



Deposited via The University of Leeds.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/id/eprint/142531/>

Version: Accepted Version

---

**Article:**

Huyghe, JR, Bien, SA, Harrison, TA et al. (2019) Discovery of common and rare genetic risk variants for colorectal cancer. *Nature Genetics*, 51 (1). pp. 76-87. ISSN: 1061-4036

<https://doi.org/10.1038/s41588-018-0286-6>

---

© This is a U.S. government work and not under copyright protection in the U.S.; foreign copyright protection may apply 2018. This is an author produced version of a paper published in *Nature Genetics*. Uploaded in accordance with the publisher's self-archiving policy. <https://doi.org/10.1038/s41588-018-0286-6>.

**Reuse**

Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

**Takedown**

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing [eprints@whiterose.ac.uk](mailto:eprints@whiterose.ac.uk) including the URL of the record and the reason for the withdrawal request.

# 1 Discovery of common and rare genetic risk variants for colorectal cancer

2  
3 Jeroen R Huyghe<sup>1\*</sup>, Stephanie A Bien<sup>1\*</sup>, Tabitha A Harrison<sup>1\*</sup>, Hyun Min Kang<sup>2</sup>, Sai Chen<sup>2</sup>,  
4 Stephanie L Schmit<sup>3</sup>, David V Conti<sup>4</sup>, Conghui Qu<sup>1</sup>, Jihyoun Jeon<sup>5</sup>, Christopher K Edlund<sup>4</sup>,  
5 Peyton Greenside<sup>6</sup>, Michael Wainberg<sup>7</sup>, Fredrick R Schumacher<sup>8</sup>, Joshua D Smith<sup>9</sup>, David M  
6 Levine<sup>10</sup>, Sarah C Nelson<sup>10</sup>, Nasa A Sinnott-Armstrong<sup>11</sup>, Demetrius Albanes<sup>12</sup>, M Henar  
7 Alonso<sup>13-15</sup>, Kristin Anderson<sup>16</sup>, Coral Arnau-Collell<sup>17</sup>, Volker Arndt<sup>18</sup>, Christina Bamia<sup>19,20</sup>,  
8 Barbara L Banbury<sup>1</sup>, John A Baron<sup>21</sup>, Sonja I Berndt<sup>12</sup>, Stéphane Bézieau<sup>22</sup>, D Timothy Bishop<sup>23</sup>,  
9 Juergen Boehm<sup>24</sup>, Heiner Boeing<sup>25</sup>, Hermann Brenner<sup>18,26,27</sup>, Stefanie Brezina<sup>28</sup>, Stephan Buch<sup>29</sup>,  
10 Daniel D Buchanan<sup>30-32</sup>, Andrea Burnett-Hartman<sup>33</sup>, Katja Butterbach<sup>18</sup>, Bette J Caan<sup>34</sup>, Peter T  
11 Campbell<sup>35</sup>, Christopher S Carlson<sup>1,36</sup>, Sergi Castellví-Bel<sup>17</sup>, Andrew T Chan<sup>37-42</sup>, Jenny Chang-  
12 Claude<sup>43,44</sup>, Stephen J Chanock<sup>12</sup>, Maria-Dolores Chirlaque<sup>14,45</sup>, Sang Hee Cho<sup>46</sup>, Charles M  
13 Connolly<sup>1</sup>, Amanda J Cross<sup>47,48</sup>, Katarina Cuk<sup>18</sup>, Keith R Curtis<sup>1</sup>, Albert de la Chapelle<sup>49</sup>,  
14 Kimberly F Doheny<sup>50</sup>, David Duggan<sup>51</sup>, Douglas F Easton<sup>52,53</sup>, Sjoerd G Elias<sup>54</sup>, Faye Elliott<sup>23</sup>,  
15 Dallas R English<sup>55,56</sup>, Edith JM Feskens<sup>57</sup>, Jane C Figueiredo<sup>58,59</sup>, Rocky Fischer<sup>60</sup>, Liesel M  
16 FitzGerald<sup>56,61</sup>, David Forman<sup>62</sup>, Manish Gala<sup>37,39</sup>, Steven Gallinger<sup>63</sup>, W James Gauderman<sup>4</sup>,  
17 Graham G Giles<sup>55,56</sup>, Elizabeth Gillanders<sup>64</sup>, Jian Gong<sup>1</sup>, Phyllis J Goodman<sup>65</sup>, William M  
18 Grady<sup>66</sup>, John S Grove<sup>67</sup>, Andrea Gsur<sup>28</sup>, Marc J Gunter<sup>68</sup>, Robert W Haile<sup>69</sup>, Jochen Hampe<sup>29</sup>,  
19 Heather Hampel<sup>70</sup>, Sophia Harlid<sup>71</sup>, Richard B Hayes<sup>72</sup>, Philipp Hofer<sup>28</sup>, Michael Hoffmeister<sup>18</sup>,  
20 John L Hopper<sup>55,73</sup>, Wan-Ling Hsu<sup>10</sup>, Wen-Yi Huang<sup>12</sup>, Thomas J Hudson<sup>74</sup>, David J Hunter<sup>41,75</sup>,  
21 Gemma Ibañez-Sanz<sup>13,76,77</sup>, Gregory E Idos<sup>4</sup>, Roxann Ingersoll<sup>50</sup>, Rebecca D Jackson<sup>78</sup>, Eric J  
22 Jacobs<sup>35</sup>, Mark A Jenkins<sup>55</sup>, Amit D Joshi<sup>39,41</sup>, Corinne E Joshi<sup>79</sup>, Temitope O Keku<sup>80</sup>, Timothy J  
23 Key<sup>81</sup>, Hyeong Rok Kim<sup>82</sup>, Emiko Kobayashi<sup>1</sup>, Laurence N Kolonel<sup>83</sup>, Charles Kooperberg<sup>1</sup>,  
24 Tilman Kühn<sup>43</sup>, Sébastien Küry<sup>22</sup>, Sun-Seog Kweon<sup>84,85</sup>, Susanna C Larsson<sup>86</sup>, Cecelia A  
25 Laurie<sup>10</sup>, Loic Le Marchand<sup>67</sup>, Suzanne M Leal<sup>87</sup>, Soo Chin Lee<sup>88,89</sup>, Flavio Lejbkowitz<sup>90-92</sup>,  
26 Mathieu Lemire<sup>74</sup>, Christopher I Li<sup>1</sup>, Li Li<sup>93</sup>, Wolfgang Lieb<sup>94</sup>, Yi Lin<sup>1</sup>, Annika Lindblom<sup>95,96</sup>,  
27 Noralane M Lindor<sup>97</sup>, Hua Ling<sup>50</sup>, Tin L Louie<sup>10</sup>, Satu Männistö<sup>98</sup>, Sanford D Markowitz<sup>99</sup>,  
28 Vicente Martín<sup>14,100</sup>, Giovanna Masala<sup>101</sup>, Caroline E McNeil<sup>102</sup>, Marilena Melas<sup>4</sup>, Roger L  
29 Milne<sup>55,56</sup>, Lorena Moreno<sup>17</sup>, Neil Murphy<sup>68</sup>, Robin Myte<sup>71</sup>, Alessio Naccarati<sup>103,104</sup>, Polly A  
30 Newcomb<sup>1,36</sup>, Kenneth Offit<sup>105,106</sup>, Shuji Ogino<sup>40,41,107,108</sup>, N Charlotte Onland-Moret<sup>54</sup>, Barbara  
31 Pardini<sup>104,109</sup>, Patrick S Parfrey<sup>110</sup>, Rachel Pearlman<sup>70</sup>, Vittorio Perduca<sup>111,112</sup>, Paul D P  
32 Pharoah<sup>52</sup>, Mila Pinchev<sup>91</sup>, Elizabeth A Platz<sup>79</sup>, Ross L Prentice<sup>1</sup>, Elizabeth Pugh<sup>50</sup>, Leon  
33 Raskin<sup>113</sup>, Gad Rennert<sup>91,92,114</sup>, Hedy S Rennert<sup>91,92,114</sup>, Elio Riboli<sup>115</sup>, Miguel Rodríguez-  
34 Barranco<sup>14,116</sup>, Jane Romm<sup>50</sup>, Lori C Sakoda<sup>1,117</sup>, Clemens Schafmayer<sup>118</sup>, Robert E Schoen<sup>119</sup>,  
35 Daniela Seminara<sup>64</sup>, Mitul Shah<sup>53</sup>, Tameka Shelford<sup>50</sup>, Min-Ho Shin<sup>84</sup>, Katerina Shulman<sup>120</sup>,  
36 Sabina Sieri<sup>121</sup>, Martha L Slattery<sup>122</sup>, Melissa C Southey<sup>123</sup>, Zsafia K Stadler<sup>124</sup>, Christa  
37 Stegmaier<sup>125</sup>, Yu-Ru Su<sup>1</sup>, Catherine M Tangen<sup>65</sup>, Stephen N Thibodeau<sup>126</sup>, Duncan C Thomas<sup>4</sup>,  
38 Sushma S Thomas<sup>1</sup>, Amanda E Toland<sup>127</sup>, Antonia Trichopoulou<sup>19,20</sup>, Cornelia M Ulrich<sup>24</sup>,  
39 David J Van Den Berg<sup>4</sup>, Franzel JB van Duijnhoven<sup>57</sup>, Bethany Van Guelpen<sup>71</sup>, Henk van  
40 Kranen<sup>128</sup>, Joseph Vijai<sup>124</sup>, Kala Visvanathan<sup>79</sup>, Pavel Vodicka<sup>103,129,130</sup>, Ludmila  
41 Vodickova<sup>103,129,130</sup>, Veronika Vymetalkova<sup>103,129,130</sup>, Korbinian Weigl<sup>18,27,131</sup>, Stephanie J  
42 Weinstein<sup>12</sup>, Emily White<sup>1</sup>, Aung Ko Win<sup>32,55</sup>, C Roland Wolf<sup>132</sup>, Alicja Wolk<sup>86,133</sup>, Michael O  
43 Woods<sup>134</sup>, Anna H Wu<sup>4</sup>, Syed H Zaidi<sup>74</sup>, Brent W Zanke<sup>135</sup>, Qing Zhang<sup>136</sup>, Wei Zheng<sup>137</sup>, Peter  
44 C Scacheri<sup>138</sup>, John D Potter<sup>1</sup>, Michael C Bassik<sup>11</sup>, Anshul Kundaje<sup>7,11</sup>, Graham Casey<sup>139</sup>, Victor  
45 Moreno<sup>13-15,77</sup>, Goncalo R Abecasis<sup>2</sup>, Deborah A Nickerson<sup>9</sup>§, Stephen B Gruber<sup>4</sup>§, Li Hsu<sup>1,10</sup>§,  
46 Ulrike Peters<sup>1,36</sup>§

- 47
- 48 1. Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle,  
49 Washington, USA.
- 50 2. Department of Biostatistics and Center for Statistical Genetics, University of Michigan,  
51 Ann Arbor, Michigan, USA.
- 52 3. Department of Cancer Epidemiology, H. Lee Moffitt Cancer Center and Research Institute,  
53 Tampa, Florida, USA.
- 54 4. Department of Preventive Medicine, USC Norris Comprehensive Cancer Center, Keck  
55 School of Medicine, University of Southern California, Los Angeles, California, USA.
- 56 5. Department of Epidemiology, University of Michigan, Ann Arbor, Michigan, USA.
- 57 6. Biomedical Informatics Program, Stanford University, Stanford, California, USA.
- 58 7. Department of Computer Science, Stanford University, Stanford, California, USA.
- 59 8. Department of Population and Quantitative Health Sciences, Case Western Reserve  
60 University, Cleveland, Ohio, USA.
- 61 9. Department of Genome Sciences, University of Washington, Seattle, Washington, USA.
- 62 10. Department of Biostatistics, University of Washington, Seattle, Washington, USA.
- 63 11. Department of Genetics, Stanford University, Stanford, California, USA.
- 64 12. Division of Cancer Epidemiology and Genetics, National Cancer Institute, National  
65 Institutes of Health, Bethesda, Maryland, USA.
- 66 13. Cancer Prevention and Control Program, Catalan Institute of Oncology-IDIBELL,  
67 L'Hospitalet de Llobregat, Barcelona, Spain.
- 68 14. CIBER de Epidemiología y Salud Pública (CIBERESP), Madrid, Spain.
- 69 15. Department of Clinical Sciences, Faculty of Medicine, University of Barcelona, Barcelona,  
70 Spain.
- 71 16. Division of Epidemiology and Community Health, University of Minnesota, Minneapolis,  
72 Minnesota, USA.
- 73 17. Gastroenterology Department, Hospital Clínic, Institut d'Investigacions Biomèdiques  
74 August Pi i Sunyer (IDIBAPS), Centro de Investigación Biomédica en Red de  
75 Enfermedades Hepáticas y Digestivas (CIBEREHD), University of Barcelona, Barcelona,  
76 Spain.
- 77 18. Division of Clinical Epidemiology and Aging Research, German Cancer Research Center  
78 (DKFZ), Heidelberg, Germany.
- 79 19. Hellenic Health Foundation, Athens, Greece.
- 80 20. WHO Collaborating Center for Nutrition and Health, Unit of Nutritional Epidemiology and  
81 Nutrition in Public Health, Department of Hygiene, Epidemiology and Medical Statistics,  
82 School of Medicine, National and Kapodistrian University of Athens, Greece.
- 83 21. Department of Medicine, University of North Carolina School of Medicine, Chapel Hill,  
84 North Carolina, USA.
- 85 22. Service de Génétique Médicale, Centre Hospitalier Universitaire (CHU) Nantes, Nantes,  
86 France.
- 87 23. Leeds Institute of Medical Research at St James's, University of Leeds, Leeds, UK.
- 88 24. Huntsman Cancer Institute and Department of Population Health Sciences, University of  
89 Utah, Salt Lake City, Utah, USA.
- 90 25. Department of Epidemiology, German Institute of Human Nutrition (DIfE), Potsdam-  
91 Rehbrücke, Germany.

- 92 26. Division of Preventive Oncology, German Cancer Research Center (DKFZ) and National  
93 Center for Tumor Diseases (NCT), Heidelberg, Germany.
- 94 27. German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ),  
95 Heidelberg, Germany.
- 96 28. Institute of Cancer Research, Department of Medicine I, Medical University of Vienna,  
97 Vienna, Austria.
- 98 29. Department of Medicine I, University Hospital Dresden, Technische Universität Dresden  
99 (TU Dresden), Dresden, Germany.
- 100 30. Colorectal Oncogenomics Group, Department of Clinical Pathology, The University of  
101 Melbourne, Parkville, Victoria, Australia.
- 102 31. University of Melbourne Centre for Cancer Research, Victorian Comprehensive Cancer  
103 Centre, Parkville, Victoria, Australia.
- 104 32. Genomic Medicine and Family Cancer Clinic, Royal Melbourne Hospital, Parkville,  
105 Victoria, Australia.
- 106 33. Institute for Health Research, Kaiser Permanente Colorado, Denver, Colorado, USA.
- 107 34. Division of Research, Kaiser Permanente Medical Care Program, Oakland, California,  
108 USA.
- 109 35. Behavioral and Epidemiology Research Group, American Cancer Society, Atlanta,  
110 Georgia, USA.
- 111 36. Department of Epidemiology, University of Washington, Seattle, Washington, USA.
- 112 37. Division of Gastroenterology, Massachusetts General Hospital and Harvard Medical  
113 School, Boston, Massachusetts, USA.
- 114 38. Channing Division of Network Medicine, Brigham and Women's Hospital and Harvard  
115 Medical School, Boston, Massachusetts, USA.
- 116 39. Clinical and Translational Epidemiology Unit, Massachusetts General Hospital and  
117 Harvard Medical School, Boston, Massachusetts, USA.
- 118 40. Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA.
- 119 41. Department of Epidemiology, Harvard T.H. Chan School of Public Health, Harvard  
120 University, Boston, Massachusetts, USA.
- 121 42. Department of Immunology and Infectious Diseases, Harvard T.H. Chan School of Public  
122 Health, Harvard University, Boston, Massachusetts, USA.
- 123 43. Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg,  
124 Germany.
- 125 44. Cancer Epidemiology Group, University Medical Centre Hamburg-Eppendorf, University  
126 Cancer Centre Hamburg (UCCH), Hamburg, Germany.
- 127 45. Department of Epidemiology, Regional Health Council, IMIB-Arrixaca, Murcia  
128 University, Murcia, Spain.
- 129 46. Department of Hematology-Oncology, Chonnam National University Hospital, Hwasun,  
130 South Korea.
- 131 47. Department of Epidemiology and Biostatistics, Imperial College London, London, UK.
- 132 48. Department of Surgery and Cancer, Imperial College London, London, UK. .
- 133 49. Department of Cancer Biology and Genetics and the Comprehensive Cancer Center, The  
134 Ohio State University, Columbus, Ohio, USA.
- 135 50. Center for Inherited Disease Research (CIDR), Institute of Genetic Medicine, Johns  
136 Hopkins University, Baltimore, Maryland, USA.

- 137 51. Translational Genomics Research Institute - An Affiliate of City of Hope, Phoenix,  
138 Arizona, USA.
- 139 52. Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK.
- 140 53. Centre for Cancer Genetic Epidemiology, Department of Oncology, University of  
141 Cambridge, Cambridge, UK.
- 142 54. Julius Center for Health Sciences and Primary Care, University Medical Center  
143 Utrecht, Utrecht University, Utrecht, The Netherlands.
- 144 55. Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global  
145 Health, The University of Melbourne, Melbourne, Victoria, Australia.
- 146 56. Cancer Epidemiology and Intelligence Division, Cancer Council Victoria, Melbourne,  
147 Victoria, Australia.
- 148 57. Division of Human Nutrition and Health, Wageningen University and Research,  
149 Wageningen, The Netherlands.
- 150 58. Department of Medicine, Samuel Oschin Comprehensive Cancer Institute, Cedars-Sinai  
151 Medical Center, Los Angeles, California, USA.
- 152 59. Department of Preventive Medicine, Keck School of Medicine, University of Southern  
153 California, Los Angeles, California, USA.
- 154 60. University of Michigan Comprehensive Cancer Center, Ann Arbor, Michigan, USA.
- 155 61. Menzies Institute for Medical Research, University of Tasmania, Hobart, Tasmania,  
156 Australia.
- 157 62. International Agency for Research on Cancer, World Health Organization, Lyon, France.
- 158 63. Lunenfeld Tanenbaum Research Institute, Mount Sinai Hospital, University of Toronto,  
159 Toronto, Ontario, Canada.
- 160 64. Division of Cancer Control and Population Sciences, National Cancer Institute, Bethesda,  
161 Maryland, USA.
- 162 65. SWOG Statistical Center, Fred Hutchinson Cancer Research Center, Seattle, Washington,  
163 USA.
- 164 66. Clinical Research Division, Fred Hutchinson Cancer Research Center, Seattle,  
165 Washington, USA.
- 166 67. University of Hawaii Cancer Research Center, Honolulu, Hawaii, USA.
- 167 68. Nutrition and Metabolism Section, International Agency for Research on Cancer, World  
168 Health Organization, Lyon, France.
- 169 69. Division of Oncology, Department of Medicine, Stanford University, Stanford, California,  
170 USA.
- 171 70. Division of Human Genetics, Department of Internal Medicine, The Ohio State University  
172 Comprehensive Cancer Center, Columbus, Ohio, USA.
- 173 71. Department of Radiation Sciences, Oncology Unit, Umeå University, Umeå, Sweden.
- 174 72. Division of Epidemiology, Department of Population Health, New York University School  
175 of Medicine, New York, New York, USA.
- 176 73. Department of Epidemiology, School of Public Health and Institute of Health and  
177 Environment, Seoul National University, Seoul, South Korea.
- 178 74. Ontario Institute for Cancer Research, Toronto, Ontario, Canada.
- 179 75. Nuffield Department of Population Health, University of Oxford, Oxford, UK.
- 180 76. Gastroenterology Department, Bellvitge University Hospital, L'Hospitalet de Llobregat,  
181 Barcelona, Spain.

- 182 77. Colorectal Cancer Group, ONCOBELL Program, Bellvitge Biomedical Research Institute-  
183 IDIBELL, Hospitalet de Llobregat, Barcelona, Spain.
- 184 78. Department of Medicine, Division of Endocrinology, Diabetes and Metabolism, The Ohio  
185 State University, Columbus, Ohio, USA.
- 186 79. Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Johns  
187 Hopkins University, Baltimore, Maryland, USA.
- 188 80. Center for Gastrointestinal Biology and Disease, University of North Carolina, Chapel Hill,  
189 North Carolina, USA.
- 190 81. Cancer Epidemiology Unit, Nuffield Department of Population Health, University of  
191 Oxford, Oxford, UK.
- 192 82. Department of Surgery, Chonnam National University Hwasun Hospital and Medical  
193 School, Hwasun, Korea.
- 194 83. Office of Public Health Studies, University of Hawaii Manoa, Honolulu, Hawaii, USA.
- 195 84. Department of Preventive Medicine, Chonnam National University Medical School,  
196 Gwangju, Korea.
- 197 85. Jeonnam Regional Cancer Center, Chonnam National University Hwasun Hospital,  
198 Hwasun, Korea.
- 199 86. Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden.
- 200 87. Center for Statistical Genetics, Department of Molecular and Human Genetics, Baylor  
201 College of Medicine, Houston, Texas, USA.
- 202 88. Department of Haematology-Oncology, National University Cancer Institute, Singapore.
- 203 89. Cancer Science Institute of Singapore, National University of Singapore, Singapore.
- 204 90. The Clalit Health Services, Personalized Genomic Service, Carmel, Haifa, Israel.
- 205 91. Department of Community Medicine and Epidemiology, Lady Davis Carmel Medical  
206 Center, Haifa, Israel.
- 207 92. Clalit National Cancer Control Center, Haifa, Israel.
- 208 93. Center for Community Health Integration and Case Comprehensive Cancer Center, Case  
209 Western Reserve University, Cleveland, Ohio, USA.
- 210 94. Institute of Epidemiology, PopGen Biobank, Christian-Albrechts-University Kiel, Kiel,  
211 Germany.
- 212 95. Department of Clinical Genetics, Karolinska University Hospital, Stockholm, Sweden.
- 213 96. Department of Molecular Medicine and Surgery, Karolinska Institutet, Stockholm,  
214 Sweden.
- 215 97. Department of Health Science Research, Mayo Clinic, Scottsdale, Arizona, USA.
- 216 98. Department of Public Health Solutions, National Institute for Health and Welfare, Helsinki,  
217 Finland.
- 218 99. Departments of Medicine and Genetics, Case Comprehensive Cancer Center, Case Western  
219 Reserve University, and University Hospitals of Cleveland, Cleveland, Ohio, USA.
- 220 100. Biomedicine Institute (IBIOMED), University of León, León, Spain.
- 221 101. Cancer Risk Factors and Life-Style Epidemiology Unit, Institute of Cancer Research,  
222 Prevention and Clinical Network - ISPRO, Florence, Italy.
- 223 102. USC Norris Comprehensive Cancer Center, University of Southern California, Los  
224 Angeles, California, USA.
- 225 103. Department of Molecular Biology of Cancer, Institute of Experimental Medicine of the  
226 Czech Academy of Sciences, Prague, Czech Republic.
- 227 104. Italian Institute for Genomic Medicine (IIGM), Turin, Italy.

- 228 105. Clinical Genetics Service, Department of Medicine, Memorial Sloan Kettering Cancer  
229 Center, New York, New York, USA.
- 230 106. Department of Medicine, Weill Cornell Medical College, New York, New York, USA.
- 231 107. Program in MPE Molecular Pathological Epidemiology, Department of Pathology,  
232 Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA.
- 233 108. Department of Oncologic Pathology, Dana-Farber Cancer Institute, Boston, Massachusetts,  
234 USA.
- 235 109. Department of Medical Sciences, University of Turin, Turin, Italy.
- 236 110. The Clinical Epidemiology Unit, Memorial University Medical School, Newfoundland,  
237 Canada.
- 238 111. Laboratoire de Mathématiques Appliquées MAP5 (UMR CNRS 8145), Université Paris  
239 Descartes, Paris, France.
- 240 112. CESP (Inserm U1018), Facultés de Médecine Université Paris-Sud, UVSQ, Université  
241 Paris-Saclay, Gustave Roussy, Villejuif, France.
- 242 113. Division of Epidemiology, Vanderbilt Epidemiology Center, Vanderbilt University School  
243 of Medicine, Nashville, Tennessee, USA.
- 244 114. Ruth and Bruce Rappaport Faculty of Medicine, Technion-Israel Institute of Technology,  
245 Haifa, Israel.
- 246 115. School of Public Health, Imperial College London, London, UK.
- 247 116. Escuela Andaluza de Salud Pública. Instituto de Investigación Biosanitaria  
248 ibs.GRANADA. Hospitales Universitarios de Granada/Universidad de Granada, Granada,  
249 Spain.
- 250 117. Division of Research, Kaiser Permanente Northern California, Oakland, California, USA.
- 251 118. Department of General and Thoracic Surgery, University Hospital Schleswig-Holstein,  
252 Campus Kiel, Kiel, Germany.
- 253 119. Department of Medicine and Epidemiology, University of Pittsburgh Medical Center,  
254 Pittsburgh, Pennsylvania, USA.
- 255 120. Oncology Unit, Hillel Yaffe Medical Center, Hadera, Israel.
- 256 121. Epidemiology and Prevention Unit, Fondazione IRCCS Istituto Nazionale dei Tumori,  
257 Milan, Italy.
- 258 122. Department of Internal Medicine, University of Utah, Salt Lake City, Utah, USA.
- 259 123. Genetic Epidemiology Laboratory, Department of Pathology, The University of  
260 Melbourne, Melbourne, Australia.
- 261 124. Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, New York,  
262 USA.
- 263 125. Saarland Cancer Registry, Saarbrücken, Germany.
- 264 126. Division of Laboratory Genetics, Department of Laboratory Medicine and Pathology,  
265 Mayo Clinic, Rochester, Minnesota, USA.
- 266 127. Departments of Cancer Biology and Genetics and Internal Medicine, Comprehensive  
267 Cancer Center, The Ohio State University, Columbus, Ohio, USA.
- 268 128. National Institute for Public Health and the Environment (RIVM), Bilthoven, The  
269 Netherlands.
- 270 129. Institute of Biology and Medical Genetics, First Faculty of Medicine, Charles University,  
271 Prague, Czech Republic.
- 272 130. Faculty of Medicine and Biomedical Center in Pilsen, Charles University, Pilsen, Czech  
273 Republic.

- 274 131. Medical Faculty, University of Heidelberg, Germany.  
275 132. School of Medicine, University of Dundee, Dundee, Scotland.  
276 133. Department of Surgical Sciences, Uppsala University, Uppsala, Sweden.  
277 134. Memorial University of Newfoundland, Discipline of Genetics, St. John's, Canada.  
278 135. Division of Hematology, University of Toronto, Toronto, Ontario, Canada.  
279 136. Genomics Shared Resource, Fred Hutchinson Cancer Research Center, Seattle,  
280 Washington, USA.  
281 137. Division of Epidemiology, Department of Medicine, Vanderbilt-Ingram Cancer Center,  
282 Vanderbilt Epidemiology Center, Vanderbilt University School of Medicine, Nashville,  
283 Tennessee, USA.  
284 138. Department of Genetics and Genome Sciences, Case Western Reserve University School  
285 of Medicine, Case Comprehensive Cancer Center, Cleveland, Ohio, USA.  
286 139. Center for Public Health Genomics, University of Virginia, Charlottesville, Virginia, USA.  
287

288 \*These authors contributed equally to this work.

289 §These authors jointly supervised this work.

290 Correspondence should be addressed to U.P. (upeters@fredhutch.org).

291

292 **To further dissect the genetic architecture of colorectal cancer (CRC), we performed**  
293 **whole-genome sequencing of 1,439 cases and 720 controls, imputed discovered sequence**  
294 **variants and Haplotype Reference Consortium panel variants into genome-wide association**  
295 **study data, and tested for association in 34,869 cases and 29,051 controls. Findings were**  
296 **followed up in an additional 23,262 cases and 38,296 controls. We discovered a strongly**  
297 **protective 0.3% frequency variant signal at *CHD1*. In a combined meta-analysis of 125,478**  
298 **individuals, we identified 40 new independent signals at  $P < 5 \times 10^{-8}$ , bringing the number of**  
299 **known independent signals for CRC to approximately 100. New signals implicate lower-**  
300 **frequency variants, Krüppel-like factors, Hedgehog signaling, Hippo-YAP signaling, long**  
301 **noncoding RNAs, somatic drivers, and support a role of immune function. Heritability**  
302 **analyses suggest that CRC risk is highly polygenic, and larger, more comprehensive studies**  
303 **enabling rare variant analysis will improve understanding of underlying biology, and**  
304 **impact personalized screening strategies and drug development.**

305

306 Colorectal cancer (CRC) is the fourth leading cancer-related cause of death worldwide<sup>1</sup> and  
307 presents a major public health burden. Up to 35% of inter-individual variability in CRC risk has  
308 been attributed to genetic factors<sup>2,3</sup>. Family-based studies have identified rare high-penetrance  
309 mutations in at least a dozen genes but, collectively, these account for only a small fraction of

310 familial risk<sup>4</sup>. Over the past decade, genome-wide association studies (GWAS) for sporadic  
311 CRC, which constitutes the majority of cases, have identified approximately 60 association  
312 signals at over 50 loci<sup>5-22</sup>. Yet, most of the genetic factors contributing to CRC risk remain  
313 undefined. This severely hampers our understanding of biological processes underlying CRC. It  
314 also limits CRC precision prevention, including individualized preventive screening  
315 recommendations and development of cancer prevention drugs. The contribution of rare  
316 variation to sporadic CRC is particularly poorly understood.

317

318 To expand the catalog of CRC risk loci and improve our understanding of rare variants, genes,  
319 and pathways influencing sporadic CRC risk, and risk prediction, we performed the largest and  
320 most comprehensive whole-genome sequencing (WGS) study and GWAS meta-analysis for  
321 CRC to date, combining data from three consortia: the Genetics and Epidemiology of Colorectal  
322 Cancer Consortium (GECCO), the Colorectal Cancer Transdisciplinary Study (CORECT), and  
323 the Colon Cancer Family Registry (CCFR). Our study almost doubles the number of individuals  
324 analyzed, incorporating GWAS results from >125,000 individuals, and substantially expands and  
325 strengthens our understanding of biological processes underlying CRC risk.

326

## 327 **RESULTS**

### 328 **Study Overview**

329 We performed WGS of 1,439 CRC cases and 720 controls of European ancestry at low coverage  
330 (3.8-8.6×). We detected, called, and estimated haplotype phase for 31.8 million genetic variants,  
331 including 1.7 million short insertion-deletion variants (indels) (Online Methods). These data  
332 include many rare variants not studied by GWAS. Based on other large-scale WGS studies  
333 employing a similar design, we expected to have near-complete ascertainment of single  
334 nucleotide variants (SNVs) with minor allele count (MAC) greater than five (minor allele  
335 frequency (MAF) >0.1%), and high accuracy at heterozygous genotypes<sup>23,24</sup>. We tested 14.4  
336 million variants with MAC ≥5 for CRC association using logistic regression (Online Methods)  
337 but did not find any significant associations. To increase power to detect associations with rare  
338 and low-frequency variants of modest effect, we imputed variants from the sequencing  
339 experiment into 34,869 cases and 29,051 controls of predominantly European (91.7%) and East  
340 Asian ancestry (8.3%) from 30 existing GWAS studies (Online Methods and **Supplementary**

341 **Table 1**). By design, two thirds of sequenced individuals were CRC cases, thereby enriching the  
342 panel for rare or low-frequency alleles that increase CRC risk. We contributed our sequencing  
343 data to the Haplotype Reference Consortium (HRC)<sup>25</sup> and imputed the 30 existing GWAS  
344 studies to the HRC panel, which comprises haplotypes for 32,488 individuals. Results of these  
345 GWAS meta-analyses (referred to as Stage 1 meta-analysis; Online Methods) informed the  
346 design of a custom Illumina array comprising the OncoArray, a custom array to identify cancer  
347 risk loci<sup>26</sup>, and 15,802 additional variants selected based on Stage 1 meta-analysis results. We  
348 genotyped 12,007 cases and 12,000 controls of European ancestry with this custom array, and  
349 combined them with an additional 11,255 cases and 26,296 controls with GWAS data, resulting  
350 in a Stage 2 meta-analysis of 23,262 CRC cases and 38,296 controls (Online Methods,  
351 **Supplementary Fig. 1**, and **Supplementary Table 1**). Next, we performed a combined (Stage 1  
352 + Stage 2) meta-analysis of up to 58,131 cases and 67,347 controls. This meta-analysis was  
353 based on the HRC-panel-imputed data because, given its large size, this panel results in superior  
354 imputation quality and enables accurate imputation of variants with MAFs as low as 0.1%<sup>25</sup>.  
355 Here, we report new association signals discovered through our custom genotyping experiment  
356 and replicating in Stage 2 at the Bonferroni significance threshold of  $P < 7.8 \times 10^{-6}$  (Online  
357 Methods), as well as distinct association signals passing the genome-wide significance (GWS)  
358 threshold of  $P < 5 \times 10^{-8}$  in the combined meta-analysis of up to 125,478 individuals.

359

### 360 **CRC risk loci**

361 In the combined meta-analysis, we identified 30 new CRC risk loci reaching GWS and >500kb  
362 away from previously reported CRC risk variants (**Table 1**; **Supplementary Fig. 2 and 3**).  
363 Twenty-two of these were represented on our custom genotyping panel, either by the lead variant  
364 (15 loci) or by a variant in linkage disequilibrium (LD) ( $r^2 > 0.7$ ). Of these 22 variants,  
365 eight attained the Bonferroni significance threshold in the Stage 2 meta-analysis (**Table 1**).

366

367 Among these eight loci is the first rare variant signal identified for sporadic CRC, involving five  
368 0.3% frequency variants at 5q21.1, near genes *CHDI* and *RGMB*. SNP rs145364999, intronic to  
369 *CHDI*, had high quality genotyping (**Supplementary Fig. 4**). The variant was well imputed in  
370 the remaining sample sets (imputation quality  $r^2$  ranged from 0.66 to 0.87; **Supplementary**  
371 **Table 2**) and there was no evidence of heterogeneity of effects (heterogeneity  $P=0.63$ ;

372 **Supplementary Table 2).** The rare allele confers a strong protective effect (allelic odds ratio  
373 (OR)=0.52 in Stage 2; 95% confidence interval (CI)=0.40-0.68). Chromatin remodeling factor  
374 CHD1 provides an especially plausible candidate and has been shown to be a synthetically-  
375 essential gene<sup>27</sup> that is occasionally deleted in some cancers, but always retained in PTEN-  
376 deficient cancers<sup>28</sup>. The resulting mutually exclusive deletion pattern of *CHD1* and *PTEN* has  
377 been observed in prostate, breast, and CRC TCGA data<sup>28</sup>. We hypothesize that the rare allele  
378 confers a protective effect through lowering *CHD1* expression, which is required for nuclear  
379 factor- $\kappa\beta$  (NF- $\kappa\beta$ ) pathway activation and growth in cancer cells driven by loss of the tumor  
380 suppressor *PTEN*<sup>28</sup>. However, we cannot rule out involvement of nearby candidate gene *RGMB*  
381 that encodes a co-receptor for bone morphogenetic proteins BMP2 and BMP4, both of which are  
382 linked to CRC risk through GWAS<sup>9,11</sup>. Additionally, *RGMB* has been shown to bind to PD-L2<sup>29</sup>,  
383 a known ligand of PD-1, an immune checkpoint blockade inhibitor targeted by cancer  
384 immunotherapy<sup>30</sup>.

385  
386 The vast majority of new association signals involve common variants. We found associations  
387 near strong candidate genes for CRC risk in pathways or gene families not previously implicated  
388 by GWAS. Locus 13q22.1, represented by lead SNP rs78341008 (MAF 7.2%;  $P=3.2\times 10^{-10}$ ), is  
389 near *KLF5*, a known CRC oncogene that can be activated by somatic hotspot mutations or super-  
390 enhancer duplications<sup>31,32</sup>. *KLF5* encodes transcription factor Krüppel-like factor 5 (KLF5),  
391 which promotes cell proliferation and is highly expressed in intestinal crypt stem cells. We also  
392 found an association at 19p13.11, near *KLF2*. *KLF2* expression in endothelial cells is critical for  
393 normal blood vessel function<sup>33,34</sup>. Down-regulated *KLF2* expression in colon tumor tissues  
394 contributes to structurally and functionally abnormal tumor blood vessels, resulting in impaired  
395 blood flow and hypoxia in tumors<sup>35</sup>. Another locus at 9q31.1 is near *LPARI*, which encodes a  
396 receptor for lysophosphatidic acid (LPA). LPA-induced expression of hypoxia-inducible factor 1  
397 (HIF-1 $\alpha$ ), a key regulator of cellular adaptation to hypoxia and tumorigenesis, depends on  
398 *KLF5*<sup>36</sup>. Additionally, LPA activates multiple signaling pathways and stimulates proliferation of  
399 colon cancer cells by activation of *KLF5*<sup>37</sup>. Another locus (7p13) is near *SNHG15*, encoding a  
400 long non-coding RNA (lncRNA) that epigenetically represses *KLF2* to promote pancreatic  
401 cancer proliferation<sup>38</sup>.

402

403 We found two loci near members of the Hedgehog (Hh) signaling pathway. Aberrant activation  
404 of this pathway, caused by somatic mutations or changes in expression, can drive tumorigenesis  
405 in many tumors<sup>39</sup>. Notably, downregulated stromal cell Hh signaling reportedly accelerates  
406 colonic tumorigenesis in mice<sup>40</sup>. Locus 3q13.2, represented by low-frequency lead SNP  
407 rs72942485 (MAF 2.2%;  $P=2.1\times 10^{-8}$ ), overlaps *BOC*, encoding a Hh coreceptor molecule. In  
408 medulloblastoma, upregulated *BOC* promotes Hh-driven tumor progression through Cyclin D1-  
409 induced DNA damage<sup>41</sup>. In pancreatic cancer, a complex role for stromal *BOC* expression in  
410 tumorigenesis and angiogenesis has been reported<sup>42</sup>. Locus 4q31.21 is near *HHIP*, encoding an  
411 inhibitor of Hh signaling. Of note, the Hh signaling pathway was also significantly enriched in  
412 our pathway analysis (described below).

413  
414 Locus 11q22.1 is near *YAP1*, which encodes a critical downstream regulatory target in the Hippo  
415 signaling pathway that is gaining recognition as a pivotal player in organ size control and  
416 tumorigenesis<sup>43</sup>. *YAP1* is highly expressed in intestinal crypt stem cells, and in transgenic mice,  
417 overexpression resulted in severe intestinal dysplasia and loss of differentiated cell types<sup>44</sup>,  
418 reminiscent of phenotypes observed in mice and humans with deleterious germline *APC*  
419 mutations. Further, Hypoxia-inducible factor 2 $\alpha$  (HIF-2 $\alpha$ ) promotes colon cancer growth by up-  
420 regulating YAP1 activity<sup>45</sup>.

421  
422 We provide further evidence for a link between immune function and CRC pathogenesis, and  
423 implicate the major histocompatibility complex (MHC) in CRC risk. We identified a locus near  
424 genes *HLA-DRB1/HLA-DQA1*, which is associated with immune-mediated diseases<sup>46</sup>.

425  
426 We identified two new loci near known tumor suppressor genes. Locus 4q24 is near *TET2*, a  
427 chromatin-remodeling gene frequently somatically mutated in multiple cancers, including colon  
428 cancer<sup>47</sup>, and overlapping GWAS signals for multiple other cancers<sup>48-50</sup>. The *CDKN2B-*  
429 *CDKN2A-ANRIL* locus at 9p21.3 is a well-established hot spot of pleiotropic GWAS  
430 associations for many complex diseases including coronary artery disease<sup>51</sup>, type 2 diabetes<sup>52</sup>,  
431 and cancers<sup>50,53,54-56</sup>. Interestingly, lead variant rs1537372 is in high LD ( $r^2=0.82$ ) with variants  
432 associated with coronary artery disease<sup>51</sup> and endometriosis<sup>57</sup>, but not with the other cancer-  
433 associated variants. *CDKN2A/B* encode cyclin-dependent kinase inhibitors that regulate the cell

434 cycle. *CDKN2A* is one of the most commonly inactivated genes in cancer, and is a high  
435 penetrance gene for melanoma<sup>58,59</sup>. *CDKN2B* activation is tightly controlled by the cytokine  
436 TGF- $\beta$ , further linking this signaling pathway with CRC tumorigenesis<sup>60</sup>.

437  
438 Our findings implicate genes in pathways with established roles in CRC pathogenesis. We  
439 identified loci at *SMAD3* and *SMAD9*, members of the TGF- $\beta$  signaling pathway that includes  
440 genes linked to familial CRC syndromes (e.g., *SMAD4* and *BMPRIA*) and several GWAS-  
441 implicated genes (e.g., *SMAD7*, *BMP2*, *BMP4*)<sup>61</sup>. We identified another locus near TGF- $\beta$   
442 Receptor 1 (*TGFBRI*). Nearby gene *GALNT12* reportedly harbors inactivating germline and  
443 somatic mutations in human colon cancers<sup>62</sup> and, therefore, could also be the regulated effector  
444 gene. We identified a locus at 14q23.1 near *DACT1*, a member of the Wnt- $\beta$ -catenin pathway  
445 with genes previously linked to familial CRC syndromes (*APC*<sup>63</sup>), and several GWAS-implicated  
446 genes (e.g., *CTNGB1*<sup>18</sup> and *TCF7L2*<sup>17</sup>). Genes related to telomere biology were linked by other  
447 GWAS: *TERC*<sup>10</sup> and *TERT*<sup>22</sup>, encoding the RNA and protein subunit of telomerase respectively,  
448 and *FEN1*<sup>17</sup>, involved in telomere stability<sup>64</sup>. A new locus at 20q13.33 harbors another gene  
449 related to telomere biology, *RTEL1*. This gene is involved in DNA double-strand break repair,  
450 and overlaps GWAS signals for cancers<sup>55,65</sup> and inflammation-related phenotypes, including  
451 inflammatory bowel disease<sup>66</sup> and atopic dermatitis<sup>67</sup>.

452  
453 Of 61 signals at 56 loci previously associated with CRC at GWS, 42 showed association  
454 evidence at  $P < 5 \times 10^{-8}$  in the combined meta-analysis, and 55 at  $P < 0.05$  in the independent  
455 Stage 2 meta-analysis (**Supplementary Table 3**). Of note, the association of rs755229494 at  
456 locus 5q22.2 ( $P=2.1 \times 10^{-12}$ ) was driven by studies with predominantly Ashkenazi Jewish ancestry  
457 and this SNP is in perfect LD with known missense SNP rs1801155 in the *APC* gene (I1307K),  
458 the minor allele of which is enriched in this population (MAF 6%), but rare in other  
459 populations<sup>68,69</sup>.

#### 460 461 **Delineating distinct association signals at CRC risk loci**

462 To identify additional independent association signals at known or new CRC risk loci, we  
463 conducted conditional analysis using individual-level data of 125,478 participants (Online  
464 Methods). At nine loci we observed 10 new independent association signals that attained  $P_j$

465  $<5 \times 10^{-8}$  in a joint multiple-variant analysis (**Table 2; Supplementary Table 4; Supplementary**  
466 **Fig. 5**). Because this analysis focused on  $<5\%$  of the genome, we also report signals at  $P_j < 1 \times 10^{-5}$   
467 <sup>5</sup> in **Supplementary Table 5**. At 22 loci, we observed 25 new suggestive associations with  $P_j$   
468  $< 1 \times 10^{-5}$ .

469  
470 At 11q13.4, near *POLD3* and *CHRD2*, we identified a new low-frequency variant (lead SNP  
471 rs61389091, MAF 3.94%) separated by a recombination hotspot from the known common  
472 variant signal<sup>12</sup> (LD  $r^2$  between lead SNPs  $< 0.01$ ). At 5p15.33, we identified another lower-  
473 frequency variant association (lead SNP rs78368589, MAF 5.97%), which was independent from  
474 the previously reported common variant signal 56kb away near *TERT* and *CLPTMIL* (LD  $r^2$  with  
475 lead SNP rs2735940  $< 0.01$ )<sup>22</sup>. Variants in this region were linked to many cancer types,  
476 including lung, prostate, breast, and ovarian cancer<sup>70</sup>.

477  
478 The remaining eight new signals involved common variants. At new locus 2q33.1, near genes  
479 *PLCLI* and *SATB2*, two statistically independent associations (LD  $r^2$  between two lead SNPs  
480  $< 0.01$ ) are separated by a recombination hotspot (**Supplementary Fig. 5**). In the MHC region,  
481 we identified a conditionally independent signal near genes involved in NF- $\kappa$ B signaling,  
482 including the gene encoding tumor necrosis factor- $\alpha$ , genes for the stress-signaling proteins  
483 MICA/MICB, and *HLA-B*. Locus 20p12.3, near *BMP2*, harbored four distinct association signals  
484 (**Figure 1**), two of which were reported previously<sup>10,11</sup> (**Supplementary Table 5**). All four SNPs  
485 selected in the model were in pairwise linkage equilibrium (maximum LD  $r^2 = 0.039$ , between  
486 rs189583 and rs994308). Our conditional analysis further confirmed that the signal  $\sim 1$ -Mb  
487 centromeric of *BMP2*, near gene *HAOI*, is independent. At 8q24.21 near *MYC*, the locus  
488 showing the second strongest statistical evidence of association in the combined meta-analysis  
489 (lead SNP rs6983267;  $P = 3.4 \times 10^{-64}$ ), we identified a second independent signal (lead SNP  
490 rs4313119,  $P_j = 2.1 \times 10^{-9}$ ; LD  $r^2$  with rs6983267  $< 0.001$ ). At the recently reported locus  
491 5p13.1<sup>22</sup>, near the non-coding RNA gene *LINC00603*, we identified an additional signal (lead  
492 SNP rs7708610) that was partly masked by the reported signal in the single-variant analysis due  
493 to the negative correlation between rs7708610 and rs12514517 ( $r = -0.18$ ;  $r^2 = 0.03$ ). This  
494 caused significance for both SNPs to increase markedly when fitted jointly (rs7708610,  
495 unconditional  $P = 1.5 \times 10^{-5}$  and  $P_j = 3.8 \times 10^{-9}$ ). At 12p13.32 near *CCND2*, we identified a new

496 signal (lead SNP rs3217874,  $P_J = 2.4 \times 10^{-9}$ ) and confirmed two previously associated signals<sup>13-15</sup>  
497 (**Supplementary Text**). At the *GREMI* locus on 15q13.3, two independent signals were  
498 previously described<sup>11</sup>. Our analyses suggest that this locus harbors three signals. A new signal  
499 represented by SNP rs17816465 is conditionally independent from the other two signals ( $P_J =$   
500  $1.4 \times 10^{-10}$ , conditioned on rs2293581 and rs12708491; LD with conditioning SNPs  $r^2 < 0.01$ ;  
501 **Supplementary Text**).

502  
503 Additionally, signals with  $P_J$  values approaching GWS were observed at new locus 3q13.2 near  
504 *BOC* (rs13086367, unconditional  $P = 6.7 \times 10^{-8}$ ,  $P_J = 6.9 \times 10^{-8}$ , MAF=47.4%), 96kb from the low-  
505 frequency signal represented by rs72942485 (unconditional  $P = 2.1 \times 10^{-8}$ ,  $P_J = 1.3 \times 10^{-8}$ ,  
506 MAF=2.2%); at known locus 10q22.3 near *ZMIZ1* (rs1250567, unconditional  $P = 3.1 \times 10^{-8}$ ,  $P_J =$   
507  $7.2 \times 10^{-8}$ , MAF=45.1%); and at new locus 13q22.1 near *KLF5* (rs45597035, unconditional  $P =$   
508  $2.7 \times 10^{-9}$ ,  $P_J = 8.1 \times 10^{-8}$ , MAF=34.4%) (**Supplementary Table 5**). Furthermore, we clarify  
509 previously reported independent association signals (**Supplementary Text**).

#### 511 **Associations of CRC risk variants with other traits**

512 Nineteen of the GWS association signals for CRC were in high LD ( $r^2 > 0.7$ ) with at least one  
513 SNP in the NHGRI-EBI GWAS Catalog<sup>46</sup> that has significant association in GWAS of other  
514 traits. Notable overlap included SNPs associated with other cancers, immune-related traits (e.g.,  
515 tonsillectomy, inflammatory bowel disease, and circulating white blood cell traits), obesity traits,  
516 blood pressure, and other cardiometabolic traits (**Supplementary Table 6**).

#### 518 **Mechanisms underlying CRC association signals**

519 To further localize variants driving the 40 newly identified signals, we used association evidence  
520 to define credible sets of variants that are 99% likely to contain the causal variant (Online  
521 Methods). The 99% credible set size for new loci ranged from one (17p12) to 93 (2q33.1). For  
522 11 distinct association signals, the set included ten or fewer variants (**Supplementary Table 7**).  
523 At locus 17p12, we narrowed the candidate variant to rs1078643, located in exon 1 of the  
524 lncRNA *LINC00675* that is primarily expressed in gastrointestinal tissues. Small credible sets  
525 were observed for locus 4q31.21 (two variants, indexed by synonymous SNP rs11727676 in  
526 *HHIP*), and signals at known loci near *GREMI* (one variant) and *CCND2* (two variants).

527  
528 We performed functional annotation of credible set variants to nominate putative causal variants.  
529 Eight sets contained coding variants but only the synonymous SNP in *HHIP* had a high posterior  
530 probability of driving the association (**Supplementary Table 8**). Next, we examined overlap of  
531 credible sets with regulatory genomic annotations from 51 existing CRC-relevant datasets to  
532 examine non-coding functions (Online Methods). Also, to better refine regulatory elements in  
533 active enhancers, we performed ATAC-seq to measure chromatin accessibility in four colonic  
534 crypts and used resulting data to annotate GWAS signals.

535  
536 Of the 40 sets, 36 overlapped with active enhancers identified by histone mark H3K27ac  
537 measured in normal colonic crypt epithelium, CRC cell lines, or CRC tissue (**Supplementary**  
538 **Table 8; Supplementary Fig. 6**). Twenty of these 36 overlapped with super-enhancers. Notably,  
539 when compared with epigenomics data from normal colonic crypt epithelium, all 36 sets  
540 overlapped enhancers with gained or lost activity in one or more CRC specimens. Eleven of  
541 these sets overlapped enhancers recurrently gained or lost in  $\geq 20$  CRC cell lines.

542  
543 The locus at GWAS hot spot 9p21 overlaps a super-enhancer, and the credible set is entirely  
544 intronic to *ANRIL*, alias *CDKN2B-AS1*. The Genotype-Tissue Expression (GTEx) data show that  
545 the antisense lncRNA *ANRIL* is exclusively expressed in transverse colon and small intestine.  
546 Interestingly, ANRIL recruits SUZ12 and EHZ2 to epigenetically silence tumor suppressor genes  
547 *CDKN2A/B*<sup>71</sup>.

548  
549 Noncoding somatic driver mutations or focal amplifications have been reported in regions  
550 regulating expression of *MYC*<sup>72</sup>, *TERT*<sup>73</sup>, and *KLF5*<sup>31</sup>, now implicated by GWAS for CRC. We  
551 checked whether GWAS-identified association signals co-localize with these regions and found  
552 that the *KLF5* signal overlaps the somatically amplified super-enhancer flanked by *KLF5* and  
553 *KLF12* (**Figure 2**). Also, the previously reported signal in the *TERT* promotor region<sup>22</sup> overlaps  
554 with the recurrent somatically mutated region in multiple cancers<sup>73</sup>.

555  
556 To test whether CRC associations are non-randomly distributed across genomic features, we  
557 used GARFIELD<sup>74</sup>. Focusing on DNase I hypersensitive site (DHS) peaks that identify open

558 chromatin, we observed significant enrichment across many cell types, particularly fetal tissues,  
559 with strongest enrichment observed in fetal gastrointestinal tissues, CD20<sup>+</sup> primary cells (B  
560 cells), and embryonic stem cells (**Supplementary Fig. 7; Supplementary Table 9**).

561  
562 We used MAGENTA<sup>75</sup> to identify pathways or gene sets enriched for associations with CRC,  
563 assessing two gene *P*-value cutoffs: 95th and 75th percentiles. At the 75th percentile, we  
564 observed enrichment of multiple KEGG cancer pathways at a false discovery rate (FDR) of 0.05.  
565 This was not observed for the 95th percentile cutoff and suggests that many more loci that are  
566 shared with other cancer types remain to be identified in larger studies. Using the 75th (95th)  
567 percentile cutoff, at FDR 0.05 and 0.20, we found enrichment of 7 (5) and 53 (24) gene sets,  
568 respectively. Established pathways related to TGF- $\beta$ /SMAD and BMP signaling were among the  
569 top enriched pathways. Other notable enriched pathways included Hedgehog signaling, basal cell  
570 carcinoma, melanogenesis, cell cycle, S phase, and telomere maintenance (**Supplementary**  
571 **Table 10**).

572

### 573 **Polygenicity of colorectal cancer and contribution of rare variants**

574 To estimate the contribution of rare variants (MAF  $\leq$ 1%) to CRC heritability, we used the LD-  
575 and MAF-stratified component GREML (GREML-LDMS) method implemented in GCTA<sup>76</sup>  
576 (Online Methods). Assuming a lifetime risk of 4.3%, we estimated that all imputed autosomal  
577 variants explain 21.6% (95% CI=17.5-25.7%) of the variation in liability for CRC, with almost  
578 half of this contributed by rare variants ( $h_g^2=9.7\%$ , 95% CI=6.2-13.3%; likelihood ratio test  
579  $P=0.003$ ); the estimated liability-scale heritability for variants with MAF  $>1\%$  is 11.8% (95%  
580 CI=8.9-14.7%). Our overall estimate falls within the range of heritability reported by large twin  
581 studies<sup>2</sup>. Because heritability estimates for rare variants are sensitive to potential biases due to  
582 technical effects or population stratification<sup>77</sup> and the contribution of rare variants is probably  
583 underestimated due to limitations of genotype imputation, results should be interpreted with  
584 caution. Overall, findings suggest that missing heritability is not large, but that many rare and  
585 common variants have yet to be identified.

586

### 587 **Familial relative risk explained by GWAS-identified variants**

588 Adjusting for winner's curse<sup>78</sup>, the familial relative risk (RR) to first-degree relatives ( $\lambda_0$ )  
589 attributable to GWAS-identified variants rose from 1.072 for the 55 previously described  
590 autosomal risk variants that showed evidence for replication at  $P < 0.05$ , to 1.092 after inclusion  
591 of 40 new signals, and increased further to 1.098 when we included 25 suggestive association  
592 signals reported in **Supplementary Table 5** (Online Methods). Assuming a  $\lambda_0$  of 2.2, the 55  
593 established signals account for 8.8% of familial RR explained (95% CI: 8.1-9.4). Established  
594 signals combined with 40 newly discovered signals account for 11.2% (95% CI: 10.5-12.0), and  
595 adding 25 suggestive signals increases this to 11.9% (95% CI: 11.1-12.7).

596

### 597 **Implications for stratified screening prevention**

598 We demonstrate how using a polygenic risk score (PRS) derived from 95 independent  
599 association signals could impact clinical guidelines for preventive screening. The difference in  
600 recommended starting age for screening for those in the highest 1% (and 10%) percentiles of risk  
601 compared with lowest percentiles is 18 years (and 10 years) for men, and 24 years (and 12 years)  
602 for women (**Figure 3**; Online Methods). **Supplementary Table 11** gives risk allele frequency  
603 (RAF) estimates in different populations for variants included in the PRS. As expected, RAFs  
604 vary across populations. Furthermore, differences in LD between tagging and true causal variants  
605 across populations can result in less prediction accuracy and subsequent lower predictive power  
606 of the PRS in non-European populations. Accordingly, it will be important to develop ancestry-  
607 specific PRSs that incorporate detailed fine-mapping results for each GWAS signal.

608

### 609 **DISCUSSION**

610 To further define the genetic architecture of sporadic CRC, we performed low-coverage WGS  
611 and imputation into a large set of GWAS data. We discovered 40 new CRC signals and  
612 replicated 55 previously reported signals. We found the first rare variant signal for sporadic  
613 CRC, which represents the strongest protective rare allelic effect identified to date. Our analyses  
614 highlight new genes and pathways contributing to underlying CRC risk and suggest roles for  
615 Krüppel-like factors, Hedgehog signaling, Hippo-YAP signaling, and immune function. Multiple  
616 loci provide new evidence for an important role of lncRNAs in CRC tumorigenesis<sup>79</sup>. Functional  
617 genomic annotations support that most sporadic CRC genetic risk lies in non-coding genomic  
618 regions. We further show how newly discovered variants can lead to improved risk prediction.

619  
620 This study underscores the critical importance of large-scale GWAS collaboration. While  
621 discovery of the rare variant signal was only possible through increased coverage and improved  
622 imputation accuracy enabled by imputation panels, sample size was pivotal for discovery of new  
623 CRC loci. Results suggest that CRC exhibits a highly polygenic architecture, much of which  
624 remains undefined. This also suggests that continued GWAS efforts, together with increasingly  
625 comprehensive imputation panels that allow for improved low-frequency and rare genetic variant  
626 imputation, will uncover more CRC risk variants. In addition, to investigate sites that are not  
627 imputable, large-scale deep sequencing will be needed. Importantly, the prevailing European bias  
628 in CRC GWAS limits the generalizability of findings and the application of PRSs in non-  
629 European (especially African) populations<sup>80</sup>. Therefore, a broader representation of ancestries in  
630 CRC GWAS is necessary.

631  
632 Studies of somatic genomic alterations in cancer have mostly focused on the coding genome and  
633 identification of noncoding drivers has proven to be challenging<sup>73</sup>. Yet, noncoding somatic driver  
634 mutations or focal amplifications in regulatory regions impacting expression have been reported  
635 for *MYC*<sup>72</sup>, *TERT*<sup>73</sup>, and *KLF5*<sup>31</sup>. The observed overlap between GWAS-identified CRC risk loci  
636 and somatic driver regions strongly suggests that expanding the search of somatic driver  
637 mutations to noncoding regulatory elements will yield additional discoveries and that searches  
638 for somatic drivers can be guided by GWAS findings.

639  
640 Additionally, we found loci near proposed drug targets, including *CHDI*, implicated by the rare  
641 variant signal, and *KLF5*. To date, cancer drug target discovery research has almost exclusively  
642 focused on properties of cancer cells, yielding drugs that target proteins either highly expressed  
643 or expressed in a mutant form due to frequent recurrent somatic missense mutations (e.g.,  
644 *BRAF*<sup>V600E</sup>) or gene fusion events. In stark contrast with other common complex diseases, cancer  
645 GWAS results are not being used extensively to inform drug target selection. It has been  
646 estimated that selecting targets supported by GWAS could double the success rate in clinical  
647 development<sup>81</sup>. Our discoveries corroborate that not using GWAS results to inform drug  
648 discovery is a missed opportunity, not only for treating cancers, but also for chemoprevention in  
649 high-risk individuals.

650

651 In summary, in the largest genome-wide scan for sporadic CRC risk thus far, we identified the  
652 first rare variant signal for sporadic CRC, and almost doubled the number of known association  
653 signals. Our findings provide a substantial number of new leads that may spur downstream  
654 investigation into the biology of CRC risk, and that will impact drug development and clinical  
655 guidelines, such as personalized screening decisions.

656

### 657 **Acknowledgements**

658 A full list of acknowledgements appears in the Supplementary Notes.

659

### 660 **Author contributions**

661 J.R.H., S.A.B. and T.A.H. contributed equally, and D.A.N., S.B.G., L.H. and U.P. jointly  
662 supervised this research. J.R.H., S.A.B., T.A.H., H.M.K., D.V.C., M.W., F.R.S., J.D.S., D.A.,  
663 M.H.A., K.A., C.A.-C., V.A., C.B., J.A.B., S.I.B., S.B., D.T.B., J.B., H. Boeing, H. Brenner, S.  
664 Brezina, S. Buch, D.D.B., A.B.-H., K.B., B.J.C., P.T.C., S.C.-B., A.T.C., J.C.-C., S.J.C., M.-  
665 D.C., S.H.C., A.J.C., K.C., A.d.I.C., D.F.E., S.G.E., F.E., D.R.E., E.J.M.F., J.C.F., D.F., S.G.,  
666 G.G.G., E.G., P.J.G., J.S.G., A.G., M.J.G., R.W.H., J.H., H.H., R.B.H., P.H., M.H., J.L.H., W.-  
667 Y.H., T.J.H., D.J.H., R.J., E.J.J., M.A.J., T.O.K., T.J.K., H.R.K., L.N.K., C.K., S.K., S.-S.K.,  
668 L.L.M., S.C.L., C.I.L., L.L., A.L., N.M.L., S.M., S.D.M., V.M., G.M., M.M., R.L.M., L.M.,  
669 R.M., A.N., P.A.N., K.O., N.C.O.-M., B.P., P.S.P., R.P., V.P., P.D.P.P., E.A.P., R.L.P., G.R.,  
670 H.S.R., E.R., M.R.-B., C.S., R.E.S., D.S., M.-H.S., S.S., M.L.S., C.M.T., S.N.T., A.T., C.M.U.,  
671 F.J.B.v.D., B.V.G., H.v.K., J.V., K.V., P.V., L.V., V.V., E.W., C.R.W., A.W., M.O.W., A.H.W.,  
672 B.W.Z., W.Z., P.C.S., J.D.P., M.C.B., G.C., V.M., G.R.A., D.A.N., S.B.G., L.H. and U.P.  
673 conceived and designed the experiments. T.A.H., M.W., J.D.S., K.F.D., D.D., R.I., E.K., H.L.,  
674 C.E.M., E.P., J.R., T.S., S.S.T., D.J.V.D.B., M.C.B., D.A.N. performed the experiments. J.R.H.,  
675 H.M.K., S.C., S.L.S., D.V.C., C.Q., J.J., C.K.E., P.G., F.R.S., D.M.L., S.C.N., N.A.S.-A.,  
676 C.A.L., M.L., T.L.L., Y.-R.S., A.K., G.R.A., L.H. performed statistical analysis. J.R.H., S.A.B.,  
677 T.A.H., H.M.K., S.C., S.L.S., D.V.C., C.Q., J.J., C.K.E., P.G., M.W., F.R.S., D.M.L., S.C.N.,  
678 N.A.S.-A., B.L.B., C.S.C., C.M.C., K.R.C., J.G., W.-L.H., C.A.L., S.M.L., M.L., Y.L., T.L.L.,  
679 M.S., Y.-R.S., A.K., G.R.A., L.H., U.P. analyzed the data. H.M.K., C.K.E., D.A., M.H.A., K.A.,  
680 C.A.-C., V.A., C.B., J.A.B., S.I.B., S.B., D.T.B., J.B., H. Boeing, H. Brenner, S. Brezina, S.

681 Buch, D.D.B., A.B.-H., K.B., B.J.C., P.T.C., S.C.-B., A.T.C., J.C.-C., S.J.C., M.-D.C., S.H.C.,  
682 A.J.C., K.C., A.d.I.C., D.F.E., S.G.E., F.E., D.R.E., E.J.M.F., J.C.F., R.F., L.M.F., D.F., M.G.,  
683 S.G., W.J.G., G.G.G., P.J.G., W.M.G., J.S.G., A.G., M.J.G., R.W.H., J.H., H.H., S.H., R.B.H.,  
684 P.H., M.H., J.L.H., W.-Y.H., T.J.H., D.J.H., G.I.-S., G.E.I., R.J., E.J.J., M.A.J., A.D.J., C.E.J.,  
685 T.O.K., T.J.K., H.R.K., L.N.K., C.K., T.K., S.K., S.-S.K., S.C.L., L.L.M., S.C.L., F.L., C.I.L.,  
686 L.L., W.L., A.L., N.M.L., S.M., S.D.M., V.M., G.M., M.M., R.L.M., L.M., N.M., R.M., A.N.,  
687 P.A.N., K.O., S.O, N.C.O.-M., B.P., P.S.P., R.P., V.P., P.D.P.P., M.P., E.A.P., R.L.P., L.R.,  
688 G.R., H.S.R., E.R., M.R.-B., L.C.S., C.S., R.E.S., M.S., M.-H.S., K.S., S.S., M.L.S., M.C.S.,  
689 Z.K.S., C.S., C.M.T., S.N.T., D.C.T., A.E.T., A.T., C.M.U., F.J.B.v.D., B.V.G., H.v.K., J.V.,  
690 K.V., P.V., L.V., V.V., K.W., S.J.W., E.W., A.K.W., C.R.W., A.W., M.O.W., A.H.W., S.H.Z.,  
691 B.W.Z., Q.Z., W.Z., P.C.S., J.D.P., M.C.B., A.K., G.C., V.M., G.R.A., S.B.G. and U.P.  
692 contributed reagents/materials/analysis tools. J.R.H., S.A.B., T.A.H., J.J., L.H. and U.P. wrote  
693 the paper.

694

#### 695 **Competing Interests Statement**

696 Goncalo R Abecasis has received compensation from 23andMe and Helix. He is currently an  
697 employee of Regeneron Pharmaceuticals. Heather Hampel performs collaborative research with  
698 Ambry Genetics, InVitae Genetics, and Myriad Genetic Laboratories, Inc., is on the scientific  
699 advisory board for InVitae Genetics and Genome Medical, and has stock in Genome Medical.  
700 Rachel Pearlman has participated in collaborative funded research with Myriad Genetics  
701 Laboratories and Invitae Genetics but has no financial competitive interest.

702

#### 703 **REFERENCES**

704

- 705 1. Ferlay, J. *et al.* Cancer incidence and mortality worldwide: sources, methods and major  
706 patterns in GLOBOCAN 2012. *Int J Cancer* **136**, E359-86 (2015).
- 707 2. Lichtenstein, P. *et al.* Environmental and heritable factors in the causation of cancer--  
708 analyses of cohorts of twins from Sweden, Denmark, and Finland. *N Engl J Med* **343**, 78-  
709 85 (2000).
- 710 3. Czene, K., Lichtenstein, P. & Hemminki, K. Environmental and heritable causes of cancer  
711 among 9.6 million individuals in the Swedish Family-Cancer Database. *Int J Cancer* **99**,  
712 260-266 (2002).
- 713 4. Sud, A., Kinnersley, B. & Houlston, R. S. Genome-wide association studies of cancer:  
714 current insights and future perspectives. *Nat Rev Cancer* **17**, 692-704 (2017).
- 715 5. Tomlinson, I. *et al.* A genome-wide association scan of tag SNPs identifies a susceptibility

- 716 variant for colorectal cancer at 8q24.21. *Nat Genet* **39**, 984–988 (2007).
- 717 6. Broderick, P. *et al.* A genome-wide association study shows that common alleles of  
718 SMAD7 influence colorectal cancer risk. *Nat Genet* **39**, 1315–1317 (2007).
- 719 7. Tomlinson, I. P. M. *et al.* A genome-wide association study identifies colorectal cancer  
720 susceptibility loci on chromosomes 10p14 and 8q23.3. *Nat Genet* **40**, 623–630 (2008).
- 721 8. Tenesa, A. *et al.* Genome-wide association scan identifies a colorectal cancer susceptibility  
722 locus on 11q23 and replicates risk loci at 8q24 and 18q21. *Nat Genet* **40**, 631–637 (2008).
- 723 9. COGENT Study *et al.* Meta-analysis of genome-wide association data identifies four new  
724 susceptibility loci for colorectal cancer. *Nat Genet* **40**, 1426–1435 (2008).
- 725 10. Houlston, R. S. *et al.* Meta-analysis of three genome-wide association studies identifies  
726 susceptibility loci for colorectal cancer at 1q41, 3q26.2, 12q13.13 and 20q13.33. *Nat Genet*  
727 **42**, 973–977 (2010).
- 728 11. Tomlinson, I. P. M. *et al.* Multiple common susceptibility variants near BMP pathway loci  
729 GREM1, BMP4, and BMP2 explain part of the missing heritability of colorectal cancer.  
730 *PLoS Genet* **7**, e1002105 (2011).
- 731 12. Dunlop, M. G. *et al.* Common variation near CDKN1A, POLD3 and SHROOM2  
732 influences colorectal cancer risk. *Nat Genet* **44**, 770–776 (2012).
- 733 13. Peters, U. *et al.* Identification of Genetic Susceptibility Loci for Colorectal Tumors in a  
734 Genome-Wide Meta-analysis. *Gastroenterology* **144**, 799–807.e24 (2013).
- 735 14. Jia, W.-H. *et al.* Genome-wide association analyses in East Asians identify new  
736 susceptibility loci for colorectal cancer. *Nat Genet* **45**, 191–196 (2013).
- 737 15. Whiffin, N. *et al.* Identification of susceptibility loci for colorectal cancer in a genome-  
738 wide meta-analysis. *Hum Mol Genet* **23**, 4729–4737 (2014).
- 739 16. Wang, H. *et al.* Trans-ethnic genome-wide association study of colorectal cancer identifies  
740 a new susceptibility locus in VTI1A. *Nat Commun* **5**, 4613 (2014).
- 741 17. Zhang, B. *et al.* Large-scale genetic study in East Asians identifies six new loci associated  
742 with colorectal cancer risk. *Nat Genet* **46**, 533–542 (2014).
- 743 18. Schumacher, F. R. *et al.* Genome-wide association study of colorectal cancer identifies six  
744 new susceptibility loci. *Nat Commun* **6**, 7138 (2015).
- 745 19. Al-Tassan, N. A. *et al.* A new GWAS and meta-analysis with 1000Genomes imputation  
746 identifies novel risk variants for colorectal cancer. *Sci Rep* **5**, 10442 (2015).
- 747 20. Orlando, G. *et al.* Variation at 2q35 (PNKD and TMBIM1) influences colorectal cancer  
748 risk and identifies a pleiotropic effect with inflammatory bowel disease. *Hum Mol Genet*  
749 **25**, 2349–2359 (2016).
- 750 21. Zeng, C. *et al.* Identification of susceptibility loci and genes for colorectal cancer risk.  
751 *Gastroenterology* **150**, 1633–1645 (2016).
- 752 22. Schmit, S. L. *et al.* Novel common genetic susceptibility loci for colorectal cancer. *J Natl*  
753 *Cancer Inst* 1–12 (2018). doi:10.1093/jnci/djy099
- 754 23. Fuchsberger, C. *et al.* The genetic architecture of type 2 diabetes. *Nature* **536**, 41–47  
755 (2016).
- 756 24. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation.  
757 *Nature* **526**, 68–74 (2015).
- 758 25. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat*  
759 *Genet* **48**, 1279–1283 (2016).
- 760 26. Amos, C. I. *et al.* The oncoarray consortium: A network for understanding the genetic  
761 architecture of common cancers. *Cancer Epidemiol Biomarkers Prev* **26**, 126–135 (2017).

- 762 27. Zhao, D. & DePinho, R. A. Synthetic essentiality: Targeting tumor suppressor deficiencies  
763 in cancer. *Bioessays* **39**, (2017).
- 764 28. Zhao, D. *et al.* Synthetic essentiality of chromatin remodelling factor CHD1 in PTEN-  
765 deficient cancer. *Nature* **542**, 484–488 (2017).
- 766 29. Xiao, Y. *et al.* RGMb is a novel binding partner for PD-L2 and its engagement with PD-L2  
767 promotes respiratory tolerance. *J Exp Med* **211**, 943–959 (2014).
- 768 30. Topalian, S. L. *et al.* Safety, activity, and immune correlates of anti-PD-1 antibody in  
769 cancer. *N Engl J Med* **366**, 2443–2454 (2012).
- 770 31. Zhang, X. *et al.* Somatic superenhancer duplications and hotspot mutations lead to  
771 oncogenic activation of the KLF5 transcription factor. *Cancer Discov* **8**, 108–125 (2018).
- 772 32. Giannakis, M. *et al.* Genomic Correlates of Immune-Cell Infiltrates in Colorectal  
773 Carcinoma. *Cell Rep* **15**, 857–865 (2016).
- 774 33. Dekker, R. J. *et al.* KLF2 provokes a gene expression pattern that establishes functional  
775 quiescent differentiation of the endothelium. *Blood* **107**, 4354–4363 (2006).
- 776 34. Boon, R. A. *et al.* KLF2 suppresses TGF-beta signaling in endothelium through induction  
777 of Smad7 and inhibition of AP-1. *Arterioscler Thromb Vasc Biol* **27**, 532–539 (2007).
- 778 35. Chakroborty, D. *et al.* Dopamine stabilizes tumor blood vessels by up-regulating  
779 angiopoietin 1 expression in pericytes and Kruppel-like factor-2 expression in tumor  
780 endothelial cells. *Proc Natl Acad Sci U S A* **108**, 20730–20735 (2011).
- 781 36. Lee, S.-J. *et al.* Regulation of hypoxia-inducible factor 1 $\alpha$  (HIF-1 $\alpha$ ) by lysophosphatidic  
782 acid is dependent on interplay between p53 and Krüppel-like factor 5. *J Biol Chem* **288**,  
783 25244–25253 (2013).
- 784 37. Zhang, H. *et al.* Lysophosphatidic acid facilitates proliferation of colon cancer cells via  
785 induction of Krüppel-like factor 5. *J Biol Chem* **282**, 15541–15549 (2007).
- 786 38. Ma, Z. *et al.* Long non-coding RNA SNHG15 inhibits P15 and KLF2 expression to  
787 promote pancreatic cancer proliferation through EZH2-mediated H3K27me3. *Oncotarget*  
788 **8**, 84153–84167 (2017).
- 789 39. Evangelista, M., Tian, H. & de Sauvage, F. J. The hedgehog signaling pathway in cancer.  
790 *Clin Cancer Res* **12**, 5924–5928 (2006).
- 791 40. Gerling, M. *et al.* Stromal Hedgehog signalling is downregulated in colon cancer and its  
792 restoration restrains tumour growth. *Nat Commun* **7**, 12321 (2016).
- 793 41. Mille, F. *et al.* The Shh receptor Boc promotes progression of early medulloblastoma to  
794 advanced tumors. *Dev Cell* **31**, 34–47 (2014).
- 795 42. Mathew, E. *et al.* Dosage-dependent regulation of pancreatic cancer growth and  
796 angiogenesis by hedgehog signaling. *Cell Rep* **9**, 484–494 (2014).
- 797 43. Zhao, B., Li, L., Lei, Q. & Guan, K.-L. The Hippo-YAP pathway in organ size control and  
798 tumorigenesis: an updated version. *Genes Dev* **24**, 862–874 (2010).
- 799 44. Camargo, F. D. *et al.* YAP1 increases organ size and expands undifferentiated progenitor  
800 cells. *Curr Biol* **17**, 2054–2060 (2007).
- 801 45. Ma, X., Zhang, H., Xue, X. & Shah, Y. M. Hypoxia-inducible factor 2 $\alpha$  (HIF-2 $\alpha$ ) promotes  
802 colon cancer growth by potentiating Yes-associated protein 1 (YAP1) activity. *J Biol Chem*  
803 **292**, 17046–17056 (2017).
- 804 46. MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association  
805 studies (GWAS Catalog). *Nucleic Acids Res* **45**, D896–D901 (2017).
- 806 47. Seshagiri, S. *et al.* Recurrent R-spondin fusions in colon cancer. *Nature* **488**, 660–664  
807 (2012).

- 808 48. Song, F. *et al.* Identification of a melanoma susceptibility locus and somatic mutation in  
809 TET2. *Carcinogenesis* **35**, 2097–2101 (2014).
- 810 49. Eeles, R. A. *et al.* Identification of seven new prostate cancer susceptibility loci through a  
811 genome-wide association study. *Nat Genet* **41**, 1116–1121 (2009).
- 812 50. Michailidou, K. *et al.* Association analysis identifies 65 new breast cancer risk loci. *Nature*  
813 **551**, 92–94 (2017).
- 814 51. Schunkert, H. *et al.* Large-scale association analysis identifies 13 new susceptibility loci  
815 for coronary artery disease. *Nat Genet* **43**, 333–338 (2011).
- 816 52. Scott, L. J. *et al.* A genome-wide association study of type 2 diabetes in Finns detects  
817 multiple susceptibility variants. *Science* **316**, 1341–1345 (2007).
- 818 53. Al Olama, A. A. *et al.* A meta-analysis of 87,040 individuals identifies 23 new  
819 susceptibility loci for prostate cancer. *Nat Genet* **46**, 1103–1109 (2014).
- 820 54. Timofeeva, M. N. *et al.* Influence of common genetic variation on lung cancer risk: meta-  
821 analysis of 14 900 cases and 29 485 controls. *Hum Mol Genet* **21**, 4980–4995 (2012).
- 822 55. Shete, S. *et al.* Genome-wide association study identifies five susceptibility loci for glioma.  
823 *Nat Genet* **41**, 899–904 (2009).
- 824 56. Bishop, D. T. *et al.* Genome-wide association study identifies three loci associated with  
825 melanoma risk. *Nat Genet* **41**, 920–925 (2009).
- 826 57. Sapkota, Y. *et al.* Meta-analysis identifies five novel loci associated with endometriosis  
827 highlighting key genes involved in hormone metabolism. *Nat Commun* **8**, 15539 (2017).
- 828 58. Cannon-Albright, L. A. *et al.* Assignment of a locus for familial melanoma, MLM, to  
829 chromosome 9p13-p22. *Science* **258**, 1148–1152 (1992).
- 830 59. Hussussian, C. J. *et al.* Germline p16 mutations in familial melanoma. *Nat Genet* **8**, 15–21  
831 (1994).
- 832 60. Seoane, J. *et al.* TGFbeta influences Myc, Miz-1 and Smad to control the CDK inhibitor  
833 p15INK4b. *Nat Cell Biol* **3**, 400–408 (2001).
- 834 61. Jung, B., Staudacher, J. J. & Beauchamp, D. Transforming Growth Factor  $\beta$  Superfamily  
835 Signaling in Development of Colorectal Cancer. *Gastroenterology* **152**, 36–52 (2017).
- 836 62. Guda, K. *et al.* Inactivating germ-line and somatic mutations in polypeptide N-  
837 acetylgalactosaminyltransferase 12 in human colon cancers. *Proc Natl Acad Sci U S A* **106**,  
838 12921–12925 (2009).
- 839 63. Groden, J. *et al.* Identification and characterization of the familial adenomatous polyposis  
840 coli gene. *Cell* **66**, 589–600 (1991).
- 841 64. Saharia, A. *et al.* FEN1 ensures telomere stability by facilitating replication fork re-  
842 initiation. *J Biol Chem* **285**, 27057–27066 (2010).
- 843 65. Eeles, R. A. *et al.* Identification of 23 new prostate cancer susceptibility loci using the  
844 iCOGS custom genotyping array. *Nat Genet* **45**, 385–91, 391e1 (2013).
- 845 66. Liu, J. Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel  
846 disease and highlight shared genetic risk across populations. *Nat Genet* **47**, 979–986  
847 (2015).
- 848 67. Paternoster, L. *et al.* Multi-ancestry genome-wide association study of 21,000 cases and  
849 95,000 controls identifies new risk loci for atopic dermatitis. *Nat Genet* **47**, 1449–1456  
850 (2015).
- 851 68. Laken, S. J. *et al.* Familial colorectal cancer in Ashkenazim due to a hypermutable tract in  
852 APC. *Nat Genet* **17**, 79–83 (1997).
- 853 69. Niell, B. L., Long, J. C., Rennert, G. & Gruber, S. B. Genetic anthropology of the

- 854 colorectal cancer-susceptibility allele APC I1307K: evidence of genetic drift within the  
855 Ashkenazim. *Am J Hum Genet* **73**, 1250–1260 (2003).
- 856 70. Karami, S. *et al.* Telomere structure and maintenance gene variants and risk of five cancer  
857 types. *Int J Cancer* **139**, 2655–2670 (2016).
- 858 71. Congrains, A., Kamide, K., Ohishi, M. & Rakugi, H. ANRIL: molecular mechanisms and  
859 implications in human health. *Int J Mol Sci* **14**, 1278–1292 (2013).
- 860 72. Zhang, X. *et al.* Identification of focally amplified lineage-specific super-enhancers in  
861 human epithelial cancers. *Nat Genet* **48**, 176–182 (2016).
- 862 73. Rheinbay, E. *et al.* Discovery and characterization of coding and non-coding driver  
863 mutations in more than 2,500 whole cancer genomes. *BioRxiv* (2017). doi:10.1101/237313
- 864 74. Iotchkova, V. *et al.* GARFIELD - GWAS Analysis of Regulatory or Functional  
865 Information Enrichment with LD correction. *BioRxiv* (2016). doi:10.1101/085738
- 866 75. Segrè, A. V. *et al.* Common inherited variation in mitochondrial genes is not enriched for  
867 associations with type 2 diabetes or related glyceic traits. *PLoS Genet* **6**, (2010).
- 868 76. Yang, J. *et al.* Genetic variance estimation with imputed variants finds negligible missing  
869 heritability for human height and body mass index. *Nat Genet* **47**, 1114–1120 (2015).
- 870 77. Bhatia, G. *et al.* Subtle stratification confounds estimates of heritability from rare variants.  
871 *BioRxiv* (2016). doi:10.1101/048181
- 872 78. Zhong, H. & Prentice, R. L. Bias-reduced estimators and confidence intervals for odds  
873 ratios in genome-wide association studies. *Biostatistics* **9**, 621–634 (2008).
- 874 79. Cheetham, S. W., Gruhl, F., Mattick, J. S. & Dinger, M. E. Long noncoding RNAs and the  
875 genetics of cancer. *Br J Cancer* **108**, 2419–2425 (2013).
- 876 80. Popejoy, A. B. & Fullerton, S. M. Genomics is failing on diversity. *Nature* **538**, 161–164  
877 (2016).
- 878 81. Nelson, M. R. *et al.* The support of human genetic evidence for approved drug indications.  
879 *Nat Genet* **47**, 856–860 (2015).

880

881

882 **FIGURE LEGENDS**

883

884 **Figure 1 Conditionally independent association signals at the *BMP2* locus.** Regional  
885 association plot showing the unconditional  $-\log_{10}(P\text{-value})$  for the association with CRC risk in  
886 the combined meta-analysis of up to 125,478 individuals, as a function of genomic position  
887 (Build 37) for each variant in the region. The lead variants are indicated by a diamond symbol  
888 and its positions are indicated by dashed vertical lines. The color-labeling and shape of all other  
889 variants indicate the lead variant with which they are in strongest LD. The two new genome-  
890 wide significant signals are indicated by an asterisk.

891

892 **Figure 2 Functional genomic annotation of new CRC risk locus overlapping *KLF5* super-**  
893 **enhancer. Top:** Regional association plot showing the unconditional  $-\log_{10}(P\text{-value})$  for the  
894 association with CRC risk in the combined meta-analysis of up to 125,478 individuals, as a  
895 function of genomic position (Build 37) for each variant in the region. The lead variants are  
896 indicated by a diamond symbol and its positions are indicated by dashed vertical lines. The  
897 color-labeling and shape of all other variants indicate the lead variant with which they are in  
898 strongest LD. **Bottom:** UCSC genome browser annotations for region overlapping the super-  
899 enhancer flanked by *KLF5* and *KLF12*, and spanning variants in LD with rs78341008, and with  
900 two conditionally independent association signals indexed by rs45597035 and rs1924816. The  
901 region is annotated with the following tracks (from top to bottom): UCSC gene annotations;  
902 epigenomic profiles showing MACS2 peak calls as transparent overlays for different samples  
903 taken from non-diseased colonic crypt cells or colon tissue (purple) and from different primary  
904 CRC cell lines or tumor samples (teal); position of the lead variants and variants in LD with the  
905 lead; variants in the 99% credible set; the union of super-enhancers called using the ROSE  
906 package; gray bars highlight the targeted enhancers (e1, e3, and e4) previously shown by Zhang  
907 *et al.*<sup>31</sup> to have combinatorial effects on *KLF5* expression. ATAC-seq data newly generated for  
908 this study show high resolution annotation of putative binding regions within the active super-  
909 enhancer further fine-mapping putative causal variants at each of the three signals.

910

911 **Figure 3 Recommended age to start CRC screening based on a polygenic risk score (PRS).**  
912 The PRS was constructed using the 95 known and newly discovered variants. The horizontal  
913 lines represent the recommended age for the first endoscopy for an average-risk person in the

914 current screening guideline for CRC. The risk threshold to determine the age for the first  
915 screening was set as the average of 10-year CRC risks for a 50-year-old man (1.25%) and  
916 woman (0.68%), i.e.  $(1.25\%+0.68\%)/2 = 0.97\%$ , who have not previously received an  
917 endoscopy. Details are given in the Online Methods.  
918

9 **Table 1** New CRC risk loci reaching genome-wide significance ( $P < 5 \times 10^{-8}$ ) in the combined (Stage 1 and Stage 2) meta-analysis.

Locus	Nearby gene(s)	rsID lead variant	Chr.	Position (Build 37)	Alleles (risk/other)	RAF (%)	Stage 1 meta-analysis: up to 34,869 cases and 29,051 controls			Stage 2 meta-analysis: up to 23,262 cases and 38,296 controls			Combined meta-analysis: up to 58,131 cases and 67,347 controls		
							OR	95% CI	<i>P</i>	OR	95% CI	<i>P</i>	OR	95% CI	<i>P</i>
<b>Rare variants</b>															
5q21.1	<i>RGMB; CHD1</i>	rs145364999*	5	98,206,082	T/A	99.69	1.57	1.20-2.05	$9.0 \times 10^{-4}$	1.93	1.48-2.52	$1.0 \times 10^{-6}$	1.74	1.45-2.10	$6.3 \times 10^{-9}$
<b>Low-frequency variants</b>															
3q13.2	<i>BOC</i>	rs72942485	3	112,999,560	G/A	98.02	1.16	1.07-1.26	$2.5 \times 10^{-4}$	1.23	1.12-1.35	$1.5 \times 10^{-5}$	1.19	1.12-1.26	$2.1 \times 10^{-8}$
<b>Common variants</b>															
1p34.3	<i>FHL3</i>	rs4360494 <sup>§</sup>	1	38,455,891	G/C	45.39	1.05	1.03-1.08	$2.9 \times 10^{-5}$	1.06	1.03-1.08	$3.3 \times 10^{-5}$	1.05	1.04-1.07	$3.8 \times 10^{-9}$
1p32.3	<i>TTC22; PCSK9</i>	rs12144319*	1	55,246,035	C/T	25.48	1.07	1.04-1.10	$1.4 \times 10^{-6}$	1.07	1.04-1.10	$5.5 \times 10^{-6}$	1.07	1.05-1.09	$3.3 \times 10^{-11}$
2q24.2	<i>MARCH7; TANC1</i>	rs448513 <sup>§</sup>	2	159,964,552	C/T	32.60	1.06	1.03-1.08	$1.9 \times 10^{-5}$	1.05	1.02-1.08	$5.8 \times 10^{-4}$	1.05	1.03-1.07	$4.4 \times 10^{-8}$
2q33.1	<i>SATB2</i>	rs983402*	2	199,781,586	T/C	33.12	1.05	1.03-1.08	$7.2 \times 10^{-5}$	1.08	1.05-1.11	$1.0 \times 10^{-8}$	1.07	1.05-1.09	$7.7 \times 10^{-12}$
3q22.2	<i>SLCO2A1</i>	rs10049390 <sup>§</sup>	3	133,701,119	A/G	73.53	1.06	1.03-1.09	$4.9 \times 10^{-5}$	1.07	1.04-1.10	$1.8 \times 10^{-5}$	1.06	1.04-1.08	$3.8 \times 10^{-9}$
4q24	<i>TET2</i>	rs1391441	4	106,128,760	A/G	67.20	1.05	1.02-1.07	$1.5 \times 10^{-4}$	1.06	1.03-1.09	$2.3 \times 10^{-5}$	1.05	1.03-1.07	$1.6 \times 10^{-8}$
4q31.21	<i>HHIP</i>	rs11727676	4	145,659,064	C/T	9.80	1.08	1.03-1.13	$4.5 \times 10^{-4}$	1.10	1.05-1.14	$1.5 \times 10^{-5}$	1.09	1.06-1.12	$2.9 \times 10^{-8}$
6p21.32	<i>HLA-DRB1; HLA-DQA1</i>	rs9271695*	6	32,593,080	G/A	79.54	1.09	1.06-1.13	$1.3 \times 10^{-7}$	1.09	1.05-1.12	$1.7 \times 10^{-7}$	1.09	1.07-1.12	$1.1 \times 10^{-13}$
7p13	<i>MYO1G; SNHG15; CCM2; TBRG4</i>	rs12672022 <sup>§</sup>	7	45,136,423	T/C	83.45	1.07	1.04-1.11	$1.6 \times 10^{-5}$	1.06	1.03-1.10	$4.4 \times 10^{-4}$	1.07	1.04-1.09	$2.8 \times 10^{-8}$
9p21.3	<i>ANRIL; CDKN2A; CDKN2B</i>	rs1537372 <sup>§</sup>	9	22,103,183	G/T	56.92	1.05	1.02-1.07	$1.4 \times 10^{-4}$	1.06	1.03-1.08	$2.4 \times 10^{-5}$	1.05	1.03-1.07	$1.4 \times 10^{-8}$
9q22.33	<i>GALNT12; TGFBR1</i>	rs34405347 <sup>§</sup>	9	101,679,752	T/G	90.34	1.08	1.04-1.13	$5.5 \times 10^{-5}$	1.09	1.04-1.13	$1.5 \times 10^{-4}$	1.09	1.05-1.12	$3.1 \times 10^{-8}$
9q31.3	<i>LPAR1</i>	rs10980628	9	113,671,403	C/T	21.06	1.05	1.02-1.09	$3.1 \times 10^{-4}$	1.08	1.05-1.11	$1.3 \times 10^{-6}$	1.07	1.04-1.09	$2.8 \times 10^{-9}$
11q22.1	<i>YAP1</i>	rs2186607	11	101,656,397	T/A	51.78	1.05	1.03-1.08	$1.1 \times 10^{-5}$	1.05	1.03-1.08	$3.3 \times 10^{-5}$	1.05	1.04-1.07	$1.5 \times 10^{-9}$
12q12	<i>PRICKLE1; YAF2</i>	rs11610543 <sup>§</sup>	12	43,134,191	G/A	50.13	1.05	1.03-1.08	$1.1 \times 10^{-5}$	1.06	1.03-1.08	$2.8 \times 10^{-5}$	1.05	1.04-1.07	$1.3 \times 10^{-9}$
12q13.3	<i>STAT6; LRPI; NAB2</i>	rs4759277	12	57,533,690	A/C	35.46	1.07	1.04-1.09	$8.4 \times 10^{-7}$	1.04	1.02-1.07	$1.6 \times 10^{-3}$	1.05	1.04-1.07	$9.4 \times 10^{-9}$
13q13.3	<i>SMAD9</i>	rs7333607*	13	37,462,010	G/A	23.50	1.09	1.06-1.12	$2.5 \times 10^{-8}$	1.07	1.04-1.10	$4.4 \times 10^{-6}$	1.08	1.06-1.10	$6.3 \times 10^{-13}$
13q22.1	<i>KLF5</i>	rs78341008 <sup>§</sup>	13	73,791,554	C/T	7.19	1.13	1.07-1.18	$1.4 \times 10^{-6}$	1.11	1.05-1.16	$4.8 \times 10^{-5}$	1.12	1.08-1.16	$3.2 \times 10^{-10}$
13q34	<i>COL4A2; COL4A1; RAB20</i>	rs8000189	13	111,075,881	T/C	64.01	1.05	1.02-1.07	$2.1 \times 10^{-4}$	1.07	1.04-1.10	$1.3 \times 10^{-6}$	1.06	1.04-1.08	$1.8 \times 10^{-9}$
14q23.1	<i>DACT1</i>	rs17094983 <sup>§</sup>	14	59,189,361	G/A	87.73	1.10	1.07-1.15	$8.4 \times 10^{-8}$	1.08	1.04-1.12	$9.0 \times 10^{-5}$	1.09	1.06-1.12	$4.6 \times 10^{-11}$

15q22.33	<i>SMAD3</i>	rs56324967*	15	67,402,824	C/T	67.57	1.07	1.04-1.10	$2.2 \times 10^{-7}$	1.08	1.05-1.11	$9.8 \times 10^{-8}$	1.07	1.05-1.09	$1.1 \times 10^{-13}$
16q23.2	<i>MAF</i>	rs9930005§	16	80,043,258	C/A	43.03	1.05	1.03-1.08	$1.3 \times 10^{-5}$	1.05	1.02-1.07	$4.0 \times 10^{-4}$	1.05	1.03-1.07	$2.1 \times 10^{-8}$
17p12	<i>LINC00675</i>	rs1078643*	17	10,707,241	A/G	76.36	1.07	1.04-1.10	$9.2 \times 10^{-6}$	1.09	1.05-1.12	$1.1 \times 10^{-7}$	1.08	1.05-1.10	$6.6 \times 10^{-12}$
17q24.3	<i>LINC00673</i>	rs983318§	17	70,413,253	A/G	25.26	1.07	1.04-1.10	$1.2 \times 10^{-6}$	1.05	1.02-1.08	$8.0 \times 10^{-4}$	1.06	1.04-1.08	$5.6 \times 10^{-9}$
17q25.3	<i>RAB40B;</i> <i>METRLN</i>	rs75954926*	17	81,061,048	G/A	65.68	1.10	1.07-1.13	$9.4 \times 10^{-11}$	1.09	1.06-1.12	$4.8 \times 10^{-9}$	1.09	1.07-1.11	$3.0 \times 10^{-18}$
19p13.11	<i>KLF2</i>	rs34797592§	19	16,417,198	T/C	11.82	1.09	1.05-1.13	$8.2 \times 10^{-6}$	1.09	1.05-1.13	$1.2 \times 10^{-5}$	1.09	1.06-1.12	$4.2 \times 10^{-10}$
19q13.43	<i>TRIM28</i>	rs73068325	19	59,079,096	T/C	18.26	1.06	1.03-1.09	$2.1 \times 10^{-4}$	1.07	1.04-1.11	$5.0 \times 10^{-5}$	1.07	1.04-1.09	$4.2 \times 10^{-8}$
20q13.12	<i>TOX2;</i> <i>HNF4A</i>	rs6031311§	20	42,666,475	T/C	75.91	1.07	1.04-1.10	$1.7 \times 10^{-6}$	1.05	1.02-1.08	$7.6 \times 10^{-4}$	1.06	1.04-1.08	$6.8 \times 10^{-9}$
20q13.33	<i>TNFRSF6B;</i> <i>RTEL1</i>	rs2738783§,¶	20	62,308,612	T/G	20.29	1.07	1.04-1.10	$2.6 \times 10^{-6}$	1.05	1.02-1.08	$3.3 \times 10^{-3}$	1.06	1.04-1.08	$5.3 \times 10^{-8}$

0 Lead variant is the most associated variant at the locus. rsIDs based on NCBI dbSNP Build 150. Alleles are on the + strand. Chr.: Chromosome. RAF: Risk allele frequency, based  
1 on stage 2 data. OR, odds ratio estimate for the risk allele. All *P*-values reported in this table are based on fixed-effects inverse variance-weighted meta-analysis.

2 \*Indicates that variant or LD proxy ( $r^2 > 0.7$ ) was selected for our custom genotyping panel and formally replicates in the Stage 2 meta-analysis at a Bonferroni significance  
3 threshold of  $P < 7.8 \times 10^{-6}$ .

4 §Indicates that variant or LD proxy ( $r^2 > 0.7$ ) was selected for our custom genotyping panel but did not attain Bonferroni significance in the Stage 2 meta-analysis.

5 ¶This SNP reached genome-wide significance in the combined (Stage 1 + Stage 2) sample-size weighted meta-analysis based on likelihood ratio test results ( $P = 4.9 \times 10^{-8}$ ).

6  
7  
8  
9  
0  
1  
2  
3  
4  
5  
6  
7  
8  
9  
0  
1  
2  
3  
4  
5

6 **Table 2 Additional new conditionally independent association signals at known and newly identified CRC risk loci that reach genome-wide**  
7 **significance ( $P < 5 \times 10^{-8}$ ) in the combined meta-analysis of up to 125,478 individuals.**

Locus	Nearby gene(s)	rsID lead variant	Chr.	Position (Build 37)	Alleles (risk/other)	RAF (%)	OR <sub>unconditional</sub>	95% CI	$P_{\text{unconditional}}$	Joint multiple-variant analysis			
										Conditioning variant(s)	OR <sub>conditional</sub>	95% CI	$P_{\text{conditional}}$
<b>Low-frequency variants</b>													
11q13.4	<i>POLD3</i>	rs61389091	11	74,427,921	C/T	96.06	1.23	1.18-1.29	$1.2 \times 10^{-18}$	rs7121958*, rs7946853	1.21	1.16-1.27	$3.7 \times 10^{-16}$
<b>Common variants</b>													
2q33.1	<i>SATB2</i>	rs11884596	2	199,612,407	C/T	38.23	1.06	1.04-1.08	$1.1 \times 10^{-9}$	rs983402	1.06	1.04-1.07	$3.6 \times 10^{-9}$
5p15.33	<i>TERT</i> ; <i>CLPTMIL</i>	rs78368589	5	1,240,204	T/C	5.97	1.14	1.10-1.18	$9.4 \times 10^{-12}$	rs2735940*	1.12	1.08-1.16	$4.1 \times 10^{-9}$
5p13.1	<i>LINC00603</i> ; <i>PTGER4</i>	rs7708610	5	40,102,443	A/G	35.64	1.04	1.02-1.06	$1.5 \times 10^{-5}$	rs12514517*	1.06	1.04-1.08	$3.8 \times 10^{-9}$
6p21.32	<i>HLA-B</i> ; <i>MICA</i> ; <i>MICB</i> ; <i>NFKBIL1</i> ; <i>TNF</i>	rs2516420	6	31,449,620	C/T	92.63	1.10	1.06-1.13	$1.3 \times 10^{-7}$	rs9271695, rs116685461, rs116353863	1.12	1.08-1.16	$2.0 \times 10^{-10}$
8q24.21	<i>MYC</i>	rs4313119	8	128,571,855	G/T	74.86	1.06	1.04-1.08	$1.0 \times 10^{-9}$	rs6983267*, rs7013278	1.06	1.04-1.08	$2.1 \times 10^{-9}$
12p13.32	<i>CCND2</i>	rs3217874	12	4,400,808	T/C	42.82	1.08	1.06-1.10	$1.2 \times 10^{-17}$	rs3217810*, rs35808169*	1.06	1.04-1.08	$2.4 \times 10^{-9}$
15q13.3	<i>GREM1</i>	rs17816465	15	33,156,386	A/G	20.55	1.07	1.04-1.09	$6.8 \times 10^{-9}$	rs2293581*, rs12708491*	1.07	1.05-1.10	$1.4 \times 10^{-10}$
20p12.3	<i>BMP2</i>	rs28488	20	6,762,221	T/C	63.88	1.06	1.04-1.08	$2.6 \times 10^{-11}$	rs189583*, rs4813802*, rs994308	1.07	1.05-1.09	$2.6 \times 10^{-14}$
20p12.3	<i>BMP2</i>	rs994308	20	6,603,622	C/T	59.39	1.08	1.06-1.10	$4.8 \times 10^{-18}$	rs189583*, rs4813802*, rs28488	1.06	1.05-1.08	$8.6 \times 10^{-12}$

8 Lead variant is the most associated variant at the locus in the conditional analysis. rsIDs based on NCBI dbSNP Build 150. Alleles are on the + strand. Chr.: Chromosome. RAF:  
9 Risk allele frequency, based on stage 2 data. OR, odds ratio estimates are for the risk allele. Conditioning variants are the lead variant of other conditionally independent  
0 association signals with  $P < 1 \times 10^{-5}$  within 1-Mb of the new association signal. Because of extensive LD we used a 2-Mb distance for the MHC region (6p21.32). All lead variants  
1 for the new association signals are in linkage equilibrium with any previously reported CRC risk variants at the locus ( $r^2 < 0.10$ ).  
2 \*Indicates that the conditioning variant is either the index variant, or a variant in LD with the index variant reported in previous GWAS. Details and full results are provided in  
3 Supplementary Table 5.  
4

955 **ONLINE METHODS**

956 **Study samples.**

957 After quality control (QC), this study included whole-genome sequencing (WGS) data for 1,439  
958 colorectal cancer (CRC) cases and 720 controls from 5 studies, and GWAS array data for 58,131  
959 CRC or advanced adenoma cases (3,674; 6.3% of cases) and 67,347 controls from 45 studies  
960 from GECCO, CORECT, and CCFR. The Stage 1 meta-analysis comprised existing genotyping  
961 data from 30 studies that were included in previously published CRC GWAS<sup>13,18,22</sup>. After QC,  
962 the Stage 1 meta-analysis included 34,869 cases and 29,051 controls. Study participants were  
963 predominantly of European ancestry (31,843 cases and 26,783 controls; 91.7% of participants).  
964 Because it was shown previously that the vast majority of known CRC risk variants are shared  
965 between Europeans and East Asians<sup>17</sup>, we included 3,026 cases and 2,268 controls of East Asian  
966 ancestry to increase power for discovery. The Stage 2 meta-analysis comprised newly generated  
967 genotype data involving 4 genotyping projects and 22 studies. After QC, the Stage 2 meta-  
968 analysis included 23,262 cases and 38,296 controls, all of European ancestry. Studies, sample  
969 selection, and matching are described in the **Supplementary Text. Supplementary Table 1**  
970 provides details on sample numbers, and demographic characteristics of study participants. All  
971 participants provided written informed consent, and each study was approved by the relevant  
972 research ethics committee or institutional review board. Four normal colon mucosa biopsies for  
973 ATAC-seq were obtained from patients with a normal colon at colonoscopy at the Institut  
974 d'Investigació Biomèdica de Bellvitge (IDIBELL), Spain. Patients signed informed consent, and  
975 the protocol was approved by the Bellvitge Hospital Ethics Committee (Colscreen protocol  
976 PR084/16).

977

978 **Whole-genome sequencing.**

979 We performed low-pass WGS of 2,192 samples from 5 studies at the University of Washington  
980 Northwest Genomics Center (Seattle, WA, USA). Cases and controls were processed and  
981 sequenced together. Libraries were prepared with ThruPLEX DNA-seq kits (Rubicon Genomics)  
982 and paired-end sequencing performed using Illumina HiSeq 2500 sequencers. Reads were  
983 mapped to human reference genome (GRCh37 assembly) using Burrows-Wheeler aligner BWA  
984 v0.6.2<sup>82</sup>. Fold genomic coverage averaged 5.3× (range: 3.8-8.6×). We used the GotCloud  
985 population-based multi-sample variant calling pipeline<sup>83</sup> for post-processing of BAM files with

986 initial alignments, and to detect and call single nucleotide variants (SNVs) and short insertions  
987 and deletions (indels). After removing duplicated reads and recalibrating base quality scores, QC  
988 checks included sample contamination detection. Variants were jointly called across all samples.  
989 To identify high-quality sites, the GotCloud pipeline performs a two-step filtering process. First,  
990 lower quality variants are identified by applying individual variant quality statistic filters. Next,  
991 variants failing multiple filters are used as negative examples to train a support vector machine  
992 (SVM) classifier. Finally, we performed a haplotype-aware genotype refinement step via  
993 Beagle<sup>84</sup> and ThunderVCF<sup>85</sup> on the SVM-filtered VCF files. After further sample QC, we  
994 excluded samples with estimated DNA contamination >3% (16), duplicated samples (5) or  
995 related individuals (1), sex discrepancies (0), and samples with low concordance with GWAS  
996 array data (11). We checked for ancestry outliers by performing principal components analysis  
997 (PCA) after merging in data for shared, linkage disequilibrium (LD)-pruned SNVs for 1,092  
998 individuals from the 1000 Genomes Project<sup>86</sup>. After QC, sequences were available for 1,439  
999 CRC cases and 720 controls of European ancestry.

1000

#### 1001 **GWAS genotype data and quality control.**

1002 Details of genotyping and QC for studies included in the Stage 1 meta-analysis are described  
1003 elsewhere<sup>13,18,22</sup>. **Supplementary Table 1** provides details of genotyping platforms used. Before  
1004 association analysis, we pooled individual-level genotype data of all Stage 1 studies for a subset  
1005 of SNPs to enable identification of unexpected duplicates and close relatives. We calculated  
1006 identity by descent (IBD) for each pair of samples using KING-robust<sup>87</sup> and excluded duplicates  
1007 and individuals that are second-degree or more closely related. As part of Stage 2, 28,805  
1008 individuals from 19 studies were newly genotyped on a custom Illumina array based on the  
1009 Infinium OncoArray-500K<sup>26</sup> and a panel of 15,802 successfully manufactured custom variants  
1010 (described in **Supplementary Text**). An additional 8,725 individuals from 5 studies were  
1011 genotyped on the Illumina HumanOmniExpressExome-8v1-2 array. Genotyping and calling for  
1012 both projects were performed at the Center for Inherited Disease Research (CIDR) at Johns  
1013 Hopkins University. Genotypic data that passed initial QC at CIDR subsequently underwent QC  
1014 at the University of Washington Genetic Analysis Center (UW GAC) using standardized  
1015 methods detailed in Laurie *et al.*<sup>88</sup>. The median call rate for the custom Infinium OncoArray-  
1016 500K data was 99.97%, and error rate estimated from 301 sample duplicate pairs was 9.99e-7. A

1017 relatively low number of samples (246) had a missing call rate >2%, with the highest being  
1018 3.48%, and were included in analysis. For the HumanOmniExpressExome-8v1-2 data, median  
1019 call rate was 99.96%, and the error rate estimated from 179 sample duplicate pairs was 2.65e-6.  
1020 Thirty samples had a missing call rate >2%, with the highest being 3.79%, and were included in  
1021 analysis. We excluded samples with discrepancies between reported and genotypic sex based on  
1022 X chromosome heterozygosity and the means of sex chromosome probe intensities, unintentional  
1023 duplicates, and close relatives defined as individuals that are second-degree or more closely  
1024 related. After further excluding individuals of non-European ancestry as determined by PCA (see  
1025 below), the custom OncoArray data included in analysis comprised 11,852 CRC cases and  
1026 11,895 controls, and the HumanOmniExpressExome-8v1-2 array data included in analysis  
1027 comprised 4,439 CRC cases and 4,115 controls. Only variants passing QC were used for  
1028 imputation. We excluded variants failing CIDR technical filters or UW GAC quality filters,  
1029 which included missing call rate >2%, discordant calls in sample duplicates, and departures from  
1030 Hardy-Weinberg equilibrium (HWE) ( $P < 1e-4$ ) based on European-ancestry controls. The Stage  
1031 2 analysis also included genotype data from the CORSA study (**Supplementary Text**). In total,  
1032 2,354 individuals were genotyped using the Affymetrix Axiom Genome-Wide Human CEU 1  
1033 Array. We called genotypes using the AxiomGT1 algorithm. All samples had missing call rate  
1034 <3%. We excluded samples with discrepancies between reported and genotypic sex (20), close  
1035 relatives defined as individuals that are second-degree or more closely related (94), as inferred  
1036 using KING-robust<sup>87</sup>, and individuals of non-European ancestry (6) as inferred from PCA. After  
1037 QC, data included in analysis comprised 1,460 cases and 774 controls. Prior to phasing and  
1038 imputation, we filtered out SNPs with missing call rate >2%, or HWE  $P < 1e-4$ . Imputed  
1039 genotype data were obtained from UK Biobank and QC and imputation are described  
1040 elsewhere<sup>89</sup>. A nested case-control dataset was constructed as described in the **Supplementary**  
1041 **Text**. We excluded individuals of non-European ancestry as inferred from PCA, and randomly  
1042 dropped one individual from each pair that were more closely related than third-degree relatives  
1043 as inferred using KING-robust. This resulted in excluding 137 samples. In total, 5,356 CRC  
1044 (5,004) or advanced adenoma (352) cases and 21,407 matched controls were included in the  
1045 replication analysis.

1046

1047 **Principal components analysis.**

1048 After excluding close relatives, we performed PCA using PLINK1.9<sup>90</sup> on LD-pruned sets of  
1049 autosomal SNPs obtained by removing regions with extensive long-range LD<sup>91,92</sup>, SNPs with  
1050 minor allele frequency (MAF) <5%, or HWE  $P < 1e-4$ , or any missingness, and carrying out LD  
1051 pruning using the PLINK option ‘-indep-pairwise 50 5 0.2’. To identify population outliers we  
1052 merged in 1,092 individuals from 1000 Genomes Project Phase III and performed PCA using the  
1053 intersection of variants<sup>93</sup>.

1054

### 1055 **Genotype imputation.**

1056 The 2,159 whole-genome sequences described above were used to create a phased imputation  
1057 reference panel. After estimating haplotypes for all GWAS array data sets using SHAPEIT2<sup>94</sup>,  
1058 we used minimac3<sup>95</sup> to impute from this reference panel (19.6 million variants with minor allele  
1059 count (MAC) >1) into the GWAS datasets described above. We also imputed to the Haplotype  
1060 Reference Consortium (HRC) panel<sup>25</sup> (39.2 million variants) using the University of Michigan  
1061 Imputation Server<sup>95</sup>. To improve imputation accuracy for Stage 1 data sets, phasing and  
1062 imputation were performed after pooling studies/genotype projects that used the same, or very  
1063 similar, genotyping platforms (**Supplementary Table 1**). For Stage 2, we performed phasing  
1064 and imputation separately for each genotyping project data set and imputed to the HCR panel.

1065

### 1066 **Statistical analyses.**

1067 *Association testing of sequence data.*

1068 We tested variants with  $MAC \geq 5$  for CRC association using Firth’s bias-reduced logistic  
1069 regression as implemented in EPACTS ([genome.sph.umich.edu/wiki/EPACTS](http://genome.sph.umich.edu/wiki/EPACTS)) and adjusted for  
1070 sex, age, study, and 3 principal components (PCs) calculated from an LD-pruned set of  
1071 genotypes. We performed rare variant aggregate tests at the gene and enhancer level using the  
1072 Mixed effects Score Test (MiST)<sup>96</sup>. This unified test is a linear combination between  
1073 unidirectional burden and bidirectional variance component tests that performs best in terms of  
1074 statistical power across a range of architectures<sup>97</sup>.

1075

1076 *Association and meta-analysis.*

1077 Stage 1 comprised two large mega-analyses of pooled individual-level genotype data sets  
1078 (**Supplementary Table 12**). The four Stage 2 genotyping project data sets were analyzed

1079 separately. Within each data set, variants with an imputation accuracy  $r^2 \geq 0.3$  and MAC  $\geq 50$   
1080 were tested for CRC association using the imputed genotype dosage in a logistic regression  
1081 model adjusted for age, sex, and study/genotyping project-specific covariates, including PCs to  
1082 adjust for population structure (**Supplementary Table 12**). To account for residual confounding  
1083 within CORSA, we tested association with each variant using a linear mixed model and kinship  
1084 matrix calculated from the data, as implemented in EMMA<sup>98</sup>. To enable meta-analysis, we then  
1085 calculated approximate allelic log odds ratios (OR) and corresponding standard errors as  
1086 described in Cook *et al.*<sup>99</sup>.

1087 Next, we combined association summary statistics across analyses via fixed-effects inverse  
1088 variance-weighted meta-analysis. Because Wald tests can be notably anti-conservative for rare  
1089 variant associations, we also performed likelihood ratio-based tests, followed by sample-size  
1090 weighted meta-analysis, as implemented in METAL<sup>100</sup>. In total, 16,900,397 variants were  
1091 analyzed. To examine residual population stratification, we inspected quantile-quantile plots of  
1092 test statistics (**Supplementary Figure 8**), and calculated genomic control inflation statistics  
1093 ( $\lambda_{GC}$ ).  $\lambda_{GC}$  for the combined meta-analysis was 1.105, and for Stage 1 and 2 meta-analyses was  
1094 1.071 and 1.075, respectively. Because  $\lambda_{GC}$  increases with sample size for polygenic phenotypes,  
1095 even in the absence of confounding biases<sup>101</sup>, we investigated the effect of confounding due to  
1096 residual population stratification using LD score regression<sup>102</sup>. Because of limitations of LD  
1097 score regression, this analysis is restricted to common variants (MAF  $\geq 1\%$ ) for which  $\lambda_{GC}$  was  
1098 1.188 in the combined meta-analysis. The LD score regression intercept was 1.067, which is  
1099 substantially less than  $\lambda_{GC}$ , indicating at most a small contribution of bias and that inflation in  $\chi^2$   
1100 statistics results mostly from polygenicity. We also calculated  $\lambda_{1,000}$  which is the equivalent  
1101 inflation statistic for a study with 1,000 cases and 1,000 controls<sup>103</sup>. For the combined meta-  
1102 analysis,  $\lambda_{1,000}$  was 1.004 and for both Stage 1 and 2 meta-analyses this was 1.003.

1103

1104 *Significance threshold for the replication genotyping experiment.*

1105 To protect against probe design failure, we built redundancy into the custom genotyping panel by  
1106 including LD proxies of independently associated variants selected for follow-up. To determine  
1107 the number of independent tests, we performed LD clumping of the 9,198 analyzed variants that  
1108 were selected for replication genotyping based on the Stage 1 meta-analysis, and that survived

1109 filters described above. Using an  $r^2$  threshold of 0.1 this translated to 6,438 independent tests and  
1110 a Bonferroni significance threshold of  $0.05/6,438=7.8\times 10^{-6}$ .

1111

1112 *Conditional and joint multiple-variant analysis.*

1113 To identify additional distinct association signals at CRC loci, we performed a series of  
1114 conditional meta-analyses. At each locus attaining  $P < 5\times 10^{-8}$ , we included the genotype dosage  
1115 for the variant showing the strongest statistical evidence for association in the region in the  
1116 combined meta-analysis, as an additional covariate in the respective logistic regression models.  
1117 Association summary statistics for each variant in the region were then combined across studies  
1118 by a fixed-effects meta-analysis. If at least one association signal attained a significance level of  
1119  $P < 1\times 10^{-5}$  in this meta-analysis, we performed a second round of conditional meta-analysis,  
1120 adding the variant showing the strongest statistical evidence for association in the region in the  
1121 first round of conditional meta-analysis as a covariate to the logistic regression models used in  
1122 the first round. We repeated this procedure and kept adding variants to the model until no  
1123 additional variants at the locus attained  $P < 1\times 10^{-5}$ . Finally, we performed a joint multiple-variant  
1124 analysis in which we jointly estimated the effects of variants selected in each step and tested for  
1125 each variant whether the  $P$ -value from the joint multiple-variant analysis ( $P_j$ ) was  $< 1\times 10^{-5}$ .  
1126 Analyses were performed on 2-Mb windows centered on the most associated variant in the  
1127 unconditional analysis. If windows overlapped, we performed the analysis on the collapsed  
1128 genomic region. Because of extensive LD, we used a 4-Mb window for the MHC region.

1129

1130 **Definition of known loci.**

1131 We compiled a list of 62 previously reported genome-wide significant CRC association signals  
1132 from the literature (**Supplementary Table 3**). Because of improved power and coverage of our  
1133 study, we identified the most associated variant at each signal, and used these lead variants for  
1134 further analyses, rather than the previously reported index variant.

1135

1136 **Refinement of association signals.**

1137 To refine new association signals, we constructed credible sets that were 99% likely, based on  
1138 posterior probability, to contain the causal disease-associated SNP<sup>104</sup>. In brief, for each distinct  
1139 signal, we retained a candidate set of variants by identifying all analyzed variants with  $r^2 \geq 0.1$

1140 with the most associated variant within a 2-Mb window centered on the most associated variant.  
1141 We calculated approximate Bayes' factors (ABF)<sup>105</sup> for each variant as:

1142

$$1143 \quad ABF = \sqrt{1-r} e^{rz^2/2}$$

1144

1145 where  $r = 0.04/(s.e.^2+0.04)$ ,  $z = \beta/s.e.$ , and  $\beta$  and s.e. are the log OR estimate and its standard  
1146 error from the combined meta-analysis. For loci with multiple distinct signals, results are based  
1147 on conditional meta-analysis, adjusting for all other index variants in the region. We then  
1148 calculated the posterior probability of being causal as  $ABF/T$  where  $T$  is the sum of ABF values  
1149 over all candidate variants. Next, variants were ranked in decreasing order by posterior  
1150 probabilities and the 99% credible set was obtained by including variants with the highest  
1151 posterior probabilities until the cumulative posterior probability  $\geq 99\%$ .

1152

### 1153 **Functional genomic annotation.**

1154 To nominate variants for future laboratory follow-up, we performed bioinformatic analysis at  
1155 each new signal using our functional annotation database, and a custom UCSC analysis data hub.  
1156 Using ANNOVAR<sup>106</sup>, we annotated lead variants and variants in LD ( $r^2 \geq 0.4$ ) with the lead  
1157 variant, relative to features pertaining to i) gene-centric function (PolyPhen2<sup>107</sup>), ii) genome-  
1158 wide functional prediction scores (CADD<sup>108</sup>, DANN<sup>109</sup>, EigenPC<sup>110</sup>), iii) disease relatedness  
1159 (GWAS catalog<sup>46</sup>), and iv) CRC-relevant regulatory functions (enhancer, repressor, DNA  
1160 accessible, and transcription factor binding site (TFBS)<sup>111,112</sup>; **Supplementary Table 13**).

1161 **Supplementary Table 8** summarizes variant annotations relative to the CCDS Project<sup>113</sup>, and  
1162 reference genome GRCh37. Variants were maintained in **Supplementary Table 8** if they met  
1163 any of the following conditions: DANN score  $\geq 0.9$ , CADD phred score  $\geq 20$ , Eigen-PC phred  
1164 score  $\geq 17$ , PolyPhen2 “probably damaging”, “stop loss”, “stop gain”, “splicing”, or were  
1165 positioned in a predicted regulatory element. We visually inspected loci overlapping with CRC-  
1166 relevant functional genomic annotations. Variants positioned in enhancers with aberrant CRC  
1167 activity were identified by comparing epigenomes of non-diseased colorectal tissues/colonic  
1168 crypt cells to epigenomes of primary CRC cell lines (data accessible at NCBI GEO database,  
1169 accession GSE77737). We prioritized target genes for loci with predicted regulatory function.  
1170 Evidence suggests that Topological Association Domains (TADs) can be used to map physical

1171 boundaries on gene promoter interactions with distal regulatory elements<sup>114-116</sup>. As such, we used  
1172 GMI12878 Hi-C Chromosome Conformation Capture data to identify gene promoters that were  
1173 in the same TADs as risk loci using the WashU Epigenome Browser  
1174 (<https://epigenomegateway.wustl.edu/>). Genes in this list were further prioritized based on  
1175 biological relevancy and expression quantitative trait loci (eQTL) data from Genotype-Tissue  
1176 Expression (GTEx)<sup>117</sup> using HaploReg v4.1<sup>118</sup>.

1177

#### 1178 **ATAC-seq assay.**

1179 We generated high resolution maps of DNA accessible regions in normal colon mucosa samples  
1180 using the Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq). Using the  
1181 updated omni-ATAC protocol for archival samples, we performed ATAC-seq in four colon  
1182 mucosa biopsies from the ICO-biobank taken from participants undergoing screening at  
1183 IDIBELL, Spain. Biopsies were cryopreserved by slow freezing using a solution of 10% DMSO,  
1184 90% media, and Mr. Frosty Cryo 1°C Freezing Containers (Thermo Scientific). ATAC-seq was  
1185 implemented as prescribed with two exceptions. Instead of dounce homogenizer we used a tissue  
1186 lyser and stainless bead system, pulverizing at 40Hz for 2 mins and pulsing at 50Hz for 10-20  
1187 seconds. Secondly, Illumina library quantification was performed using picogreen quantitation  
1188 and TapeStation instead of KAPA quantitative qPCR. Libraries were sequenced to an average of  
1189 25M paired end reads using Illumina HiSeq 2500. The ENCODE data processing pipeline was  
1190 implemented ([https://github.com/kundajelab/atac\\_dnase\\_pipelines](https://github.com/kundajelab/atac_dnase_pipelines)) aligning to hg19<sup>119</sup>. QC  
1191 results are summarized in **Supplementary Table 14**.

1192

#### 1193 **Regulatory and functional information enrichment analysis.**

1194 We used GARFIELD<sup>74</sup> to identify cell types, tissues, and functional genomic features relevant to  
1195 CRC risk. This method tests for enrichment of association in features primarily extracted from  
1196 ENCODE and Roadmap Epigenomics Project data, while accounting for sources of confounding,  
1197 including LD. We applied default settings and used the author-supplied data which is suitable for  
1198 analysis of GWAS results based on European-ancestry individuals.

1199

#### 1200 **Pathway and gene set enrichment analysis.**

1201 We used MAGENTA to test predefined gene sets (e.g., KEGG pathways) for enrichment for  
1202 CRC risk associations<sup>75</sup>. We used combined meta-analysis results as input and applied default  
1203 settings which included removing genes that fall in the MHC region from analysis. Enrichment  
1204 was tested at two gene  $P$ -value cutoffs: 95th and 75th percentiles of all gene  $P$ -values in the  
1205 genome.

1206

### 1207 **Estimation of contribution of rare variants to heritability.**

1208 We used the LD- and MAF-stratified component GREML (GREML-LDMS) method as  
1209 implemented in GCTA<sup>76</sup> to estimate the proportion of variation in liability for CRC explained by  
1210 all imputed autosomal variants (i.e., estimate of narrow-sense heritability  $h_g^2$ ), and the proportion  
1211 contributed by rare variants (MAF  $\leq 1\%$ ). Because of computational limitations we analyzed a  
1212 subset of 11,895 cases and 14,659 controls imputed to our WGS panel. We analyzed individual-  
1213 level data for 17,649,167 imputed variants with MAC  $> 3$  and HWE test  $P \geq 10^{-6}$ . Following Yang  
1214 *et al.*<sup>76</sup>, we did not filter on imputation quality. In brief, we stratified variants into groups based  
1215 on MAF (boundaries at 0.001, 0.01, 0.1, 0.2, 0.3, 0.4) and mean LD score (boundaries at  
1216 quartiles) calculated as described in Yang *et al.*<sup>76</sup>. We then calculated genetic relationship  
1217 matrices (GRMs) for each of these 28 variant partitions and jointly estimated variance  
1218 components for these partitions, adjusting for age, sex, study, genotyping batch, and three  
1219 genotype PCs. From the variance component estimates and their variance-covariance matrix we  
1220 estimated the contribution of rare variants (MAF  $\leq 1\%$ ) and common variants (MAF  $> 1\%$ ), and  
1221 calculated standard errors using the delta method. We tested significance of the contribution of  
1222 rare variants using a likelihood ratio test. To calculate heritability on the underlying liability  
1223 scale we interpreted  $K$  as lifetime risk<sup>120</sup> and used an estimate of 4.3% (Surveillance,  
1224 Epidemiology, and End Results Program (SEER) Cancer Statistics, 2011-2013).

1225

### 1226 **Familial relative risk explained by genetic variants.**

1227 We assumed a multiplicative model within and between variants and calculated the proportion of  
1228 familial relative risk (RR) explained by a given set of genetic variants as  $\frac{\sum_i \log \lambda_i}{\log \lambda_0}$ , where  $\lambda_0$  is the  
1229 overall familial RR to first-degree relatives of cases.  $\lambda_i$  is the familial RR due to variant  $i$   
1230 calculated as  $\lambda_i = \frac{p_i r_i^2 + q_i}{(p_i r_i + q_i)^2}$ , where  $p_i$  is the risk allele frequency for variant  $i$ ,  $q_i = 1 - p_i$ , and  $r_i$

1231 is the estimated per allele OR<sup>9,121</sup>. We adjusted the OR estimates of new association signals for  
 1232 winner's curse following Zhong and Prentice<sup>78</sup>. We represented previously identified association  
 1233 signals by the variant showing the strongest statistical evidence of association in the combined  
 1234 meta-analysis, and assumed that winner's curse was negligible. We assumed  $\lambda_0$  to be 2.2<sup>122</sup>.  
 1235 Using the delta method, we computed the variance for the proportion of familial RR as follows:  
 1236

$$1237 \quad \sum_i \text{Var}(r_i) \left[ \frac{1}{\log \lambda_0} \frac{1}{\lambda_i} \frac{2p_i q_i (r_i - 1)}{(p_i r_i + q_i)^3} \right]^2.$$

1238  
 1239 **Absolute risk of CRC incidence and starting age of first screening.**

1240 We constructed a polygenic risk score (PRS) as a weighted sum of expected risk allele frequency  
 1241 for common genetic variants, using the per allele OR for each variant as weights. OR estimates  
 1242 for newly discovered variants were adjusted for winner's curse to avoid potential inflation<sup>78</sup>.  
 1243 Assuming all genetic variants are independent, let  $X$  denote a PRS constructed based on  $K$   
 1244 variants:  $X = \sum_{i=1}^K \hat{\beta}_i Z_i$ , where  $\hat{\beta}_i$  and  $Z_i$  are the estimated OR and the number of risk alleles for  
 1245 variant  $i$ . We assumed  $X$  follows a normal distribution  $N(\mu, \sigma^2)$ , where the estimates of mean  
 1246 and variance are computed as following:

$$1247 \quad \hat{\mu} = \sum_{i=1}^K \hat{\beta}_i \times 2 \times \hat{p}_i \text{ and } \hat{\sigma}^2 = \sum_{i=1}^K \hat{\beta}_i^2 \times 2 \times \hat{p}_i \times (1 - \hat{p}_i),$$

1248 where  $\hat{p}_i$  is the risk allele frequency for variant  $i = 1, \dots, K$ . Then the baseline hazard at each  
 1249 age  $t$ ,  $\widehat{\lambda}_0(t)$ , is computed as following:

$$1250 \quad \widehat{\lambda}_0(1) = \lambda^*(1) \frac{\int f(x) dx}{\int e^x f(x) dx}$$

$$1251 \quad \widehat{\lambda}_0(t) = \lambda^*(t) \frac{\int \exp(-\sum_{i=1}^{t-1} \widehat{\lambda}_0(i) e^x) f(x) dx}{\int \exp(-\sum_{i=1}^{t-1} \widehat{\lambda}_0(i) e^x) e^x f(x) dx} \text{ for } t = 2, \dots, 100,$$

1252 and  $\lambda^*(t)$  are the incidence rates for non-Hispanic whites who have not taken an endoscopy  
 1253 before, derived from population incidence rates during 1992-2005 from the SEER Registry.  
 1254 Using these baseline hazard rates, we estimated the 10-year absolute risk of developing CRC  
 1255 given age and a PRS as previously described<sup>123</sup>. By setting a risk threshold as the average of the  
 1256 10-year CRC risk for a 50-year old man (1.25%) and woman (0.68%), i.e.,  
 1257  $(1.25\% + 0.68\%) / 2 = 0.97\%$ , who have not previously received an endoscopy<sup>124</sup>, we estimated the

1258 recommended starting age of first screening given the PRS. Variants and OR estimates used in  
1259 these analyses are given in **Supplementary Table 15**.

1260

#### 1261 **Data availability.**

1262 All whole-genome sequence data have been deposited at the database of Genotypes and  
1263 Phenotypes (dbGaP), which is hosted by the U.S. National Center for Biotechnology Information  
1264 (NCBI), under accession number phs001554.v1.p1. All custom Infinium OncoArray-500K array  
1265 data for the studies in the Stage 2 meta-analysis have been deposited at dbGaP under accession  
1266 number phs001415.v1.p1. All Illumina HumanOmniExpressExome-8v1-2 array data for the  
1267 studies in the Stage 2 meta-analysis have been deposited at dbGaP under accession number  
1268 phs001315.v1.p1. Genotype data for the studies included in the Stage 1 meta-analysis have been  
1269 deposited at dbGaP under accession number phs001078.v1.p1. The UK Biobank resource was  
1270 accessed through application number 8614.

1271

#### 1272 **Reporting Summary.**

1273 Further information on experimental design is available in the Life Sciences Reporting Summary  
1274 linked to this article.

1275

#### 1276 **METHODS-ONLY REFERENCES**

1277

- 1278 82. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler  
1279 transform. *Bioinformatics* **25**, 1754–1760 (2009).
- 1280 83. Jun, G., Wing, M. K., Abecasis, G. R. & Kang, H. M. An efficient and scalable analysis  
1281 framework for variant extraction and refinement from population-scale DNA sequence  
1282 data. *Genome Res* **25**, 918–925 (2015).
- 1283 84. Browning, B. L. & Yu, Z. Simultaneous genotype calling and haplotype phasing improves  
1284 genotype accuracy and reduces false-positive associations for genome-wide association  
1285 studies. *Am J Hum Genet* **85**, 847–861 (2009).
- 1286 85. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping  
1287 and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**,  
1288 2987–2993 (2011).
- 1289 86. 1000 Genomes Project Consortium *et al.* A map of human genome variation from  
1290 population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
- 1291 87. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies.  
1292 *Bioinformatics* **26**, 2867–2873 (2010).
- 1293 88. Laurie, C. C. *et al.* Quality control and quality assurance in genotypic data for genome-

- 1294 wide association studies. *Genet Epidemiol* **34**, 591–602 (2010).
- 1295 89. Bycroft, C. *et al.* Genome-wide genetic data on ~500,000 UK Biobank participants.  
1296 *BioRxiv* (2017). doi:10.1101/166298
- 1297 90. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer  
1298 datasets. *Gigascience* **4**, 7 (2015).
- 1299 91. Price, A. L. *et al.* Long-range LD can confound genome scans in admixed populations. *Am*  
1300 *J Hum Genet* **83**, 132–135 (2008).
- 1301 92. Weale, M. E. Quality control for genome-wide association studies. *Methods Mol Biol* **628**,  
1302 341–372 (2010).
- 1303 93. 1000 Genomes Project Consortium *et al.* An integrated map of genetic variation from  
1304 1,092 human genomes. *Nature* **491**, 56–65 (2012).
- 1305 94. Delaneau, O., Howie, B., Cox, A. J., Zagury, J.-F. & Marchini, J. Haplotype estimation  
1306 using sequencing reads. *Am J Hum Genet* **93**, 687–696 (2013).
- 1307 95. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat Genet* **48**,  
1308 1284–1287 (2016).
- 1309 96. Sun, J., Zheng, Y. & Hsu, L. A unified mixed-effects model for rare-variant association in  
1310 sequencing studies. *Genet Epidemiol* **37**, 334–344 (2013).
- 1311 97. Moutsianas, L. *et al.* The power of gene-based rare variant methods to detect disease-  
1312 associated variation and test hypotheses about complex disease. *PLoS Genet* **11**, e1005165  
1313 (2015).
- 1314 98. Kang, H. M. *et al.* Variance component model to account for sample structure in genome-  
1315 wide association studies. *Nat Genet* **42**, 348–354 (2010).
- 1316 99. Cook, J. P., Mahajan, A. & Morris, A. P. Guidance for the utility of linear models in meta-  
1317 analysis of genetic association studies of binary phenotypes. *Eur J Hum Genet* **25**, 240–245  
1318 (2017).
- 1319 100. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of  
1320 genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
- 1321 101. Yang, J. *et al.* Genomic inflation factors under polygenic inheritance. *Eur J Hum Genet* **19**,  
1322 807–812 (2011).
- 1323 102. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from  
1324 polygenicity in genome-wide association studies. *Nat Genet* **47**, 291–295 (2015).
- 1325 103. Michailidou, K. *et al.* Large-scale genotyping identifies 41 new loci associated with breast  
1326 cancer risk. *Nat Genet* **45**, 353–61, 361e1 (2013).
- 1327 104. Wellcome Trust Case Control Consortium *et al.* Bayesian refinement of association signals  
1328 for 14 loci in 3 common diseases. *Nat Genet* **44**, 1294–1301 (2012).
- 1329 105. Wakefield, J. A Bayesian measure of the probability of false discovery in genetic  
1330 epidemiology studies. *Am J Hum Genet* **81**, 208–227 (2007).
- 1331 106. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants  
1332 from high-throughput sequencing data. *Nucleic Acids Res* **38**, e164 (2010).
- 1333 107. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human  
1334 missense mutations using PolyPhen-2. *Curr Protoc Hum Genet* **Chapter 7**, Unit7.20  
1335 (2013).
- 1336 108. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human  
1337 genetic variants. *Nat Genet* **46**, 310–315 (2014).
- 1338 109. Quang, D., Chen, Y. & Xie, X. DANN: a deep learning approach for annotating the  
1339 pathogenicity of genetic variants. *Bioinformatics* **31**, 761–763 (2015).

- 1340 110. Ionita-Laza, I., McCallum, K., Xu, B. & Buxbaum, J. D. A spectral approach integrating  
1341 functional genomic annotations for coding and noncoding variants. *Nat Genet* **48**, 214–220  
1342 (2016).
- 1343 111. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human  
1344 epigenomes. *Nature* **518**, 317–330 (2015).
- 1345 112. Corradin, O. *et al.* Combinatorial effects of multiple enhancer variants in linkage  
1346 disequilibrium dictate levels of gene expression to confer susceptibility to common traits.  
1347 *Genome Res* **24**, 1–13 (2014).
- 1348 113. Pruitt, K. D. *et al.* The consensus coding sequence (CCDS) project: Identifying a common  
1349 protein-coding gene set for the human and mouse genomes. *Genome Res* **19**, 1316–1323  
1350 (2009).
- 1351 114. Harmston, N. *et al.* Topologically associating domains are ancient features that coincide  
1352 with Metazoan clusters of extreme noncoding conservation. *Nat Commun* **8**, 441 (2017).
- 1353 115. Berlivet, S. *et al.* Clustering of tissue-specific sub-TADs accompanies the regulation of  
1354 HoxA genes in developing limbs. *PLoS Genet* **9**, e1004018 (2013).
- 1355 116. Hu, Z. & Tee, W.-W. Enhancers and chromatin structures: regulatory hubs in gene  
1356 expression and diseases. *Biosci Rep* **37**, (2017).
- 1357 117. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580–  
1358 585 (2013).
- 1359 118. Ward, L. D. & Kellis, M. HaploReg: a resource for exploring chromatin states,  
1360 conservation, and regulatory motif alterations within sets of genetically linked variants.  
1361 *Nucleic Acids Res* **40**, D930-4 (2012).
- 1362 119. Landt, S. G. *et al.* ChIP-seq guidelines and practices of the ENCODE and modENCODE  
1363 consortia. *Genome Res* **22**, 1813–1831 (2012).
- 1364 120. Witte, J. S., Visscher, P. M. & Wray, N. R. The contribution of genetic variants to disease  
1365 depends on the ruler. *Nat Rev Genet* **15**, 765–776 (2014).
- 1366 121. Cox, A. *et al.* A common coding variant in CASP8 is associated with breast cancer risk.  
1367 *Nat Genet* **39**, 352–358 (2007).
- 1368 122. Johns, L. E. & Houlston, R. S. A systematic review and meta-analysis of familial colorectal  
1369 cancer risk. *Am J Gastroenterol* **96**, 2992–3003 (2001).
- 1370 123. Hsu, L. *et al.* A model to determine colorectal cancer risk using common genetic  
1371 susceptibility loci. *Gastroenterology* **148**, 1330–9.e14 (2015).
- 1372 124. Jeon, J. *et al.* Determining risk of colorectal cancer and starting age of screening based on  
1373 lifestyle, environmental, and genetic factors. *Gastroenterology* **154**, 2152–2164.e19 (2018).





