics, School of Public Health, University of Washington, Seattle, USA, [15]Institute of Cancer and Genomic Sciences, College of Medical and Dental Sciences, University of Birmingham, Birmingham, United Kingdom, [16]Department of Public Health, Richard Doll Building, University of Oxford, Oxford, United Kingdom, [17]Department of Biomedical Informatics, Vanderbilt University

School of Medicine, Nashville, USA, [18]Department of Preventive Medicine, USC Norris Comprehensive Cancer Center, Keck School of Medicine, University of Southern California, Los Angeles, USA, [19]Center for Public Health Genomics, Department of Public Health Sciences, University of Virginia, Charlottesville, USA, [20]Department of Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, USA, [21]Case Comprehensive Cancer Center, Case Western Reserve University, Cleveland, USA, [22]The Steve and Cindy Rasmussen Institute for Genomic Medicine, Nationwide Children's Hospital, Columbus, USA, [23]Department of Community Medicine and Epidemiology, Lady Davis Carmel Medical Center, Haifa, Israel, [24]Ruth and Bruce Rappaport Faculty of Medicine, Technion-Israel Institute of Technology, Haifa, Israel, [25]Clalit National Cancer Control Center, Haifa, Israel, [26]Consortium for Biomedical Research in Epidemiology and Public Health (CIBERESP), Madrid, Spain, [27]Biomedicine Institute (IBIOMED), University of León, León, Spain, [28]Center for Colorectal Cancer, National Cancer Center Hospital, National Cancer Center, Gyeonggi-do, South Korea, [29]Department of Cancer Biomedical Science, Graduate School of Cancer Science and Policy, National Cancer Center, Gyeonggi-do, South Korea, [30]Department of Epidemiology and Health Promotion, Graduate School of Public Health, Yonsei University, Seoul, South Korea, [31]Department of Preventive Medicine, Chonnam National University Medical School, Gwangju, South Korea, [32]Department of Preventive Medicine, Seoul National University College of Medicine, Seoul, South Korea, [33]Cancer Research Institute, Seoul National University, Seoul, South Korea, [34]Department of Social and Preventive Medicine, Hallym University College of Medicine, Okcheon-dong, South Korea, [35]Division of Cancer Epidemiology and Prevention, Aichi Cancer Center Research Institute, Nagoya, Japan, [36]Division of Molecular and Clinical Epidemiology, Aichi Cancer Center Research Institute, Nagoya, Japan, [37]Department of Epidemiology, Nagoya University Graduate School of Medicine, Nagoya, Japan, [38]Laboratory of Clinical Genome Sequencing, Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, University of Tokyo, Tokyo, Japan, [39]Laboratory of Genome Technology, Human Genome Center, Institute of Medical Science, University of Tokyo, Tokyo, Japan, [40]School of Public Health, Sun Yat-sen University, Guangzhou, China, [41]State Key Laboratory of Oncogenes and Related Genes and Department of Epidemiology, Shanghai Cancer Institute, Renji Hospital, Shanghai Jiaotong University School of Medicine, Shanghai, China,

[42]State Key Laboratory of Oncology in South China, Cancer Center, Sun Yat-sen University, Guangzhou, China, [43]Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Melbourne, Australia, [44]Department of Epidemiology, School of Public Health and Institute of Health and Environment, Seoul National University, Seoul, South Korea, [45]Department of Laboratory Medicine and Pathology, Mayo Clinic Arizona, Scottsdale, USA, [46]Department of Medicine, Samuel Oschin Comprehensive Cancer Institute, Cedars-Sinai Medical Center, Los Angeles, USA, [47]Division of Oncology, Department of Medicine, Cedars-Sinai Cancer Research Center for Health Equity, Los Angeles, USA, [48]Lunenfeld Tanenbaum Research Institute, Mount Sinai Hospital, University of Toronto, Toronto, Canada, [49]Memorial University of Newfoundland, Division of Biomedical Sciences, St. John's, Canada, [50]School of Public Health, University of Washington, Seattle, USA, [51]Translational Genomics Research Institute, City of Hope National Medical Center, Phoenix, USA, [52]Institute of Medical Genetics, Cardiff University, Cardiff, United Kingdom, [53]MRC Clinical Trials Unit, Medical Research Council, United Kingdom, [54]MRC Institute for Radiation Oncology, University of Oxford, Oxford, United Kingdom, [55]Department of Oncology, University of Oxford, Oxford, United Kingdom, [56]Radcliffe Department of Medicine, University of Oxford, Oxford, United Kingdom, [57]Department of Surgical Oncology, University Hospital for Tumors, Sestre milosrdnice University Hospital Center, Zagreb, Croatia, [58]Department of Pathology, Central Finland Health Care District, Jyväskylä, Finland, [59]Central Finland Health Care District, Central Finland Health Care District, Jyväskylä, Finland, [60]Department of Health and Welfare, Finnish Institute for Health and Welfare, Helsinki, Finland, [61]Department of Medical and Clinical Genetics, University of Helsinki, Helsinki, Finland, [62]Genome-Scale Biology Research Program, University of Helsinki, Helsinki, Finland, [63]Department of Public Health and Welfare, Finnish Institute for Health and Welfare, Helsinki, Finland, [64]Finnish Cancer Registry, Institute for Statistical and Epidemiological Cancer Research, Helsinki, Finland, [65]Faculty of Social Sciences, Tampere University, Tampere, Finland, [66]Folkhälsan Research Centre, University of Helsinki, Helsinki, Finland, [67]National University of Singapore, Human Potential Translational Research Programme, Singapore, Singapore, [68]Unit of General Practice and Primary Health Care, University of Helsinki and Helsinki University Hospital, Helsinki, Finland, [69]Department of Surgery, Abdominal Centre, Helsinki University Hospital, Helsinki,

Finland, [70]University of Toronto, Department of Oncology, Toronto, Canada, [71]Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, USA, [72]Department of Nutritional Sciences, School of Public Health, University of Michigan, Ann Arbor, USA, [73]Slone Epidemiology Center at Boston University, Boston, Massachusetts, USA, [74]Department of Medicine, Boston University School of Medicine, Boston, USA, [75]Colorectal Oncogenomics Group, Department of Clinical Pathology, The University of Melbourne, Parkville, Australia, [76]University of Melbourne Centre for Cancer Research, Victorian Comprehensive Cancer Centre, Parkville, Australia, [77]Genomic Medicine and Family Cancer Clinic, The Royal Melbourne Hospital, Parkville, Australia, [78]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, USA, [79]Huntsman Cancer Institute and Department of Population Health Sciences, University of Utah, Salt Lake City, USA, [80]Cancer Epidemiology Program, H. Lee Moffitt Cancer Center and Research Institute, Tampa, USA, [81]Institute of Cancer Research, Department of Medicine I, Medical University Vienna, Vienna, Austria, [82]Department of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, USA, [83]Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany, [84]University Medical Centre Hamburg-Eppendorf, University Cancer Centre Hamburg (UCCH), Hamburg, Germany, [85]Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), Heidelberg, Germany, [86]Division of Preventive Oncology, German Cancer Research Center (DKFZ) and National Center for Tumor Diseases (NCT), Heidelberg, Germany, [87]German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg, Germany, [88]Department of Internal Medicine, University of Utah, Salt Lake City, USA, [89]Research Centre for Hauora and Health, Massey University, Wellington, New Zealand, [90]Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece, [91]Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, United Kingdom, [92]Department of Molecular Epidemiology, German Institute of Human Nutrition Potsdam-Rehbruecke, Nuthetal, Germany, [93]Institute of Nutritional Science, University of Potsdam, Potsdam, Germany, [94]Nutrition and Metabolism Branch, International Agency for Research on Cancer, World Health Organization, Lyon, France, [95]Gastroenterology Department, Hospital Clínic, Institut d'Investigacions

Biomèdiques August Pi i Sunyer (IDIBAPS), Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBEREHD), University of Barcelona, Barcelona, Spain, [96]Department of Genetics, Stanford University, Stanford, USA, [97]Department of Population and Public Health Sciences, USC Norris Comprehensive Cancer Center, Keck School of Medicine, University of Southern California, Los Angeles, USA, [98]Jeonnam Regional Cancer Center, Chonnam National University Hwasun Hospital, Hwasun, South Korea, [99]Bioinformatics and Data Science Research Center, Bina Nusantara University, Jakarta, Indonesia, [100]Leeds Institute of Medical Research at St. James's, University of Leeds, Leeds, United Kingdom, [101]Cancer Epidemiology Division, Cancer Council Victoria, Melbourne, Australia, [102]Precision Medicine, School of Clinical Sciences at Monash Health, Monash University, Clayton, Australia, [103]Department of Clinical Pathology, The University of Melbourne, Victoria, Australia, [104]Department of Medical Oncology and Center For Precision Medicine, City of Hope National Medical Center, USA, [105]Department of Pathology, University of Michigan, Ann Arbor, USA, [106]The Clalit Health Services, Personalized Genomic Service, Lady Davis Carmel Medical Center, Haifa, Israel, [107]Clinical Genetics Service, Department of Medicine, Memorial Sloan-Kettering Cancer Center, New York, USA, [108]Department of Medicine, Weill Cornell Medical College, New York, USA, [109]Kaiser Permanente Washington Health Research Institute, Seattle, USA, [110]Center for Gastrointestinal Biology and Disease, University of North Carolina, Chapel Hill, USA, [111]Department of Radiation Sciences, Oncology Unit, Umeå University, Umeå, Sweden, [112]Wallenberg Centre for Molecular Medicine, Umeå University, Umeå, Sweden, [113]Ontario Institute for Cancer Research, Toronto, Canada, [114]Division of Human Genetics, Department of Internal Medicine, The Ohio State University Comprehensive Cancer Center, Columbus, USA, [115]Division of Epidemiology, Department of Population Health, New York University School of Medicine, New York, USA, [116]Population Sciences, Disparities and Community Engagement, University of California San Diego Moores Cancer Center, La Jolla, USA, [117]Department of Family Medicine and Public Health, University of California San Diego, La Jolla, USA, [118]Fred Hutchinson Cancer Research Center, Seattle, USA, [119]Division of Research, Kaiser Permanente Northern California, Oakland, USA, [120]Department of Gastroenterology, Kaiser Permanente Medical Center, San Francisco, USA, [121]Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom,

[122]Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden, [123]Taipei Medical University, Taipei, Taiwan, [124]Department of Medicine, Keck School of Medicine, University of Southern California, Los Angeles, USA, [125]Department of Epidemiology, University of Washington School of Public Health, Seattle, USA, [126]Center for Inherited Disease Research (CIDR), Department of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, USA, [127]Clinical and Translational Epidemiology Unit, Massachusetts General Hospital and Harvard Medical School, Boston, USA, [128]Division of Gastroenterology, Massachusetts General Hospital and Harvard Medical School, Boston, USA, [129]Channing Division of Network Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, USA, [130]Broad Institute of Harvard and MIT, Cambridge, USA, [131]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Harvard University, Boston, USA, [132]Department of Immunology and Infectious Diseases, Harvard T.H. Chan School of Public Health, Harvard University, Boston, USA, [133]Comprehensive Cancer Center, University of Puerto Rico, San Juan, Puerto Rico, [134]Department of Clinical Genetics, Karolinska University Hospital, Stockholm, Sweden, [135]Department of Molecular Medicine and Surgery, Karolinska Institutet, Stockholm, Sweden, [136]Nuffield Department of Population Health, University of Oxford, Oxford, United Kingdom, [137]Department of General Surgery, University Hospital Rostock, Rostock, Germany, [138]Department of Genetics and Genome Sciences, Case Western Reserve University, Cleveland, USA, [139]Department of Computer Science, Stanford University, Stanford, USA, [140]Department of Genome Sciences, University of Washington, Seattle, USA, [141]Department of Medicine and Epidemiology, University of Pittsburgh Medical Center, Pittsburgh, USA, [142]Department of Medicine I, University Hospital Dresden, Technische Universität Dresden (TU Dresden), Dresden, Germany, [143]Department of Medicine, Memorial Sloan-Kettering Cancer Center, New York, USA, [144]Department of Molecular Biology of Cancer, Institute of Experimental Medicine of the Czech Academy of Sciences, Prague, Czech Republic, [145]Institute of Biology and Medical Genetics, First Faculty of Medicine, Charles University, Prague, Czech Republic, [146]Faculty of Medicine and Biomedical Center in Pilsen, Charles University, Pilsen, Czech Republic, [147]Department of Oncology Ludwig Center at the Sidney Kimmel Cancer Center, Johns Hopkins University School of Medicine, Baltimore, USA, [148]Sol Goldman Pancreatic Cancer Research Center, Johns Hopkins University School of Medicine,

Baltimore, USA, [149]Department of Pathology, Johns Hopkins School of Medicine, Baltimore, USA, [150]Department of Surgery, University of Tennessee Health Science Center, Memphis, USA, [151]Departments of Cancer Biology and Genetics and Internal Medicine, Comprehensive Cancer Center, The Ohio State University, Columbus, USA, [152]Departments of Medicine and Genetics, Case Comprehensive Cancer Center, Case Western Reserve University and University Hospitals of Cleveland, Cleveland, USA, [153]Division of Human Nutrition and Health, Wageningen University & Research, Wageningen, The Netherlands, [154]Division of Human Nutrition, Wageningen University and Research, Wageningen, The Netherlands, [155]Genomic Medicine Group, Galician Public Foundation of Genomic Medicine, Servicio Galego de Saude (SERGAS), Santiago de Compostela, Spain, [156]Instituto de Investigación Sanitaria de Santiago de Compostela (IDIS), Santiago de Compostela, Spain, [157]Italian Institute for Genomic Medicine (IIGM), Candiolo Cancer Institute - FPO-IRCCS, Turin, Italy, [158]Candiolo Cancer Institute - FPO-IRCCS, Candiolo, Italy, [159]Menzies Institute for Medical Research, University of Tasmania, Hobart, Australia, [160]National University Cancer Institute, Singapore; Cancer Science Institute of Singapore, National University of Singapore, Singapore, Singapore, [161]Department of Pathology, Brigham and Women's Hospital, Harvard Medical School, Boston, USA, [162]Cancer Immunology Program, Dana-Farber Harvard Cancer Center, Boston, USA, [163]Broad Institute of MIT and Harvard, Cambridge, USA, [164]Department of Biostatistics, University of Washington, Seattle, USA, [165]Service de Génétique Médicale, Centre Hospitalier Universitaire (CHU) Nantes, Nantes, France, [166]SWOG Statistical Center, Fred Hutchinson Cancer Research Center, Seattle, USA, [167]Department of Pediatrics, Nationwide Children's Hospital, The Steve and Cindy Rasmussen Institute for Genomic Medicine, Columbus, USA, [168]Division of Epidemiology, National Cancer Center Institute for Cancer Control, National Cancer Center, Tokyo, Japan, [169]Division of Cohort Research, National Cancer Center Institute for Cancer Control, National Cancer Center, Tokyo, Japan, [170]Department of Preventive Medicine, Center for Genetic Epidemiology, University of Southern California, Los Angeles, USA, [171]Cancer Center, University of Hawaii, Honolulu, USA, [172]Preventative Medicine, University of Southern California, Los Angeles, USA, [173]USC Norris Comprehensive Cancer Center, Keck School of Medicine, University of Southern California, Los Angeles, USA, [174]Van Andel Research Institute, Grand Rapids, USA, [175]MRC Human Genetics Unit, Institute of Genomics and Cancer, University

of Edinburgh, Edinburgh, United Kingdom, [176]Lothian Birth Cohorts group, Department of Psychology, University of Edinburgh, Edinburgh, United Kingdom, [177]Lothian Birth Cohorts group, Department of Psychology, University of Edinburgh, Edinburgh EH8 9JZ, United Kingdom, [178]Centre for Global Health, Usher Institute, University of Edinburgh, Edinburgh, United Kingdom, [179]Department of Pathology, National University Hospital, National University Health System, Singapore, Singapore, [180]Department of Epidemiology, University of Washington, Seattle, USA

^These authors contributed equally

*These authors jointly supervised this work

Corresponding authors: Ian PM Tomlinson (Ian.Tomlinson@ed.ac.uk); Wei Zheng (wei.zheng@vumc.org); Malcolm G Dunlop (malcolm.dunlop@ed.ac.uk); Richard S Houlston (Richard.Houlston@icr.ac.uk); Ulrike Peters (upeters@fredhutch.org).

ABSTRACT

Colorectal cancer (CRC) is a leading cause of mortality worldwide. We conducted a genome-wide association study meta-analysis of 100,204 CRC cases and 154,587 controls of European and East Asian ancestry, identifying 205 independent risk associations, of which 50 were unreported. We performed integrative genomic, transcriptomic and methylomic analyses across large bowel mucosa and other tissues. Transcriptome- and methylome-wide association studies revealed an additional 53 risk associations. We identified 155 high confidence effector genes functionally linked to CRC risk, many of which had no previously established role in CRC. These have multiple different functions, and specifically indicate that variation in normal colorectal homeostasis, proliferation, cell adhesion, migration, immunity and microbial interactions determines CRC risk. Cross-tissue analyses indicated that over a third of effector genes most likely act outside the colonic mucosa. Our findings provide insights into colorectal oncogenesis, and highlight potential targets across tissues for new CRC treatment and chemoprevention strategies.

INTRODUCTION

Colorectal cancer (CRC), which affects approximately 1.9 million people worldwide annually[1], has a strong heritable basis[2]. Our understanding of CRC genetics has been informed by genome-wide association studies (GWAS), which have so far identified 150 statistically independent risk variants[3,4]. To provide a comprehensive description of CRC genetics, we brought together the great majority of GWAS performed to date. We complemented GWAS with transcriptome- and methylome-wide association analyses (TWAS and MWAS; **Fig. 1**). Through integration of these data, we investigated the genes and mechanisms underlying established and novel CRC risk loci. We identified credible effector genes and the tissues in which they act, informing our understanding of colorectal tumorigenesis.

RESULTS

Genetic architecture of colorectal cancer

We performed a meta-analysis of CRC GWAS data sets, comprising 100,204 CRC cases and 154,587 controls (73% European and 27% East Asian ancestry) (**Supplementary Tables 1 & 2**). We identified 205 associations, including 37 single-nucleotide polymorphisms (SNPs) at novel loci (sentinel risk SNPs > 1 megabase (Mb) from another significant SNP), 13 independent novel risk SNPs in conditional analysis (**Table 1**)**,** and 155 previously reported SNPs or proxies **Table 1, Supplementary Tables 3-4, Supplementary figures 1 & 2**). There was limited heterogeneity ascribable to population effects (**Supplementary Table 2, Supplementary figure 3**), although four risk variants (rs12078075, rs57939401, rs151127921 and rs5751474) were monomorphic in East Asian participants (Table 1).

Using linkage-disequilibrium (LD) score regression (LD hub), we estimated the heritability of CRC attributable to all common genetic variants to be similar in Europeans (h^2 0.11, s.d. 0.008) and East Asians (h^2 0.09, s.d. 0.006), which translates to 73% of familial CRC risk. Restricting estimates to the 205 GWAS-significant SNPs explained 19.7% of this familial risk. We evaluated the performance of a polygenic risk score (PRS) based on these SNPs in two cohorts independent of the GWAS discovery samples[7,8]. For Europeans and East Asians, individuals in the top PRS decile exhibited odds ratios of 2.22 (95%CI: 1.92-2.57; P = 1.80 x 10^{-26}) and 1.96 (95%CI: 1.64-2.34; P = 8.9 x 10^{-14}) compared to the remaining individuals. Corresponding areas under the receiver operating characteristic curve (AUC) were 0.62 (95%CI: 0.60-0.63) and 0.60 (95%CI: 0.59-0.62).

Discovery of risk loci by TWAS and MWAS

TWAS was performed by implementing the PredictDB pipeline using mRNA expression data from 1,107 colorectal mucosa samples as reference (709 in house, 368 GTEx transverse colon) [9,10]. In addition to associations identified by GWAS or those previously reported by TWAS (*PYGL* and *TRIM4* [11,12]), we identified 15 novel associations at Bonferroni-corrected significance ($P_{\text{Bonferroni}}$, **Table 2, Supplementary Tables 5 & 6, Supplementary figure 4**). We extended the main TWAS to a transcript isoform-wide association study (TIsWAS), both to ascertain whether specific transcripts could account for TWAS associations and to identify previously unreported risk associations (**Supplementary Tables 7 & 8**). For a third of TWAS genes, a significant association with CRC risk was found for a single mRNA isoform (**Supplementary Table 7**). The TIsWAS also identified eight loci associated with CRC risk (**Table 3**). To improve power for discovery, and because some CRC risk SNPs may not exert their effects in colorectal mucosa, we also conducted a cross-tissue TWAS using our in-house RNA sequencing (RNAseq) data and the full GTEx and Depression Genes and Networks (DGN) project data (49 tissues)[13]. We identified a further 23 risk associations (**Table 4, Supplementary Tables 9-13**).

To complement the TWAS, identify further CRC risk loci and gain mechanistic insights, we extended the PredictDB pipeline to perform MWAS based on quantitative methylation data from histologically normal colorectal mucosa (**Supplementary Methods**). We found significant associations between CRC risk and methylation of individual CpGs at 69 loci (**Supplementary Tables 14 & 15**). This included seven novel independent risk loci (**Table 5**). Risk SNPs may influence CRC risk through changes in the CpG methylation status of regulatory elements leading to changes in gene expression. We therefore explored the relationship between gene expression, CpG methylation and CRC risk in colorectal mucosa for 6,722 genes with both TWAS and MWAS predictions. There was a strong tendency for genes to be represented in both TWAS and MWAS ($P < 10^{-7}$, Fisher's exact test). Subsequently, we conditioned TWAS associations on the top MWAS-significant CpG within 1Mb, finding that 67/91 (75%) genes did not retain a significant TWAS association ($P_{Bonferroni} > 5.50 \times 10^{-4}$; **Supplementary Table 16**). Our data are consistent with a model in which many CRC risk SNPs act through changes in DNA methylation, although formal causality analysis could not be performed to exclude reverse causation or possible confounders.

Effector genes and biological pathways of CRC oncogenesis

A major, largely unfulfilled aim of cancer GWAS is to identify genes and functional mechanisms that may ultimately be clinically useful targets, for example in chemoprevention. The large GWAS and TWAS datasets in this study address this aim by enabling a detailed functional analysis of the molecular mechanisms contributing to CRC risk. Since TWAS approaches do not identify causal genes directly, we used our data to compile a set of 155 credible effector genes from the independent associations identified through GWAS, TWAS, TIsWAS and MWAS (details in **Supplementary Table 17** and **Supplementary Methods**).

We identified molecular pathways enriched in effector genes using Enrichr (https://maayanlab.cloud/Enrichr/) (**Supplementary Table 18**). This analysis was complemented with DEPICT based on the GWAS SNPs (https://data.broadinstitute.org/mpg/depict/)

(**Supplementary Table 19)**. CRC effectors were principally enriched in genes regulating TGF-β/BMP, Wnt WNT and Hippo pathways. A number of the credible effector genes that map to these pathways have no established role in CRC, including the intestinal stem cell regulator *ZNRF3*[14], the TGF repressor *LEMD3*[15], and the EMT regulator *RREB1*[16].

To complement the pathway analysis, we performed gene-level functional annotation based on the principal cellular function of each effector gene as reported in the literature (**Figure 2, Supplementary Table 20**). Thirty-six genes (mostly Wnt and BMP family members) were annotated to colorectal homeostasis (i.e. cellular stemness/differentiation). Intriguingly, 16 genes (including *ARHGEF19, ARHGEF4, GNA12, RHOG, TAGLN, TSPAN8, STARD13* and *LLGL1*) were linked to cell migration through RhoA/ROCK signaling. We found eight genes (*SPSB1, PIK3C2B, DUSP1, LRIG1, GAB1, RREB1, MAPKAPK5-AS1* and *PDGFB*) to act within the Ras/Raf growth factor signaling pathway. In addition to the previously reported association at *FUT2*, the novel fucosyltransferase effector genes *FUT3* and *FUT6* supported a relationship between the gut microbiome and CRC risk[17]. Inflammation is important in CRC[18], and the TWAS association at the FADS gene cluster and *PTGES3*, specifically highlighted the role of prostaglandin metabolism in CRC risk. Finally, our data also indicated several effector genes with roles in ion transport and cytoskeletal components (**Fig. 2, Supplementary Table 20**).

Although our pathway analysis and functional annotation indicated that the colorectum was the likely target tissue of many effector genes (**Supplementary Tables 19 & 20**), some genes were associated with principal roles in other tissue types, for example neuronal cells (*LINGO4, TULP1* and *CNIH2*) and leukocytes (*TOX*, *TOX4* and *MAF,* plus many candidate genes within the MHC region) (**Supplementary Table 20**). We therefore performed a systematic analysis of effector gene tissue specificity, based on the premise that TWAS associations tend to be present in tissues in which a gene functionally affects CRC risk. Cross-tissue analysis showed that all but one effector gene exhibited a TWAS association ($FDR_{TWAS} < 0.05$) in at least one tissue and 52 (34%) genes showed an association in multiple tissues (**Supplementary figure 5**). For 26 (17%) genes, associations were confined to the colorectal mucosa (P_{TWAS} Bonferroni-significant in mucosa,

P_{TWAS} > FDR elsewhere). In contrast, 67 genes (43%) showed no evidence of a TWAS association in colorectal mucosa (FDR_{TWAS} > 0.05). Notably, 12 (8%) gene associations were present only in immune cells (**Supplementary figure 5, Supplementary Table 11**) and four (3%) were restricted to mesenchymal cells (**Supplementary figure 5, Supplementary Table 12**).

Linking colorectal cancer risk to other traits

To gain insight into the role of potentially modifiable risk factors in CRC genetics, we performed cross-trait LD score regression analyses[19] using publicly available GWAS summary statistics for 171 phenotypes. Twelve genetic correlations remained significant (two-sided Z-test, Bonferroni-corrected $P < 2.93 \times 10^{-4}$). Notably, positive associations with CRC risk (**Supplementary Table 21**) included insulin resistance (raised fasting insulin and glucose), smoking, and obesity (body mass index - BMI, waist-to-hip ratio - WHR, waist circumference), traits that have previously been reported in observational epidemiological studies to be associated with CRC risk[3,20,21]. These associations not only highlight shared biology, but also suggest that public health interventions to reduce cardiometabolic disease will additionally lower CRC burden.

DISCUSSION

We report a comprehensive genetic analysis of CRC risk in the general population. To identify the most credible effector genes for each risk variant, we performed detailed annotation using tissue-specific gene expression and other relevant data types. Our study is twice as large as previous CRC GWAS, and also includes participants of both European and East Asian ancestries, demonstrating that most loci are shared across these ancestral groups. This increased power for GWAS, coupled with complementary analyses, including TWAS and MWAS, identified 103 previously unreported risk associations and identified 155 effector genes. These data substantially expand our existing knowledge regarding the impact of common genetic variation on the heritable risk of CRC.

The availability of large, multi-omic data sets has allowed us to assign the most likely target/effector genes of GWAS and TWAS associations (**Fig. 3**), and confidence in these assignments will increase as additional functional data are reported in the literature. It is clear that pathways (*e.g.,* Wnt , BMP, Hippo) involved in normal intestinal homeostasis play important roles in CRC risk, suggesting that modulation of normal mucosal dynamics has the potential to prevent colorectal neoplasia. The gut flora is intimately involved in normal bowel homeostasis, and effector genes are likely to be involved in microbial interactions. By contrast, Ras pathway activity is thought to be more important during repair or tumorigenesis, and the Ras effector genes we have found may act after tumor initiation. Our finding of multiple risk genes involved in cell adhesion and migration naturally suggests roles in malignant progression, although effects earlier in tumorigenesis also remain plausible. Similarly, immune pathway effector genes could, in principle, have their effects on normal cell function or at any stage of tumorigenesis, from mediating day-to-day microbial interactions to killing of cells in early neoplastic transformation or established tumors.

Cross-tissue analyses indicated that the colorectal mucosa was the most likely site of action of many effector genes, but some genes are more likely to act in different tissue types. For example, it is highly likely that genes such as *HIVEP1, LIF, SH2B3, TOX* and *TOX4* (and probably genes in the MHC region) influence the development of CRC through immune cell variation, and that *EDNRB* influences risk through effects on blood vessels. An unexpected finding was that several credible effector genes have primary roles in neurogenesis, raising the intriguing possibility that the enteric nervous system is involved in CRC risk.

While germline genetics has guided the development of drugs to prevent cardiovascular disease (*e.g.* statins and PCSK9 inhibitors), such a paradigm has yet to be realized for cancer. Since almost all CRCs develop from colonic polyps, and up to 40% of the screened population will be diagnosed with one or more polyps, CRC is particularly well-suited to evaluate novel chemopreventive agents. Our findings highlight candidate targets for chemoprevention, such as gut microbiota,

prostaglandin metabolism, and signaling through the Wnt ~~WNT~~, BMP and Hippo pathways. Specific potential targets in the near term include CDK6, which is targeted by drugs in clinical use for cancer therapy, such as palbociclib and ribociclib. Similarly, Wnt ~~WNT~~ pathway activity can be targeted indirectly using porcupine inhibitors (e.g. LGK974, ETC159, CGX-1321 and RXC004) that prevent Wnt ~~WNT~~ ligand palmitoylation[22], although future approaches may more specifically target effector genes such as *WNT4* and *ZNRF3*. Hence, adapted forms of these drugs or modified dosing regimens could be repurposed for chemoprevention, possibly initially for high-risk groups, such as those with in the top PRS percentiles or Lynch Syndrome cases. Based on our data, we speculate that in the longer term, targeted approaches based on demethylation of specific CpG sites from MWAS could be effective means of prevention with minimal toxicity.

The identification of additional risk associations has the potential to provide further biological insights into CRC. However, cohort numbers required in European and East Asian populations to identify additional risk SNPs through GWAS are likely to be prohibitive. Indeed, to identify SNPs explaining 80% of the heritable risk of CRC risk loci, thus providing comprehensive biological insights, will require sample sizes in excess of 500,000 cases and at least that number of controls (**Supplementary figure 6**). This is far higher than a previous estimate[23], which was based on a small subset of the GWAS included herein. Extending GWAS to African and other populations may detect further risk SNPs, including population specific ones. Complementary approaches such as TWAS and MWAS are demonstrably useful for the discovery of further risk loci, especially if, and when, reference data sets from multiple populations are made available.

Overall, our findings demonstrate the power of multi-omics to provide new insights into the biological basis of CRC, including both the identification of candidate effector genes and support for previously unsuspected functional mechanisms. Importantly, several of the genes and pathways we have identified are potential targets for CRC treatment or chemoprevention.

Funding and acknowledgements

At the Institute of Cancer Research, this work was supported by Cancer Research UK (C1298/A25514 - RSH). Additional support was provided by the National Cancer Research Network. In Edinburgh, the work was supported by Programme Grant funding from Cancer Research UK (C348/A12076 to MGD, C6199/A16459 to IT), EU ERC Advanced Grant EVOCAN, and the infrastructure and staffing of the Edinburgh CRUK Cancer Research Centre. CFR was supported by a Marie Sklodowska-Curie Intra-European Fellowship Action (IEF-301077) for the INTERMPHEN project and received considerable help from many staff in the Department of Endoscopy at the John Radcliffe Hospital in Oxford. Support from the European Union [FP7/207–2013, grant 258236], FP7 collaborative project SYSCOL, and COST Actions EuColonGene and TransColonCan are also acknowledged [BM1206 and CA17118] (IT). We are grateful to many colleagues within UK Clinical Genetics Departments (for CORGI) and to many collaborators who participated in the VICTOR, QUASAR2 and SCOT trials. We also thank colleagues from the UK National Cancer Research Network (for NSCCG). IT acknowledges funding from Cancer Research UK Programme Grant C6199/A27327.

The work at Vanderbilt University Medical Center was supported by U.S. NIH grants R01CA188214, R37CA070867, UM1CA182910, R01CA124558, R01CA158473, and R01CA148667, as well as Anne Potter Wilson Chair funds from the Vanderbilt University School of Medicine (WZ). Sample preparation and genotyping assays at Vanderbilt University were conducted at the Survey and Biospecimen Shared Resources and Vanderbilt Microarray Shared Resource, supported in part by the Vanderbilt-Ingram Cancer Center (P30CA068485). Statistical analyses were performed on servers maintained by the Advanced Computing Center for Research and Education (ACCRE) at Vanderbilt University (Nashville, TN).

GECCO: Genetics and Epidemiology of Colorectal Cancer Consortium: National Cancer Institute, National Institutes of Health, U.S. Department of Health and Human Services (U01 CA164930, U01 CA137088, R01 CA059045, R01201407, R01CA206279). Genotyping services were provided by the Center for Inherited Disease Research (CIDR) contract number HHSN268201200008I. This

research was funded in part through the NIH/NCI Cancer Center Support Grant P30 CA015704. Scientific Computing Infrastructure at Fred Hutch funded by ORIP grant S10OD028685 (UP). Colorectal Cancer Transdisciplinary (CORECT) Study: 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 Colon CFR participant recruitment and collection of data and biospecimens used in this study were supported by the NCI, NIH (grant number U01 CA167551). OFCCR was supported through funding allocated to the Ontario Registry for Studies of Familial Colorectal Cancer (U01 CA074783). The content of this manuscript does not necessarily reflect the views or policies of the NCI or any of the collaborating centers in the Colon Cancer Family Registry (CCFR), nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government, any cancer registry, or the CCFR.

Author contributions

Study design: CFR, MNT, PJL, VM, GC, SBG, IT, WZ, MGD, RSH, UP; Patient recruitment and sample collection: CFR, CP, SMF, JPB, PGVS, XOS, JL, QC, XG, YLU, PB, JS, TAH, DVC, MM, GR, MOS, JO, DK, SJ, KJ, SSK, AES, MHS, YA, JEK, IO, WW, KEM, KOM, CT, ZR, YG, WJ, JLH, MAJ, AKW, RKP, JCF, RWH, SG, MOW, PAN, JPC, RK, TSM, RSK, DJK, IK, JB, LPM, PJ, PK, LAA, HR, EP, JGE, TC, UH, JOK, KP, TT, LR, BZ, SM, DA, JRP, DDB, EAP, NU, EMS, SBR, AG, PTC, VMS, JCC, MH, HB, MLS, JDP, MBS, MJG, NM, AC, SCB, LM, VA, MS, BEP, DTB, GGG, CHH, MCS, GEI, KJM, AFZ, JKG, KAS, FL, KO, YS, TOK, BVG, TJH, HH, RP, RBH, MEM, PP, SCL, YY, HJL, EW, LL, ATC, MCC, AL, DJH, CS, PCS, DAN, RES, JH, ZKS, PEV, LV, VV, NP, DS, AET, SDM, SJC, FvD, EJMF, MGD, AW, AN, BAP, LMF, LSC, SO,

CK, CIL, RLP, CXQ, SBE, CMT, ERM, LLM, AHW, CEM, GAC, CH, IJD, SEH, ET, SJR, MW, LYO, MAD, TUS, TY, NS, MI, VM, GC, SBG, IT, WZ, MD, RSH, UP; Molecular analysis: CFR, MNT, PJL, SLS, VDO, CP, SEB, VS, KD, SMF, PGVS, JL, QC, XG, YLU, PB, JS, JRH, TAH, DVC, CHD, MD, FRS, MM, GR, MOS, WW, JLH, DD, JPC, RK, RSK, DJK, KP, DA, SJW, EARN, JRP, EAP, KV, NU, EMS, PTC, JCC, MH, HB, MLS, MJG, AC, SCB, LM, BEP, MCS, GEI, AFZ, JKG, KAS, FL, RS, TOK, SIB, ST, DAC, PP, HJL, EW, KFD, EWP, ATC, AL, ADJ, CS, PCS, JH, CKE, DCT, AEK, FvD, EJMF, LCS, MGD, AW, LMF, SO, SAB, CK, YLI, CXQ, LLM, CQ, CEM, SEH, ET, SJR, VM, GC, SBG, IT, WZ, MD, RSH, UP; Data analysis: CFR, MNT, PJL, MT, ZC, SLS, VDO, LH, JFT, CP, KIS, VS, KD, JRH, MM, FMN, KP, ANS, ABK, CKE, WJG, DCT, YLI, CXQ, CQ, SBG, IT, WZ, MD, RSH, UP; Data interpretation: CFR, MNT, PJL, MT, ZC, SLS, VDO, LH, JFT, KIS, JRH, AKW, JCF, RWH, PTC, KKT, MJG, ANS, BEP, DAC, PP, MCC, ABK, LCS, SO, RLP, VM, GC, SBG, IT, WZ, MD, RSH, UP; Drafting or substantially revising manuscript: all authors; Supervision and funding: CFR, VM, SBG, IT, MD, RSH, UP.

Competing interests

AC is consultant to Bayer Pharma AG, Boehringer Ingelheim, and Pfizer Inc. for work unrelated to this manuscript; AS is an employee at Insitro, incl. consulting fees from BMS; HH is SAB for Invitae Genetics, Promega, and Genome Medical. Stock/Stock options for Genome Medical and GI OnDemand; JK is a consultant for Guardant Health; NP is a collaborator for Thrive and Exact, PGDx, CAGE, NeoPhore, Vidium, ManaTbio, and receives royalties for licensed technologies according to JHU rules; RKP collaborates with Eli Lilly, AbbVie, Allergan, Verily, and Alimentiv, which include consulting fees (outside of the submitted work); SAB has financial interest in Adaptive Biotechnologies; SBG is co-founder, Brogent International LLC; TSM receives research and honoraria from Merck Serono; ZKS's immediate family member serves as a consultant in Ophthalmology for Alcon, Adverum, Gyroscope Therapeutics Limited, Neurogene, and RegenexBio (outside the submitted work). VM has research projects and owns stocks of Aniling. The remaining authors declare no competing interests.

570

TABLES

Table 1. Previously unreported colorectal cancer risk associations identified by genome-wide association study analysis. *P*-values calculated from a fixed-effects meta-analysis; *, conditional SNP association, with *P*-values and ORs derived from analysis conditional on known risk loci within 1Mb; RAF, risk allele frequency; EUR, European ancestry population; EAS, East Asian ancestry population; OR, odds ratio; I^2, fraction of variance attributable to between study heterogeneity; bp, base pairs. Association statistics for European and East Asian populations are detailed in Supplementary Table 3.

SNP	Cytoband	Position (bp, GRCh37)	Risk/Alt Allele	RAF (EUR)	RAF (EAS)	OR (95% CI)	*P*-value	I^2 (%)	Closest gene (RefSeq)
rs34963268 *	1p36.12	22,710,877	G/C	0.84	0.77	1.07 (1.05-1.09)	6.28E-16	31	*ZBTB40*
rs5028523	1q24.3	172,864,224	A/G	0.53	0.05	1.04 (1.03-1.06)	1.44E-08	0	*TNFSF18*
rs12137232	1q32.1	201,885,446	G/T	0.52	0.19	1.04 (1.03-1.05)	7.71E-09	15	*LMOD1*
rs12078075	1q32.1	205,163,798	G/A	0.09	0	1.07 (1.05-1.10)	1.94E-08	0	*DSTYK*
rs2078095	1q43	240,408,346	G/A	0.28	0.23	1.04 (1.03-1.06)	2.08E-08	0	*FMN2*
rs4668039	2q24.3	169,025,379	G/A	0.2	0.52	1.04 (1.03-1.06)	3.32E-08	12	*STK39*
rs704417	3p14.1	64,252,424	T/C	0.51	0.89	1.05 (1.03-1.06)	4.35E-10	0	*PRICKLE2*
rs7623129 *	3p14.1	64,624,426	C/T	0.56	0.51	1.04 (1.02-1.05)	1.51E-08	5	*ADAMTS9*
rs2388976	4q26	115,502,406	A/G	0.44	0.45	1.04 (1.02-1.05)	1.75E-08	17	*UGT8*
rs10006803	4q31.3	151,501,208	C/G	0.5	0.45	1.04 (1.02-1.05)	2.58E-08	0	*LRBA*
rs1426947	4q34.1	175,420,523	T/C	0.42	0.66	1.04 (1.03-1.05)	7.48E-10	0	*HPGD*
rs3930345	5q14.3	82,881,255	C/T	0.8	0.75	1.05 (1.03-1.06)	6.82E-09	10	*VCAN*

rs472959	5q35.1	172,324,558	A/G	0.46	0.46	1.04 (1.03-1.05)	4.71E-09	24	*ERGIC1*
rs1294437	6p25.1	6,749,789	C/T	0.65	0.23	1.04 (1.03-1.06)	1.21E-08	0	*LY86*
rs9379084 *	6p24.3	7,231,843	G/A	0.88	0.8	1.07 (1.05-1.09)	1.79E-12	9	*RREB1*
rs209142 *	6p22.1	28,862,617	C/G	0.39	0.52	1.04 (1.02-1.05)	3.66E-08	20	*TRIM27*
rs57939401	6p21.1	45,572,071	A/G	0.1	0.13	1.07 (1.04-1.09)	3.51E-10	0	*RUNX2*
rs6912214 *	6p12.1	55,721,302	T/C	0.55	0.83	1.04 (1.03-1.05)	1.55E-08	20	*BMP5*
rs145997965 *	6q21	106,482,613	C/T	0.02	0	1.21 (1.13-1.29)	1.26E-08	0	*PRDM1*
rs6911915	6q22.1	117,809,031	C/T	0.44	0.43	1.05 (1.03-1.06)	3.99E-12	3	*DCBLD1*
rs151127921	6q23.2	133,993,925	T/C	0.02	0	1.17 (1.11-1.24)	3.19E-08	24	*EYA4*
rs1182197	7p22.2	2,863,289	A/C	0.63	0.7	1.04 (1.03-1.05)	5.32E-09	0	*GNA12*
rs12539962	7q11.23	73,167,259	C/T	0.72	0.63	1.04 (1.03-1.05)	2.96E-08	27	*ABHD11*
rs2527927	7q22.1	99,477,426	G/A	0.55	0.71	1.04 (1.03-1.06)	3.31E-10	2	*OR2AE1*
rs60911071	8p21.2	23,664,632	G/C	0.95	0.64	1.06 (1.04-1.09)	2.24E-08	0	*STC1*
rs826732	8q12.1	59,742,639	C/G	0.5	0.59	1.04 (1.03-1.06)	6.26E-10	7	*TOX*
rs11557154	9p13.3	34,107,505	T/C	0.14	0.59	1.05 (1.04-1.07)	6.02E-10	14	*DCAF12*
rs10978941	9q31.2	110,373,819	C/T	0.83	0.87	1.06 (1.04-1.08)	2.29E-12	0	*KLF4*
rs7038489 *	9q34.2	136,682,468	C/T	0.89	0.99	1.08 (1.05-1.1)	1.1E-08	48	*VAV2*
rs11789898	9q34.2	136,925,663	T/G	0.18	0.08	1.05 (1.04-1.07)	6.28E-09	36	*BRD3*
rs1775910 *	10p12.1	29,096,942	G/C	0.25	0.32	1.04 (1.03-1.06)	3.11E-08	17	*LOC100507605*
rs1773860	10p12.1	29,291,556	T/C	0.49	0.35	1.04 (1.03-1.05)	3.49E-09	6	*LOC100507605*

rs10751097	11q13.3	69,938,433	A/G	0.4	0.31	1.05 (1.03-1.06)	2.14E-12	0	*ANO1*
rs497916	11q23.3	118,758,089	T/C	0.28	0.17	1.04 (1.03-1.06)	3.37E-08	0	*CXCR5*
rs7297628	12q14.2	64,404,555	T/C	0.54	0.75	1.04 (1.03-1.05)	1.39E-08	30	*SRGAP1*
rs11178634	12q21.1	71,518,329	G/T	0.62	0.7	1.05 (1.03-1.06)	1.36E-11	34	*TSPAN8*
rs7299936 *	12q24.21	115,934,000	A/G	0.56	0.18	1.04 (1.02-1.05)	3.73E-08	0	*MED13L*
rs116964464	13q12.13	27,543,193	T/C	0.03	0.04	1.11 (1.07-1.15)	4.83E-09	3	*USP12*
rs1078563 *	13q34	110,352,851	G/C	0.33	0.28	1.04 (1.03-1.05)	1.53E-08	0	*IRS2*
rs1497077	14q22.1	52,491,655	C/T	0.66	0.76	1.04 (1.03-1.06)	3.64E-08	0	*NID2*
rs8031386	15q23	72,508,799	A/C	0.26	0.54	1.04 (1.03-1.06)	4.50E-09	12	*PKM2*
rs11247566 *	17p13.3	835,371	G/A	0.55	0.52	1.04 (1.02-1.05)	2.92E-08	35	*NXN*
rs1791373	18p11.31	3,616,779	T/A	0.43	0.14	1.04 (1.03-1.06)	1.13E-08	0	*DLGAP1*
rs10409772	19p13.3	5,840,926	A/C	0.09	0.29	1.07 (1.05-1.09)	1.33E-10	6	*FUT6*
rs9983528	21q22.3	47,772,439	A/G	0.13	0.24	1.07 (1.05-1.09)	5.10E-13	0	*PCNT*
rs4616575	22q12.1	29,406,076	T/G	0.52	0.56	1.04 (1.03-1.05)	1.49E-10	0	*ZNRF3*
rs130651	22q13.1	39,644,273	G/A	0.33	0.08	1.05 (1.03-1.07)	2.92E-10	46	*PDGFB*
rs5751474	22q13.2	43,689,542	A/G	0.79	0	1.05 (1.03-1.07)	1.80E-08	52	*SCUBE1*
rs34256596 *	22q13.2	43,778,431	A/G	0.26	0.4	1.05 (1.03-1.06)	5.86E-09	0	*MPPED1*
rs9330814 *	22q13.31	46,364,191	T/C	0.33	0.68	1.05 (1.03-1.07)	1.28E-09	33	*WNT7B*

577

Table 2. Colorectal cancer risk associations identified by a colorectal mucosa-specific transcriptome-wide association study. SMultiXcan uses a two-sided F-test to quantify the significance of the joint fit of the linear regression of the phenotype on predicted expression from multiple tissue models jointly. All associations shown were transcriptome-wide significant after Bonferroni correction for 12,017 genes with an S-MultiXcan model (*i.e.* $P = 0.05/12{,}017 = 4.16 \times 10^{-6}$ for the $P_{\text{S-MultiXcan}}$). Genes with boundaries less than 1Mb apart were considered to be in the same cluster. This resulted in 13 CRC associations, for which all TWAS-significant genes were > 1 Mb away from and independent of any GWAS-significant SNP ($P_{GWAS} < 5 \times 10^{-8}$) As expected SNPs close to genome-wide significance were found in all cases. Two further gene associations (*) were < 1Mb from a GWAS-significant SNP, but in analysis conditional on the SNP showed a minimally changed association (**Supplementary Table 6**) and remained significant at $P = 4.16 \times 10^{-6}$. # indicates the number of novel TWAS loci. z score and effect size are calculated as the mean across S-PrediXcan models from the TWAS reference data sets. n models shows the number of reference data sets for which the S-PrediXcan elastic nets produced genetically-predicted expression models, with the n indep showing the number of those models that were statistically independent. The SNP with the lowest CRC GWAS *P*-value within 1Mb of the gene is also shown.

#	ENSEMBL identifier	Gene	Chr	Start (bp, GRCh37)	End (bp, GRCh37)	$P_{\text{S-MultiXcan}}$	Mean z score	Effect size	n models	n indep	Top GWAS SNP at <1Mb	SNP position	P_{GWAS}
1	ENSG00000171621	*SPSB1*	1	9,352,939	9,429,591	2.96E-06	4.569	0.077	3	1	rs2075971	9,407,104	1.96E-07
2	ENSG00000142632	*ARHGEF19*	1	16,524,712	16,539,104	2.32E-06	-4.610	-0.046	7	1	rs2132851	16,537,752	7.20E-07
	ENSG00000237276	*ANO7P1*	1	16,542,404	16,554,522	1.27E-06	-4.801	-0.054	3	1	rs2132851	16,537,752	7.20E-07
3*	ENSG00000237190	*CDKN2AIPNL*	5	133,737,778	133,747,589	1.37E-09	1.665	0.045	3	3	rs647161	134,499,092	8.53E-18
4	ENSG00000260653	*RP11-114G11.5*	7	57,404,172	57,419,535	1.37E-06	-4.829	-0.494	1	1	rs4242307	57,477,102	2.28E-03
5	ENSG00000204175	*GPRIN2*	10	46,994,087	47,005,643	3.38E-14	-7.582	-1.709	1	1	rs10906949	47,698,776	1.58E-04
6	ENSG00000180210	*F2*	11	46,740,730	46,761,056	2.80E-07	5.136	0.257	1	1	rs7109707	46,818,814	5.30E-07

	ENSG00000123444	*KBTBD4*	11	47,595,014	47,600,561	5.48E-07	5.008	0.053	1	1	rs7109707	46,818,814	5.30E-07
7	ENSG00000213445	*SIPA1*	11	65,405,568	65,418,401	2.81E-06	-3.033	-0.046	2	2	rs570760	65,833,631	2.88E-07
8	ENSG00000166106	*ADAMTS15*	11	130,318,869	130,346,532	3.86E-06	4.515	0.125	2	2	rs7936386	130,462,505	9.18E-08
9	ENSG00000174106	*LEMD3*	12	65,563,351	65,642,107	2.15E-06	3.040	0.076	3	3	rs59829994	65,560,831	1.39E-07
10*	ENSG00000234608	*MAPKAPK5-AS1*	12	112,277,588	112,280,706	6.15E-14	3.544	0.050	6	6	rs653178	112,007,756	2.51E-24
11	ENSG00000167173	*C15orf39*	15	75,487,984	75,504,510	2.14E-07	4.036	0.100	3	2	rs17338413	75,474,936	2.15E-07
	ENSG00000260274	*RP11-817O13.8*	15	75,660,496	75,661,925	2.93E-06	3.090	0.096	2	2	rs17338413	75,474,936	2.15E-07
12	ENSG00000166822	*TMEM170A*	16	75,476,952	75,499,395	1.05E-06	-3.464	-0.041	7	4	rs4888408	75,432,824	9.14E-07
13	ENSG00000131748	*STARD3*	17	37,793,318	37,819,737	8.11E-07	4.933	0.143	1	1	rs2313171	37,833,842	2.77E-07
	ENSG00000161395	*PGAP3*	17	37,827,375	37,853,050	9.59E-07	4.777	0.043	7	1	rs2313171	37,833,842	2.77E-07
	ENSG00000141736	*ERBB2*	17	37,844,361	37,886,606	2.96E-06	2.679	0.032	3	3	rs2313171	37,833,842	2.77E-07
14	ENSG00000152217	*SETBP1*	18	42,260,138	42,648,475	3.11E-07	4.339	0.093	2	2	rs12958322	42,309,786	2.60E-07
15	ENSG00000267100	*ILF3-AS1*	19	10,762,538	10,764,520	2.70E-07	4.689	0.079	2	2	rs10408721	10,758,319	5.71E-08

591

Table 3. Colorectal cancer risk associations identified by a colorectal mucosa-specific transcript isoform-wide association study (TIsWAS). As per Table 2, SMultiXcan uses a two-sided F-test to quantify the significance of the joint fit of the linear regression of the phenotype on predicted expression from multiple tissue models jointly. All associations shown were transcriptome-wide significant after Bonferroni correction for 27,941 transcripts with an S-MultiXcan model (*i.e.* $P = 0.05/27{,}941 = 1.79 \times 10^{-6}$ for the $P_{\text{S-MultiXcan}}$). Novel associations were called when >1Mb from both a GWAS-significant SNP and a TWAS locus. As expected, all these loci showed evidence of a risk association in the full TWAS (*FDR* < 0.05, $P < 2.86 \times 10^{-3}$). Transcripts with boundaries < 1 Mb apart were considered to be in the same cluster. This resulted in seven CRC associations. One further association (*) was identified based on conditional TIsWAS analysis (**Supplementary Table 8**). Other annotations are as per **Table 2**.

#	ENSEMBL identifier	Gene	Chr	Start (bp, GRCh37)	End (bp, GRCh37)	$P_{\text{S-MultiXcan}}$	Mean z score	Effect size	n models	n indep	Top GWAS SNP at <1Mb	SNP location	P_{GWAS}
1	ENST00000609196	*ACP6*	1	147,101,453	147,131,116	6.43E-11	-1.264	-0.048	4	3	rs1541187	147,051,493	1.44E-04
	ENST00000493129	*ACP6*	1	147,127,341	147,142,574	1.65E-23	-5.781	-0.482	2	2	rs1541187	147,051,493	1.44E-04
2	ENST00000273153	*CSRNP1*	3	39,183,346	39,195,066	9.99E-07	4.891	0.099	1	1	rs4676609	39,214,256	4.63E-06
3	ENST00000274695	*CDKAL1*	6	20,534,688	21,232,635	1.29E-06	-4.841	-0.046	1	1	rs9295474	20,652,717	7.61E-08
4	ENST00000481601	*CCDC183*	9	139,694,767	139,702,192	9.60E-07	-4.490	-0.048	2	2	rs2811736	139,651,954	3.12E-05
	ENST00000464157	*ABCA2*	9	139,902,688	139,903,240	7.39E-07	-4.951	-0.235	1	1	rs2811736	139,651,954	3.12E-05
5 *	ENST00000543000	*PLEKHG6*	12	6,426,733	6,427,529	3.30E-09	6.003	0.076	3	2	rs10849433	6,406,904	6.73E-17
6	ENST00000448790	*TOX4*	14	21,945,335	21,967,315	1.22E-07	5.290	0.498	1	1	rs3811252	22,855,779	2.11E-05
7	ENST00000478981	BNIP2	15	59,955,092	59,961,148	9.91E-07	-4.893	-0.326	1	1	rs7182962	59,945,783	6.04E-08

8	ENST00000310144	*PSMC5*	17	61,904,543	61,909,379	4.18E-10	6.247	0.553	1	1	rs12449782	61,576,249	2.18E-05

601

602

603 **Table 4. Colorectal cancer risk associations identified by cross-tissue transcriptome-wide association study.** SMultiXcan uses a two-
604 sided F-test to quantify the significance of the joint fit of the linear regression of the phenotype on predicted expression from multiple
605 tissue models jointly. TWAS tests were performed separately for the following tissue categories: *"Colon_sigmoid"*: GTEx (n=318
606 samples; $P_{Bonferroni}$ = 8.12 x 10^{-6} for the $P_{S\text{-}PrediXcan}$)*;* "*Immune*": DGN + GTEx Cells_EBV-transformed_lymphocytes + GTEx Whole_Blood
607 + GTEx_Spleen (n=1,966 samples; $P_{Bonferroni}$ = 3.34 x 10^{-6} for the $P_{S\text{-}MultiXcan}$); "*Mesenchymal*": GTEx Adipose_Subcutaneous + GTEx
608 Adipose_Visceral_Omentum + GTEx Cells_Cultured_fibroblasts (n=1,533 samples; $P_{Bonferroni}$ = 3.96 x 10^{-6} for the $P_{S\text{-}MultiXcan}$);
609 "*Gastrointestinal*": the 6 in-house colorectal mucosa datasets + GTEx Pancreas + GTEx Liver + GTEx Stomach + GTEx Terminal_Ileum +
610 GTEx Oesophageal_Mucosa + GTEx Colon_Transverse (n=2,615 samples; $P_{Bonferroni}$ = 3.34 x 10^{-6} for the $P_{S\text{-}MultiXcan}$); "*All*": the 6 in-house
611 colorectal mucosa datasets + all GTEx 49 tissues + DGN (n=16,832 samples; $P_{Bonferroni}$ = 2.31 x 10^{-6} for the $P_{S\text{-}MultiXcan}$). Other annotations
612 are as per **Table 2**.
613

#	Gene	Chr	Start (bp, GRCh37)	End (bp, GRCh37)	$P_{S\text{-}MultiXcan}$	Tissue	Mean z score	Effect size	n models	n indep	Top GWAS SNP at <1Mb	SNP location	P_{GWAS}
1	*RPL5*	1	93,297,540	93,307,481	2.27E-07	All	-1.160	-0.167	2	2	rs7530780	93,130,268	4.18E-05
2	*LINGO4*	1	151,772,740	151,778,546	2.73E-08	All	1.666	0.034	27	6	rs9826	151,778,899	3.81E-06
3	*FAM98A*	2	33,808,725	33,824,429	2.98E-06	Immune	4.672	0.166	1	1	rs1448561	33,854,344	5.92E-07
4	*FBLN7*	2	112,895,962	112,945,793	1.28E-06	All	-0.711	-0.023	28	10	rs7580507	112,879,209	2.71E-07
5	*ARHGEF4*	2	131,671,559	131,804,836	2.33E-08	All	-0.243	-0.026	14	8	rs73960398	131,795,345	4.86E-06
6	*GBE1*	3	81,538,850	81,811,312	1.95E-12	All	-0.557	-0.032	8	7	rs554330436	81.039,172	1.69E-04
7	*DIRC2*	3	122,513,642	122,599,986	1.25E-06	All	0.812	0.003	16	13	rs6774610	122,521,477	6.85E-07
8	*GAB1*	4	144,258,304	144,395,721	1.11E-07	All	1.756	0.040	10	6	rs72726477	143,517,452	2.91E-05

9	*FBXO38*	5	147,763,498	147,822,399	2.11E-06	Mesenchymal	4.677	0.287	2	2	rs35548425	147,816,153	1,80E-07
10	*EPB41L2*	6	131,160,487	131,384,462	2.70E-11	Gastrointestinal	-1.720	-0.018	8	6	rs12662663	131,398,523	6.71E-08
	EPB41L2	6	131,160,487	131,384,462	2.96E-09	All	-0.108	0.024	24	11	rs12662663	131,398,523	6.71E-08
11	*CDK6*	7	92,234,235	92,465,908	8.00E-14	All	0.281	0.037	8	6	rs143120528	92,258,733	2.49E-07
12	*PSMD13*	11	236,546	252,984	3.89E-06	Mesenchymal	1.737	0.113	3	2	rs7394572	432,436	4.88E-06
	IFITM1	11	313,506	314,456	6.73E-07	All	-0.090	-0.071	33	18	rs7394572	432,436	4.88E-06
13	*RHOG*	11	3,848,208	3,862,213	1.58E-06	Gastrointestinal	-1.862	-0.232	2	2	rs10835185	3,862,343	5.97E-08
	RHOG	11	3,848,208	3,862,213	8.27E-07	Mesenchymal	-4.929	-0.476	1	1	rs10835185	3,862,343	5.97E-08
	OR51E2	11	4,701,401	4,719,084	7.44E-06	Colon Sigmoid	4.480	0.336	1	1	rs10835185	3,862,343	5.97E-08
14	*ME3*	11	86,152,150	86,383,678	2.62E-06	Gastrointestinal	-0.215	-0.125	5	5	rs74402426	86,161,656	1.89E-05
15	*TAGLN*	11	117,070,037	117,075,052	5.80E-09	All	-2.118	-0.111	14	9	rs1035237	116,727,850	5.43E-08
15	*PCSK7*	11	117,075,499	117,103,241	2.67E-06	Mesenchymal	3.281	0.311	2	2	rs1035237	116,727,850	5.43E-08
16	*CLIP1*	12	122,755,979	122,907,179	7.61E-08	All	0.664	0.026	6	5	rs1716169	123,716,930	1.58E-06
17	*ATP2C2*	16	84,402,133	84,497,793	4.44E-07	Gastrointestinal	1.903	0.021	7	5	rs7187803	84,501,660	1.07E-05
	ATP2C2	16	84,402,133	84,497,793	2.89E-07	All	0.754	0.010	23	14	rs7187803	84,501,660	1.07E-05
18	*CBFA2T3*	16	88,941,266	89,043,612	1.11E-06	Mesenchymal	4.871	0.253	1	1	rs502258	88,968,547	9.90E-06
19	*LLGL1*	17	18,128,901	18,148,149	3.05E-06	Immune	-4.667	-0.469	1	1	rs6502570	17,183,255	2.63E-06
20	*PSMC3IP*	17	40,725,329	40,729,849	2.21E-06	All	1.575	0.108	11	9	rs12949918	40,526,273	1.39E-06
	BECN1	17	40,963,673	40,985,158	1.14E-06	Immune	4.824	0.547	2	2	rs12949918	40,526,273	1.39E-06
21	*SMAD4*	18	48,554,764	48,611,415	2.75E-06	Mesenchymal	4.750	0.653	2	2	rs12958467	48,481,751	4.69E-07
22	*ATP8B1*	18	55,313,658	55,470,547	2.54E-06	Immune	-4.704	-0.203	1	1	rs8097764	55,317,896	1.49E-07

23	*LIF*	22	30,636,528	30,640,922	4.96E-06	Colon Sigmoid	-4.566	-0.201	1	1	rs12484740	30,606,927	4.97E-06

614

615

Table 5. Colorectal cancer risk associations identified by methylome-wide association study. SMultiXcan uses a two-sided F-test to quantify the significance of the joint fit of the linear regression of the phenotype on predicted expression from multiple tissue models jointly. All associations shown were methylome-wide significant after Bonferroni correction for 88,888 CpGs with an S-PrediXcan model (P = 0.05/88,888 = 5.62 x 10^{-7} for the $P_{S\text{-}MultiXcan}$). Pairs of CpGs or strings of adjacent CpGs within 1Mb of one another were considered to lie within the same cluster. Five CRC associations were found for which all CpGs were > 1 Mb away from GWAS-significant SNP (P_{GWAS} < 5 x 10^{-8}), although near a SNP close to genome-wide significance. Two further associations for 4 CpGs (*) were identified based on conditional MWAS analysis (**Supplementary Table 15**). Novel CpG hits were all independent of each other and of GWAS SNPs and TWAS genes. Other annotations are as per **Table 2**.

#	CpG	Annotated Gene	Chr	Probe location (bp, GRCh37)	Probe annotation	$P_{S\text{-}MultiXcan}$	Mean z score	Effect size	n models	n indep	Top GWAS SNP at <1Mb	SNP location	P_{GWAS}
1	cg01716680	*GJA4*	1	35,259,750	S Shore	3.41E-07	-5.099	-0.164	1	1	rs57975061	34,890,238	2.42E-06
2	cg15917621	*NRBP1*	2	27,650,478	N Shore	1.61E-07	-3.301	-0.094	2	2	rs4665972	27,598,097	1.58E-07
3	cg02609692	*LMX1B*	9	129,389,125	Island	4.24E-07	5.058	0.112	1	1	rs4075850	130,169,301	1.76E-06
4*	cg12931523	*TTLL13*	15	90,793,004	S Shore	7.74E-09	4.511	0.067	3	3	rs71407320	91,185,291	3.61E-08
	cg05239308	*TTLL13*	15	90,793,057	S Shore	1.54E-07	5.364	0.114	3	2	rs71407320	91,185,291	3.61E-08
	cg27018984	*TTLL13*	15	90,796,558	S Shelf	3.64E-09	-5.900	-0.089	1	1	rs71407320	91,185,291	3.61E-08
5	cg02086790	*AXIN1*	16	375,327	Island	2.75E-07	2.471	0.042	3	3	rs9921222	375,782	7.10E-07
6*	cg09894072	*PLA2G15*	16	68,279,487	Island	2.26E-07	5.176	0.096	1	1	rs9939049	68,812,301	1.95E-12

7	cg15135657	*LOC100631378*	19	38,346,511	S Shore	1.55E-07	-2.170	-0.032	2	2	rs55876653	39,146,780	2.10E-06

625

626

627 **Figure 1. Summary of the study data and analytical design, and the number of previously unreported CRC risk loci discovered.** The
628 figure illustrates the information for the different analyses used: GWAS (green), TWAS (blue), MWAS (yellow) used to identify
629 additional risk loci. These are later used to select credible effector genes annotated to functions and tissues.
630
631 **Figure 2. Effector genes for CRC risk and the cellular processes in which they act.** Pie chart describing the proportion and list of
632 effector genes allocated to each process.
633
634 **Figure 3. Representation of effector genes and their putative actions in the colorectum.** Diagram representing the processes that
635 the combined GWAS, TWAS and MWAS analyses have unveiled as relevant to CRC risk. Exemplar effector genes from cellular processes
636 and pathways (in capitals) are chosen to depict each category.
637
638

639

640

641

642

643 **References**

644 1. Sung H, Ferlay J, Siegel R, et al. Global cancer statistics 2020: GLOBOCAN estimates of
645 incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2021.
646 2. Jiao S, Peters U, Berndt S, et al. Estimating the heritability of colorectal cancer. Hum Mol
647 Genet. 2014;23(14):3898-3905.
648 3. Law PJ, Timofeeva M, Fernandez-Rozadilla C, et al. Association analyses identify 31 new
649 risk loci for colorectal cancer susceptibility. Nat Commun. 2019;10(1):2154.
650 4. Huyghe JR, Bien SA, Harrison TA, et al. Discovery of common and rare genetic risk variants
651 for colorectal cancer. Nat Genet. 2019;51(1):76-87.
652 5. Zhu Z, Zhang F, Hu H, et al. Integration of summary data from GWAS and eQTL studies
653 predicts complex trait gene targets. Nat Genet. 2016;48(5):481-487.
654 6. Speed D, Holmes J, Balding DJ. Evaluating and improving heritability models using
655 summary statistics. Nat Genet. 2020;52(4):458-462.
656 7. Kvale MN, Hesselson S, Hoffmann TJ, et al. Genotyping Informatics and Quality Control
657 for 100,000 Subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA)
658 Cohort. Genetics. 2015;200(4):1051-1060.
659 8. Wang H, Burnett T, Kono S, et al. Trans-ethnic genome-wide association study of
660 colorectal cancer identifies a new susceptibility locus in VTI1A. Nat Commun. 2014;5:4613.
661 9. Gamazon ER, Wheeler HE, Shah KP, et al. A gene-based association method for mapping
662 traits using reference transcriptome data. Nat Genet. 2015;47(9):1091-1098.
663 10. Barbeira AN, Pividori M, Zheng J, Wheeler HE, Nicolae DL, Im HK. Integrating predicted
664 transcriptome from multiple tissues improves association detection. PLoS Genet.
665 2019;15(1):e1007889.
666 11. Bien SA, Su YR, Conti DV, et al. Genetic variant predictors of gene expression provide new
667 insight into risk of colorectal cancer. Hum Genet. 2019;138(4):307-326.
668 12. Guo X, Lin W, Wen W, et al. Identifying Novel Susceptibility Genes for Colorectal Cancer
669 Risk From a Transcriptome-Wide Association Study of 125,478 Subjects. Gastroenterology.
670 2021;160(4):1164-1178 e1166.

671 13. Battle A, Mostafavi S, Zhu X, et al. Characterizing the genetic basis of transcriptome
672 diversity through RNA-sequencing of 922 individuals. Genome Res. 2014;24(1):14-24.
673 14. Koo BK, Spit M, Jordens I, et al. Tumour suppressor RNF43 is a stem-cell E3 ligase that
674 induces endocytosis of Wnt receptors. Nature. 2012;488(7413):665-669.
675 15. Hirano Y, Iwase Y, Ishii K, Kumeta M, Horigome T, Takeyasu K. Cell cycle-dependent
676 phosphorylation of MAN1. Biochemistry. 2009;48(7):1636-1643.
677 16. Fattet L, Yang J. RREB1 Integrates TGF-beta and RAS Signals to Drive EMT. Dev Cell.
678 2020;52(3):259-260.
679 17. Keku TO, Dulal S, Deveaux A, Jovov B, Han X. The gastrointestinal microbiota and
680 colorectal cancer. Am J Physiol Gastrointest Liver Physiol. 2015;308(5):G351-363.
681 18. Tuomisto AE, Makinen MJ, Vayrynen JP. Systemic inflammation in colorectal cancer:
682 Underlying factors, effects, and prognostic significance. World J Gastroenterol.
683 2019;25(31):4383-4404.
684 19. Zheng J, Erzurumluoglu AM, Elsworth BL, et al. LD Hub: a centralized database and web
685 interface to perform LD score regression that maximizes the potential of summary level GWAS
686 data for SNP heritability and genetic correlation analysis. Bioinformatics. 2017;33(2):272-279.
687 20. Pearson-Stuttard J, Papadimitriou N, Markozannes G, et al. Type 2 Diabetes and Cancer:
688 An Umbrella Review of Observational and Mendelian Randomization Studies. Cancer Epidemiol
689 Biomarkers Prev. 2021;30(6):1218-1228.
690 21. Kyrgiou M, Kalliala I, Markozannes G, et al. Adiposity and cancer at major anatomical sites:
691 umbrella review of the literature. BMJ. 2017;356:j477.
692 22. Liu J, Pan S, Hsieh MH, et al. Targeting Wnt-driven cancer through the inhibition of
693 Porcupine by LGK974. Proc Natl Acad Sci U S A. 2013;110(50):20224-20229.
694 23. Zhang YD, Hurson AN, Zhang H, et al. Assessment of polygenic architecture and risk
695 prediction based on common variants across fourteen cancers. Nat Commun. 2020;11(1):3353

Methods

The research presented in this study complies with all relevant ethical regulations, and has been approved by the South Central Ethics Committee (UK) (reference number 17/SC/0079).

Data availability

Summary level data for the full set of Asian and European GWAS are available through GWAS catalog (accession number GCST90129505). For individual-level data, CCFR, CORECT, CORSA_2 and GECCO are deposited in dbGaP (phs001415.v1.p1, phs001315.v1.p1, phs001078.v1.p1, phs001903.v1.p1, phs001856.v1.p1 and phs001045.v1.p1). NSCCG and COIN are available in the European Genome-phenome Archive under accession numbers EGAS00001005412 (NSCCG), EGAS00001005421 (COIN). UK Biobank data are available through http://www.ukbiobank.ac.uk/ and Finnish data through THL Biobank. Access to individual-level data for the remaining studies is controlled through oversight committees. CCFR 1 and CCFR 2 data can be requested by submitting an application for collaboration to the CCFR (forms, instructions and contact information can be located at (www.coloncfr/collaboration.org). Applications for individual level data from the QUASAR2 and SCOT clinical trials will be assessed by the Translational Research Steering Committees that oversee those studies. Individual level data from the CORGI (UK1) study will be made available subject to standard institutional agreements. Application forms for these three studies, and for Scotland Phase 1, Scotland Phase 2, SOCCS, DACHS4 and Croatia, will be provided by emailing a request to access.crc.gwas.data@outlook.com. For access to CORSA_1, please contact gecco@fredhutch.org. For Generation Scotland (GS) access is through the GS Access Committee (GSAC) (access@generationscotland.org). Applications for The Lothian Birth Cohort data should be made through https://www.ed.ac.uk/lothian-birth-cohorts/data-access-collaboration. For details of the application process for Aichi1, Aichi2, BBJ, Guanzhou1, HCES, HCES2, Korea and Shanghai cohorts, please go to https://swhs-smhs.app.vumc.org/ or contact Dr. Zheng at wei.zheng@vanderbilt.edu.

CRC-relevant epigenome data were obtained from the NCBI Gene Expression Omnibus (GEO) database under accession number GSE77737 and GSE36401.

Genetically predicted models of gene expression and methylation have been deposited in the Zenodo repository (https://zenodo.org/deposit/6472285).

Code availability

All bioinformatics and statistical analysis tools used in this study are open source, details of which are available in the Methods section and in the Reporting Summary. No custom code was used to process or analyse data. Details on URLs used can be found in the Supplementary Note.

Statistics and reproducibility

No statistical method was used to predetermine sample size. The experiments were not randomized. Data exclusion from each analysis is explained below in the corresponding sections. Informed consent was obtained for all participants in the study. A description of the different datasets and cohorts used is included in the Supplementary Note.

Criteria for declaring new CRC risk associations

Multi-omic studies present inherent difficulties for deciding on what constitutes a novel GWAS, TWAS or MWAS association. To declare statistically significant associations, for GWAS we have used the established threshold of $P = 5 \times 10^{-8}$. We applied this to both loci >1Mbp from a previously known SNP and analyses conditioned on the most significant SNP within 1Mb region. For TWAS or MWAS we also followed convention and used a Bonferroni correction $P = 0.05/N$, where N is the number of gene models successfully derived from the reference tissue. Furthermore, for TIsWAS and cross-tissue TWAS, we used Bonferroni-corrected *P*-value thresholds for significance in each of the reference tissue data sets separately, owing to the overlap in between tissue groups and the fact that many eQTLs are present across tissues. A further common practice, is that a new association should be located >1Mb from another association (from this study or previously reported), whether a genome-wide significant GWAS SNP, a TWAS gene or an MWAS CpG. However, use of the 1Mb distance convention introduces a

755 further problem in that, whilst the location of a GWAS SNP and MWAS CpG can be defined
756 precisely, the location of a gene cannot. We therefore defined a gene's boundaries by the
757 canonical transcript and novel associations must lie 1Mb from both those boundaries. Since
758 TWAS and MWAS associations can affect multiple nearby genes or CpGs (*e.g.* owing to co-
759 regulation or LD between eQTLs or mQTLs), we have conservatively assigned each TWAS and
760 MWAS association to a single locus (defined as a group of genes or CpGs that are significantly
761 associated with CRC risk and lie < 1Mb apart). Locus boundaries must be > 1Mb from another
762 association to be declared an independent risk association.
763 We have also performed conditional analyses across GWAS, TWAS and MWAS. This is standard
764 practice in GWAS (see below) [24], whereby nearby SNPs with no or limited correlation can be
765 independently associated with CRC risk. Conditioning TWAS, TIsWAS and MWAS on GWAS using
766 sMIST also allowed us to identify risk associations that were independent of the GWAS
767 associations within 1Mb, based on a $P_{conditional}$ that (i) remained Bonferroni-significant at the
768 unconditional analysis threshold, and (ii) was within one order of magnitude as $P_{unconditional}$. A
769 much larger number of TWAS and MWAS associations fulfilled only criterion (i) after conditioning
770 on a GWAS association within 1Mb (Supplementary Table 6, 8 and 15). Whilst we could not
771 exclude the possibility that some of these associations resulted from additional SNPs
772 independent of a nearby GWAS SNP for example, we conservatively did not declare these as
773 novel risk associations.
774
775 ***GWAS data analysis***
776 Meta-analysis: Within each of the 31 analytical units, we conducted logistic regression under a
777 log-additive model to examine the association between allelic dosage for each genetic variant
778 and the risk of CRC, adjusted for unit-specific covariates. Meta-analysis under a fixed-effects
779 inverse-variance weighted model was performed using META v1.7[25] . Variants in the meta-
780 analysis only included those with an imputation quality score (info/R^2) > 0.4, MAF > 0.005, and
781 seen in at least 15 analytical units. The I^2 statistic was calculated to quantify between study
782 heterogeneity and variants with I^2 > 65% were excluded. A total of 8,782,440 variants were taken
783 forward in the meta-analysis. Meta-analysis of risk estimates was conducted under an inverse

784 variance weighted, fixed-effects model[3]. None of the analytical units showed strong evidence of
785 genomic inflation (λ ranged from 0.95 to 1.28), and the λ value for the meta-analysis was 1.30
786 (λ_{1000} = 1.01) **Supplementary figure 3**). To account for any -ancestral differences between
787 analytical units, we implemented MR-MEGA v0.1.5[26] , including 10 principal components (PCs)
788 in the analysis. To measure the probability of associations being false positives, the Bayesian
789 False-Discovery Probability (BFDP)[3] was calculated based on a plausible odds ratio (OR) of 1.2
790 (based on the 95th percentile of the meta-analysis OR values) and a prior probability of
791 association of 10^{-5}.
792
793 Definition of known and novel GWAS SNP risk associations: We identified all previously reported
794 CRC associations at $P < 5 \times 10^{-8}$ by referencing the NHGRI-EBI Catalog of human GWAS and by
795 searching PubMed (performed June 2021)[3]. Additional articles were ascertained through
796 references cited in primary publications (Supplementary Table 4). Where multiple studies
797 reported associations in the same region ($r^2 > 0.1$ and within 500kb-1Mb of the index SNP), we
798 considered all variants with genome-wide significant associations. Given the improved power and
799 coverage of our study over previous works, we identified the most strongly associated variant at
800 each known signal and used lead variants for further analyses, rather than the previously
801 reported index variants (**Supplementary Table 3**). A genome-wide significant risk variant was
802 considered novel if >1Mb from a known risk variant.
803 GWAS conditional analysis: To identify independent association signals at the discovered CRC risk
804 associations, we performed conditional analyses using GCTA-COJO[24] on the meta-analysis
805 summary statistics. Analyses were performed separately for European and East Asian ancestry
806 populations, to account for LD structure differences. The conditioned data were meta-analyzed
807 together as described above, and associations with $P_{conditional} < 5 \times 10^{-8}$ were considered novel
808 secondary associations. As reference for LD estimation, we made use of genotyping data from
809 6,684 unrelated samples of East Asian ancestry, and 4,284 samples from combined UK10K and
810 European samples in 1000 Genomes.
811
812 ***Heritability analysis***

We used the LDSC regression package with default parameters as implemented in LD Hub[27] to estimate the SNP heritability from the GWAS meta-analysis summary statistics data[3]. SNPs were filtered to HapMap3 SNPS with 1000 Genomes EUR MAF above 5%. SNPs with imputation info score < 0.9, MAF < 0.01 and within the major histocompatibility complex (MHC) region (i.e. SNPs between 26Mb and 34Mb on chromosome six were excluded. Precalculated LD scores files computed using 1000 Genome European data were used.

The contribution of risk SNPs to the familial risk of CRC was calculated as $\sum_k \frac{log\lambda_k}{log\lambda_0}$, where λ_0 is the familial risk to first-degree relatives of CRC cases, assumed to be 2.2[28], and λ_k is the familial relative risk associated with SNP *k*, calculated as $\lambda_k = \frac{p_k r_k^2 + q_k}{(p_k r_k + q_k)^2}$, where p_k is the risk allele frequency for SNP *k*, $q_k = 1 - p_k$, and r_k is the estimated per-allele OR from the meta-analysis[3,29].

Pleiotropy analysis

We explored cross-trait pleiotropic effects using the LDSC regression package with default parameters[30] as implemented in LD Hub. The summary statistics for 252 phenotypes were extracted from LD Hub. For comparability of results across the traits we limited our analysis to the CRC GWAS of European ancestry. After excluding GWAS performed on non-European cohorts, traits where the LD Hub output came with the following warning messages: "Caution: using this data may yield results outside bounds due to relative low Z score of the SNP heritability of the trait" and "Caution: using this data may yield less robust results due to minor departure of the LD structure", as well as highly correlated traits, 171 phenotypes were included in the analysis. The departure of the LD structure means departure from the assumption of equal LD structure between two datasets, e.g due to differences in population structure between the study populations. SNPs from the MHC (chr6 26M~34M) region were removed for all traits prior to analysis.

Sample size prediction

To estimate the sample size required to detect a given proportion of the GWAS heritability, we made use of GENESIS software (GENetic Effect-Size distribution Inference from Summary-level data)[31], which implements a likelihood-based approach to model the effect-size distribution in conjunction with LD information, using the three-component model (mixture of two normal distributions). The percentage of GWAS heritability explained for a projected sample size was based on power calculations for the discovery of genome-wide significant SNPs[3]. The genetic variance explained was calculated as the proportion of total GWAS heritability explained by SNPs reaching genome-wide significance at a given sample size.

TWAS analysis

Gene expression models for the six in-house expression datasets were generated using the PredictDB v7 pipeline for a total of 1,077 participants[9,10]. Elastic net model building with 10-fold cross-validation was performed independently for each dataset. The elastic net models for GTEx v8 Colon Transverse were obtained from the PredictDB data repository (http://predictdb.org/) and had been generated using the same pipeline. Models were computed using HapMap2 SNPs ±1Mb from each gene, together with covariate factors estimated using PEER[32], clinical covariates when appropriate (age, sex and, where appropriate, case-control status, type of polyp and anatomic location in the colorectum), and three PCs from the individual dataset's SNP genotype data. Transcriptome-wide association tests were then performed for each dataset with the S-PrediXcan feature using summary statistics from the GWAS meta-analysis. We used individual level GWAS data from GECCO (n=8,725) to derive the LD reference covariance matrix. S-MultiXcan analysis was then undertaken across datasets. Significant associations were declared using Bonferroni correction (0.05/number of gene models from S-MultiXcan). As recommended[33], an additional filter of a TWAS association statistic, $P_{\text{S-PrediXcan}} \leq 10^{-4}$, in at least one individual reference data set was implemented to minimize potential errors due to LD mismatches. Genes localizing to the HLA/MHC region (chr6:28,477,797-33,448,354bp) were excluded.

Transcript-based TWAS analyses (TIsWAS) were likewise performed by using transcript-level data from the SOCCS, BarcUVa-Seq and GTEx Colon Transverse datasets.

Additional TWAS analyses were similarly performed using the non-colonic mucosa tissue data available from GTEx. These correspond to S-PrediXCan elastic net models from 48 additional GTEx tissues with eQTL data and the DGN whole blood cohort. Five tissue groupings were tested: *"Sigmoid colon",* corresponding to muscle and other sub-epithelial tissues; "*Immune*", comprising DGN + GTEx Cells_EBV-transformed_lymphocytes + GTEx Whole_Blood + GTEx_Spleen (n=1,966 samples); "*Mesenchymal*", comprising GTEx Adipose_Subcutaneous + GTEx Adipose_Visceral_Omentum + GTEx Cells_Cultured_fibroblasts (n=1,533 samples); "*Gastrointestinal*", comprising six in-house datasets + GTEx Pancreas + GTEx Liver + GTEx Stomach + GTEx Terminal_Ileum + GTEx Oesophageal_Mucosa + GTEx Colon_Transverse; n=2,615 samples); and "*All*", comprising the six in-house datasets + all 49 GTEx tissues + DGN (n=16,832 samples).

The predictive performance of the models for TWAS and TisWAS across the datasets was similar. For the TWAS models the number of genes successfully predicted with $R^2 > 0.01$ (equivalent of R>0.1) varied between 3308 for the BarcUVa data set and 5092 for SOCCS rectum, while GTEx Colon Transverse models were available for 6295 genes. The mean CV-based prediction R^2 for all genes varied between 0.09 (25-75th percentile 0.04-0.12) for BarcUVa to 0.19 for INTERMPHEN (0.07-0.24), compared with 0.12 (0.04-0.16) for GTEx Colon Transverse model. The numbers were slightly higher when comparing the overlapping 736 genes only. The in-house TisWAS models were constructed for a lesser number of transcripts (n=4632 for BarcUVa dataset and n=11262 for SOCCS rectum dataset) compared to GTEx Colon Transverse (n=15500), owing to greater read depth and larger sample size for GTEx. The mean R^2 for all genes varied from 0.07 (0.03-0.09) for BarcUVa to 0.16 for SOCCS colon (0.07-0.21). GTEx Colon Transverse had mean R^2 0.10 (0.03-0.12).

MWAS analysis

Methylation beta values were calculated based on the manufacturer's standard, ranging from 0 to 1. Quality control and data normalization were performed in R using the ChAMP software pipeline for the EPIC and 450K arrays[34]. Briefly, we filtered out failed probes with detection *P* >

0.02 in >5% of samples, probes with <3 reads in >5% of samples per probe and all non-CpG probes. Samples with failed probes >0.1 were also excluded from downstream analyses. We discarded all probes with SNPs within 10bp of the interrogated CpG (from 1,000 Genomes Project, CEU population)[35] , and probes that ambiguously mapped to multiple locations in the human genome with up to two mismatches[33]. We only considered probes mapping to autosomes and those overlapping between the EPIC and the 450K arrays. Normalization was achieved using the Beta MIxture Quantile (BMIQ) method. Per probe methylation models were created using the PredictDB pipeline on the normalized methylation matrix and the genotypes as per TWAS eQTL analysis. To optimize power, we restricted our analysis to 263,341-238,443 (for the 450K array) and 377,678 (for the EPIC array) probes annotated to Islands, Shores and Shelves, and discarded "Open Sea" regions. Further analysis was performed as per the TWAS. CpGs were annotated to a known GWAS signal if within 1Mb of a genome-wide significant GWAS risk SNP and otherwise considered novel. For the MWAS models the number of CpG probes successfully predicted with $R^2 > 0.01$ (equivalent of R>0.1) varied from 24325 for INTERMPHEN rectum to 30385 for COLONOMICS. The mean CV-based prediction R^2 for all genes varied from 0.14 (25th-7th percentile 0.07-0.16) for INTERMPHEN proximal dataset to 0.19 for SOCCS (0.07-0.25).

Conditional analysis using sMiST for TWAS and MWAS findings

S-MultiXcan is a powerful method for assessing predicted gene expression across multiple tissues and samples, but cannot readily undertake conditional analysis to determine independence of a TWAS or MWAS association from other GWAS, TWAS or MWAS associations. We therefore used the summary statistics-based Mixed effects Score Test (sMiST)[36] method to perform conditional analysis of TWAS, TIsWAS and MWAS data adjusting for GWAS risk SNPs. sMiST can assess the total effect, including both predicted molecular features (gene expression or methylation) and the residual direct effects of SNPs that are not explained by predicted molecular features, on CRC risk. To be consistent with S-MultiXcan, we only assessed the association of predicted molecular features. We first confirmed that there was a strong correlation between the sMiST and S-MultiXcan results, with minimal discordance (**Supplementary figure 4**). In view of this, we used sMiST to perform conditional TWAS and MWAS analysis for each of the

significantly associated genes or CpGs respectively, conditioning on the lead GWAS-significant SNP (if present) within 1Mb (**Supplementary Tables 6, 8** & **15**). We also conditioned TWAS on TWAS, TIsWAS on TIsWAS and MWAS on MWAS. We also conducted TWAS conditioned on MWAS analyses for the genes for which both significant genetically predicted expression and methylation models were produced by the PredictDB pipeline. Where multiple CpGs were annotated to the same gene, we selected the association with the lowest MWAS *P*-value. We determined the number of genes associated (at Bonferroni-corrected $P = 0.05/6{,}722 = 7.44 \times 10^{-6}$) with CRC risk in both TWAS and MWAS (n=43), TWAS-only (n=54), MWAS-only (n=91) or neither (n=6,534)."

Effector gene identification

To identify the most credible target or "effector" genes at each CRC risk locus, a pragmatic approach was utilized. After excluding the MHC region, pseudogenes and transcripts of uncertain significance (generally RPNNNN or ACNNN), the following hierarchical inclusion criteria were used.

For significant (Bonferroni-corrected $P_{TWAS} < 0.05$) TWAS genes at a locus, the gene most strongly associated with CRC risk in any tissue, as long as its P_{TWAS} was at least an order of magnitude lower than any other gene at the locus. (N=112)

For loci included under (1), additional genes that remained significant (FDR < 0.05) in conditional TWAS-TWAS analysis including the lead gene. (N=9)

At GWAS loci not included under (1), the most significant (FDR < 0.05) TWAS gene, as long as its P_{TWAS} was at least an order of magnitude lower than any other gene at the locus. (N=17)

TIsWAS analysis consistent with the approach used for TWAS as described in (1-3) above. (N=16)

Genes harboring missense or truncating variants in LD ($r^2 > 0.9$) with sentinel GWAS SNPs. (N=1)

A set of 155 genes was identified, which corresponds to about two thirds of the CRC risk loci from GWAS, TWAS and MWAS (**Supplementary Table 17**).

The area under the receiver operating characteristics curve (AUC)

We calculated the confounder adjusted AUC of PRS in discriminating individuals with and without CRC by using the propensity score weighting to account for potentially different distribution of confounders between cases and controls[37]. We adjusted for age, sex, and four PCs as confounders. We obtained the 95% confidence intervals (CI) by bootstrapping and a total of 500 bootstrap samples were generated. We calculated adjusted AUCs using the R package ROCt.

Methods-only references

24. Yang J, Ferreira T, Morris AP, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nat Genet. 2012;44(4):369-375, S361-363.

25. Liu JZ, Tozzi F, Waterworth DM, et al. Meta-analysis and imputation refines the association of 15q25 with smoking quantity. Nat Genet. 2010;42(5):436-440.

26. Magi R, Suleimanov YV, Clarke GM, et al. SCOPA and META-SCOPA: software for the analysis and aggregation of genome-wide association studies of multiple correlated phenotypes. BMC Bioinformatics. 2017;18(1):25.

27. Speed D, Balding DJ. SumHer better estimates the SNP heritability of complex traits from summary statistics. Nat Genet. 2019;51(2):277-284.

28. Johns LE, Houlston RS. A systematic review and meta-analysis of familial colorectal cancer risk. Am J Gastroenterol. 2001;96(10):2992-3003.

29. Schumacher FR, Al Olama AA, Berndt SI, et al. Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. Nat Genet. 2018;50(7):928-936.

30. Bulik-Sullivan BK, Loh PR, Finucane HK, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet. 2015;47(3):291-295.

31. Zhang Y, Qi G, Park JH, Chatterjee N. Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits. Nat Genet. 2018;50(9):1318-1326.

32. Stegle O, Parts L, Piipari M, Winn J, Durbin R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. Nat Protoc. 2012;7(3):500-507.

33. Barbeira AN, Pividori M, Zheng J, Wheeler HE, Nicolae DL, Im HK. Integrating predicted transcriptome from multiple tissues improves association detection. PLoS Genet. 2019;15(1):e1007889.

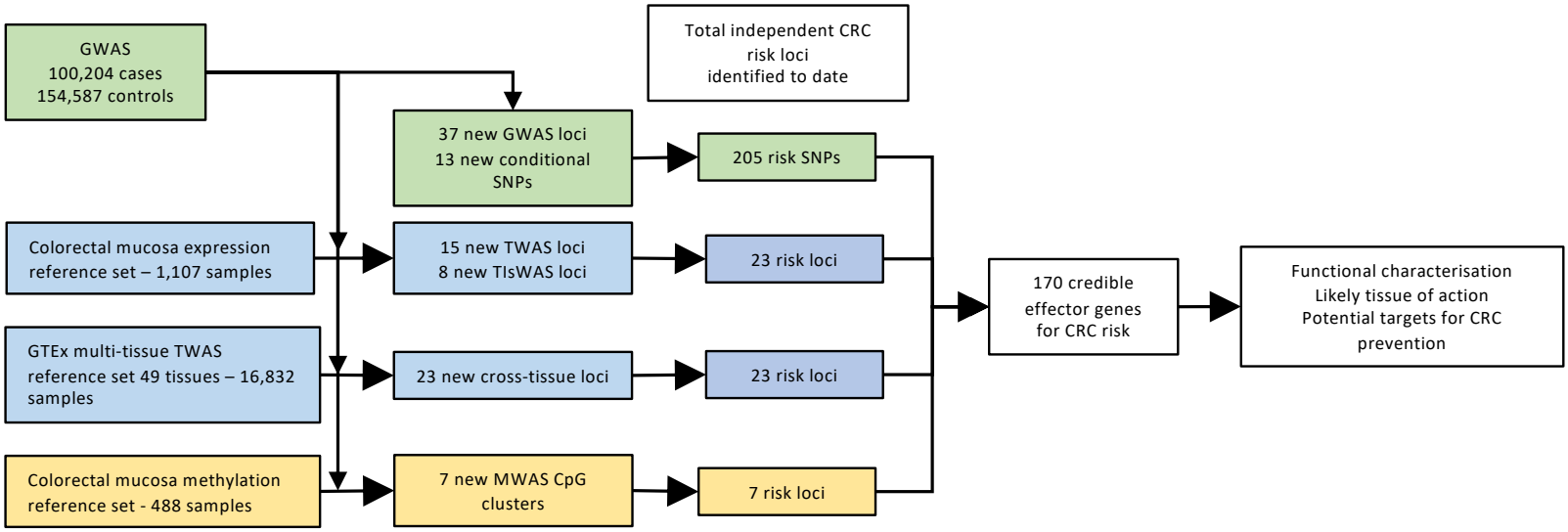
34. Tian Y, Morris TJ, Webster AP, Yang Z, Beck S, Andrew F, Teschendorff AE (2017). "ChAMP: updated methylation analysis pipeline for Illumina BeadChips." *Bioinformatics*, btx513. doi: 10.1093/bioinformatics/btx513.

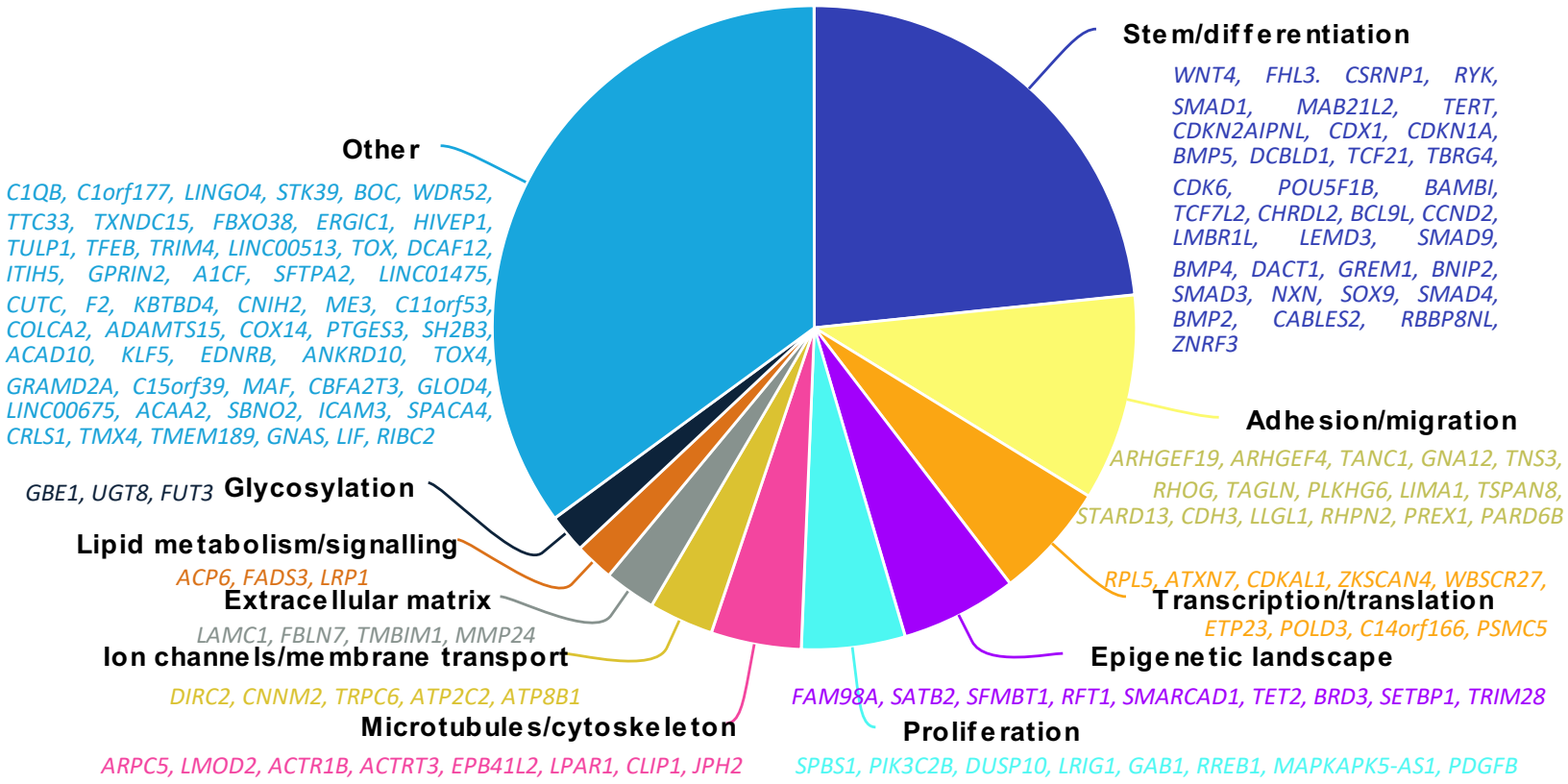
35 . Zhou W, Laird PW, Shen H. Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. Nucleic Acids Res. 2017;45(4):e22.

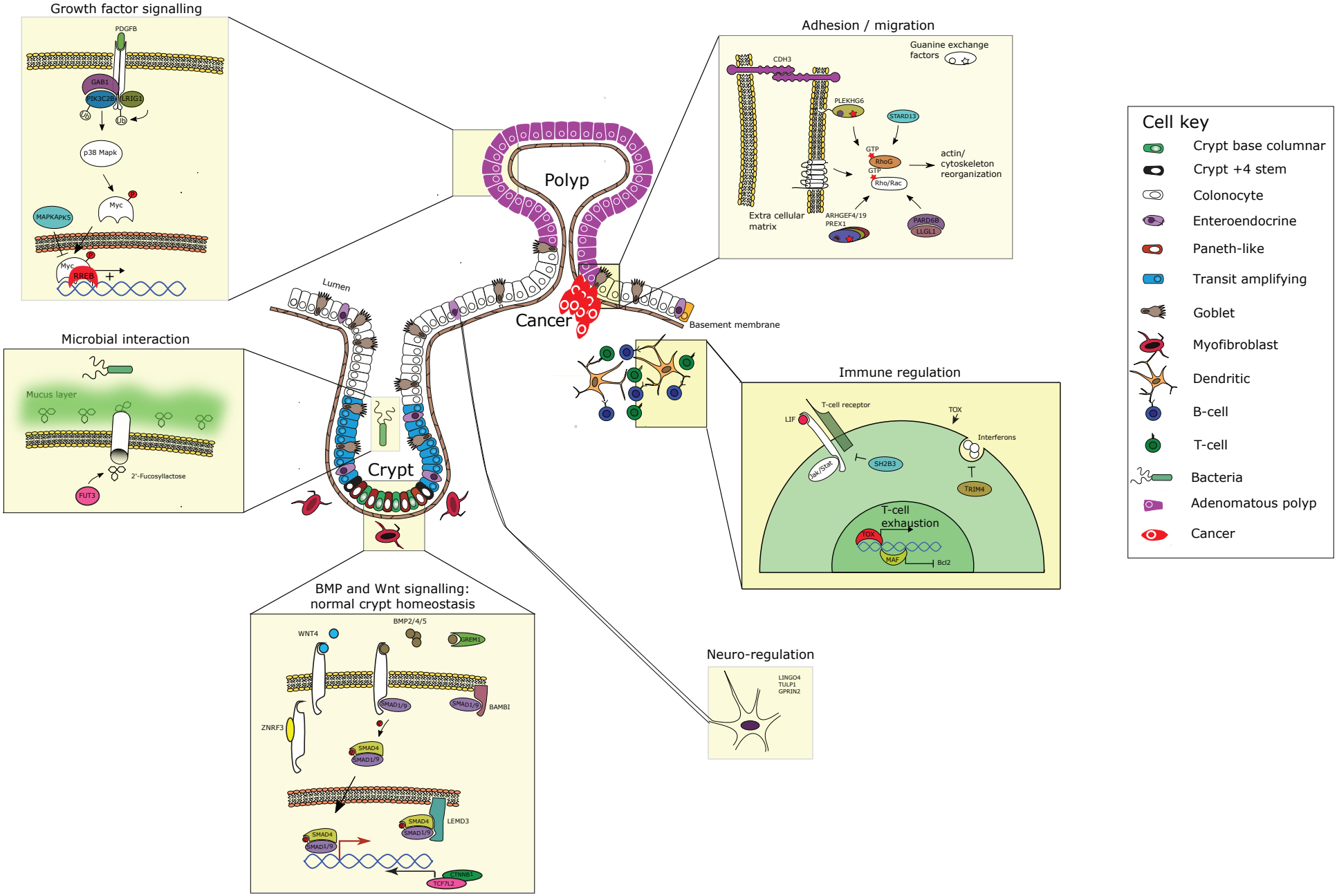
36 . Dong X, Su YR, Barfield R, et al. A general framework for functionally informed set-based analysis: Application to a large-scale colorectal cancer study. PLoS Genet. 2020;16(8):e1008947.

37 . Le Borgne F, Combescure C, Gillaizeau F, et al. Standardized and weighted time-dependent receiver operating characteristic curves to evaluate the intrinsic prognostic capacities of a marker by taking into account confounding factors. Statistical Methods in Medical Research. 2018;27(11):3397-3410. doi:10.1177/0962280217702416]

1003


GWAS
100,204 cases
154,587 controls
Total independent CRC risk loci identified to date
37 new GWAS loci
13 new conditional SNPs
205 risk SNPs
Colorectal mucosa expression reference set – 1,107 samples
15 new TWAS loci
8 new TIsWAS loci
23 risk loci
GTEx multi-tissue TWAS reference set 49 tissues – 16,832 samples
23 new cross-tissue loci
23 risk loci
Colorectal mucosa methylation reference set - 488 samples
7 new MWAS CpG clusters
7 risk loci
170 credible effector genes for CRC risk
Functional characterisation
Likely tissue of action
Potential targets for CRC prevention


Stem/differentiation
WNT4, FHL3. CSRNP1, RYK, SMAD1, MAB21L2, TERT, CDKN2AIPNL, CDX1, CDKN1A, BMP5, DCBLD1, TCF21, TBRG4, CDK6, POU5F1B, BAMBI, TCF7L2, CHRDL2, BCL9L, CCND2, LMBR1L, LEMD3, SMAD9, BMP4, DACT1, GREM1, BNIP2, SMAD3, NXN, SOX9, SMAD4, BMP2, CABLES2, RBBP8NL, ZNRF3
Adhesion/migration
ARHGEF19, ARHGEF4, TANC1, GNA12, TNS3, RHOG, TAGLN, PLKHG6, LIMA1, TSPAN8, STARD13, CDH3, LLGL1, RHPN2, PREX1, PARD6B
Transcription/translation
RPL5, ATXN7, CDKAL1, ZKSCAN4, WBSCR27, ETP23, POLD3, C14orf166, PSMC5
Epigenetic landscape
FAM98A, SATB2, SFMBT1, RFT1, SMARCAD1, TET2, BRD3, SETBP1, TRIM28
Proliferation
SPBS1, PIK3C2B, DUSP10, LRIG1, GAB1, RREB1, MAPKAPK5-AS1, PDGFB
Microtubules/cytoskeleton
ARPC5, LMOD2, ACTR1B, ACTRT3, EPB41L2, LPAR1, CLIP1, JPH2
Ion channels/membrane transport
DIRC2, CNNM2, TRPC6, ATP2C2, ATP8B1
Extracellular matrix
LAMC1, FBLN7, TMBIM1, MMP24
Lipid metabolism/signalling
ACP6, FADS3, LRP1
Glycosylation
GBE1, UGT8, FUT3
Other
C1QB, C1orf177, LINGO4, STK39, BOC, WDR52, TTC33, TXNDC15, FBXO38, ERGIC1, HIVEP1, TULP1, TFEB, TRIM4, LINC00513, TOX, DCAF12, ITIH5, GPRIN2, A1CF, SFTPA2, LINC01475, CUTC, F2, KBTBD4, CNIH2, ME3, C11orf53, COLCA2, ADAMTS15, COX14, PTGES3, SH2B3, ACAD10, KLF5, EDNRB, ANKRD10, TOX4, GRAMD2A, C15orf39, MAF, CBFA2T3, GLOD4, LINC00675, ACAA2, SBNO2, ICAM3, SPACA4, CRLS1, TMX4, TMEM189, GNAS, LIF, RIBC2


Growth factor signalling
PDGFB
GAB1
PIK3C2B
LRIG1
p38 Mapk
Myc
MAPKAPK5
RREB
Adhesion / migration
Guanine exchange factors
CDH3
PLEKHG6
STARD13
GTP
RhoG
Rho/Rac
actin/ cytoskeleton reorganization
Extra cellular matrix
ARHGEF4/19
PREX1
PARD6B
LLGL1
Microbial interaction
Mucus layer
2'-Fucosyllactose
FUT3
Polyp
Lumen
Cancer
Basement membrane
Crypt
Immune regulation
T-cell receptor
LIF
TOX
Interferons
Jak/Stat
SH2B3
TRIM4
T-cell exhaustion
MAF
Bcl2
BMP and Wnt signalling: normal crypt homeostasis
WNT4
BMP2/4/5
GREM1
SMAD1/5/9
BAMBI
ZNRF3
SMAD4
LEMD3
CTNNB1
TCF7L2
Neuro-regulation
LINGO4
TULP1
GPRIN2
Cell key
Crypt base columnar
Crypt +4 stem
Colonocyte
Enteroendocrine
Paneth-like
Transit amplifying
Goblet
Myofibroblast
Dendritic
B-cell
T-cell
Bacteria
Adenomatous polyp
Cancer