

1 **Genomic analyses identify hundreds of variants associated with age at menarche and**  
2 **support a role for puberty timing in cancer risk**

3 Felix R. Day<sup>\*1</sup>, Deborah J. Thompson<sup>\*2</sup>, Hannes Helgason<sup>\*3,4</sup>, Daniel I. Chasman<sup>5,6</sup>, Hilary  
4 Finucane<sup>7,8</sup>, Patrick Sulem<sup>3</sup>, Katherine S. Ruth<sup>9</sup>, Sean Whalen<sup>10</sup>, Abhishek K. Sarkar<sup>11,12</sup>,  
5 Eva Albrecht<sup>13</sup>, Elisabeth Altmaier<sup>14,15</sup>, Marzyeh Amini<sup>16</sup>, Caterina M. Barbieri<sup>17</sup>, Thibaud  
6 Boutin<sup>18</sup>, Archie Campbell<sup>19</sup>, Ellen Demerath<sup>20</sup>, Ayush Giri<sup>21,22</sup>, Chunyan He<sup>23,24</sup>, Jouke J.  
7 Hottenga<sup>25</sup>, Robert Karlsson<sup>26</sup>, Ivana Kolcic<sup>27</sup>, Po-Ru Loh<sup>7,28</sup>, Kathryn L. Lunetta<sup>29,30</sup>,  
8 Massimo Mangino<sup>31,32</sup>, Brumat Marco<sup>33</sup>, George McMahon<sup>34</sup>, Sarah E. Medland<sup>35</sup>, Ilja M.  
9 Nolte<sup>16</sup>, Raymond Noordam<sup>36</sup>, Teresa Nutile<sup>37</sup>, Lavinia Paternoster<sup>34,38</sup>, Natalia Perjakova<sup>39</sup>,  
10 Eleonora Porcu<sup>40</sup>, Lynda M. Rose<sup>5</sup>, Katharina E. Schraut<sup>41,42</sup>, Ayellet V. Segre<sup>43</sup>, Albert V.  
11 Smith<sup>44,45</sup>, Lisette Stolk<sup>46</sup>, Alexander Teumer<sup>47</sup>, Irene L. Andrulis<sup>48,49</sup>, Stefania Bandinelli<sup>50</sup>,  
12 Matthias W. Beckmann<sup>51</sup>, Javier Benitez<sup>52,53</sup>, Sven Bergmann<sup>54,55</sup>, Murielle Bochud<sup>56</sup>, Eric  
13 Boerwinkle<sup>57</sup>, Stig E. Bojesen<sup>58-60</sup>, Manjeet K. Bolla<sup>2</sup>, Judith S. Brand<sup>26</sup>, Hiltrud Brauch<sup>61-63</sup>,  
14 Hermann Brenner<sup>63-65</sup>, Linda Broer<sup>46</sup>, Thomas Brüning<sup>66</sup>, Julie E. Buring<sup>5,6</sup>, Harry Campbell<sup>42</sup>,  
15 Eulalia Catamo<sup>67</sup>, Stephen Chanock<sup>68</sup>, Georgia Chenevix-Trench<sup>69</sup>, Tanguy Corre<sup>54-56</sup>,  
16 Fergus J. Couch<sup>70</sup>, Diana L. Cousminer<sup>71,72</sup>, Angela Cox<sup>73</sup>, Laura Crisponi<sup>40</sup>, Kamila Czene<sup>26</sup>,  
17 George Davey-Smith<sup>34,38</sup>, Eco J.C.N de Geus<sup>25</sup>, Renée de Mutser<sup>74</sup>, Immaculata De Vivo<sup>7,75</sup>,  
18 Joe Dennis<sup>2</sup>, Peter Devilee<sup>76,77</sup>, Isabel dos-Santos-Silva<sup>78</sup>, Alison M. Dunning<sup>79</sup>, Johan G.  
19 Eriksson<sup>80</sup>, Peter A. Fasching<sup>51,81</sup>, Lindsay Fernández-Rhodes<sup>82</sup>, Luigi Ferrucci<sup>83</sup>, Dieter  
20 Flesch-Janys<sup>84,85</sup>, Lude Franke<sup>86</sup>, Marike Gabrielson<sup>26</sup>, Ilaria Gandin<sup>33</sup>, Graham G. Giles<sup>87,88</sup>,  
21 Harald Grallert<sup>14,15,89</sup>, Daniel F. Gudbjartsson<sup>3,4</sup>, Pascal Guénel<sup>90</sup>, Per Hall<sup>26</sup>, Emily  
22 Hallberg<sup>91</sup>, Ute Hamann<sup>92</sup>, Tamara B. Harris<sup>93</sup>, Catharina A. Hartman<sup>94</sup>, Gerardo Heiss<sup>82</sup>,  
23 Maartje J. Hooning<sup>95</sup>, John L. Hopper<sup>88</sup>, Frank Hu<sup>75,96</sup>, David Hunter<sup>7,75,96</sup>, M. Arfan Ikram<sup>97</sup>,  
24 Hae Kyung Im<sup>98</sup>, Marjo-Riitta Järvelin<sup>99-103</sup>, Peter K. Joshi<sup>42</sup>, David Karasik<sup>6,104</sup>, Zoltan  
25 Kutalik<sup>54,56</sup>, Genevieve LaChance<sup>31</sup>, Diether Lambrechts<sup>105,106</sup>, Claudia Langenberg<sup>1</sup>, Lenore  
26 J. Launer<sup>93</sup>, Joop S.E. Laven<sup>107</sup>, Stefania Lenarduzzi<sup>67</sup>, Jingmei Li<sup>26</sup>, Penelope A. Lind<sup>35</sup>,  
27 Sara Lindstrom<sup>108</sup>, YongMei Liu<sup>109</sup>, Jian'an Luan<sup>1</sup>, Reedik Mägi<sup>39</sup>, Arto Mannermaa<sup>110-112</sup>,  
28 Hamdi Mbarek<sup>25</sup>, Mark I. McCarthy<sup>113-115</sup>, Christa Meisinger<sup>14,116</sup>, Thomas Meitinger<sup>117</sup>,  
29 Cristina Menni<sup>31</sup>, Andres Metspalu<sup>39</sup>, Kyriaki Michailidou<sup>2,118</sup>, Lili Milani<sup>39</sup>, Roger L. Milne<sup>87,88</sup>,  
30 Grant W. Montgomery<sup>119</sup>, Anna M. Mulligan<sup>120,121</sup>, Mike A. Nalls<sup>122</sup>, Pau Navarro<sup>18</sup>, Heli  
31 Nevanlinna<sup>123</sup>, Dale R. Nyholt<sup>124</sup>, Albertine J. Oldehinkel<sup>125</sup>, Tracy A. O'Mara<sup>69</sup>, Sandosh  
32 Padmanabhan<sup>126</sup>, Aarno Palotie<sup>28,127-131</sup>, Nancy Pedersen<sup>26</sup>, Annette Peters<sup>14,89</sup>, Julian  
33 Peto<sup>78</sup>, Paul D.P. Pharoah<sup>2,79</sup>, Anneli Pouta<sup>132</sup>, Paolo Radice<sup>133</sup>, Iffat Rahman<sup>134</sup>, Susan M.  
34 Ring<sup>34,38</sup>, Antonietta Robino<sup>67</sup>, Frits R. Rosendaal<sup>74</sup>, Igor Rudan<sup>42</sup>, Rico Rueedi<sup>54,55</sup>, Daniela  
35 Ruggiero<sup>37</sup>, Cinzia F. Sala<sup>17</sup>, Marjanka K. Schmidt<sup>135,136</sup>, Robert A. Scott<sup>1</sup>, Mitul Shah<sup>79</sup>,  
36 Rossella Sorice<sup>37</sup>, Melissa C. Southey<sup>137</sup>, Ulla Sovio<sup>99,138</sup>, Meir Stampfer<sup>7,75</sup>, Maristella  
37 Steri<sup>40</sup>, Konstantin Strauch<sup>13,139</sup>, Toshiko Tanaka<sup>83</sup>, Emmi Tikkanen<sup>131,140</sup>, Nicholas J.  
38 Timpson<sup>34,38</sup>, Michela Traglia<sup>17</sup>, Thérèse Truong<sup>90</sup>, Jonathan P. Tyrer<sup>79</sup>, André G.  
39 Uitterlinden<sup>46,97</sup>, Digna R. Velez Edwards<sup>22,141,142</sup>, Veronique Vitart<sup>18</sup>, Uwe Völker<sup>143</sup>, Peter  
40 Vollenweider<sup>144</sup>, Qin Wang<sup>2</sup>, Elisabeth Widen<sup>131</sup>, Ko Willems van Dijk<sup>77,145,146</sup>, Gonkeke  
41 Willemsen<sup>25</sup>, Robert Winqvist<sup>147,148</sup>, Bruce H.R Wolffenbuttel<sup>149</sup>, Jing Hua Zhao<sup>1</sup>, Magdalena  
42 Zoledziewska<sup>40</sup>, Marek Zygmunt<sup>150</sup>, Behrooz Z. Alizadeh<sup>16</sup>, Dorret I. Boomsma<sup>25</sup>, Marina  
43 Ciullo<sup>37</sup>, Francesco Cucca<sup>40,151</sup>, Tõnu Esko<sup>28,39</sup>, Nora Franceschini<sup>82</sup>, Christian Gieger<sup>14,15,89</sup>,  
44 Vilmundur Gudnason<sup>44,45</sup>, Caroline Hayward<sup>18</sup>, Peter Kraft<sup>7,152</sup>, Debbie A. Lawlor<sup>34,38</sup>, Patrik  
45 K.E Magnusson<sup>26</sup>, Nicholas G. Martin<sup>35</sup>, Dennis O. Mook-Kanamori<sup>74,153</sup>, Ellen A. Nohr<sup>154</sup>,  
46 Ozren Polasek<sup>27</sup>, David Porteous<sup>19</sup>, Alkes L. Price<sup>7,8,28</sup>, Paul M. Ridker<sup>5,6</sup>, Harold Snieder<sup>16</sup>,  
47 Tim D. Spector<sup>31</sup>, Doris Stöckl<sup>14,155</sup>, Daniela Toniolo<sup>17</sup>, Sheila Ulivi<sup>67</sup>, Jenny A. Visser<sup>46</sup>,

This is the author's manuscript of the article published in final edited form as:

Day, F. R., Thompson, D. J., Helgason, H., McCarthy, M., LifeLines Cohort Study, InterAct Consortium, ... & PRACTICAL consortium. (2017). Genomic analyses identify hundreds of variants associated with age at menarche and support a role for puberty timing in cancer risk. <https://doi.org/10.1038/ng.3841>

48 Henry Völzke<sup>47</sup>, Nicholas J. Wareham<sup>1</sup>, James F. Wilson<sup>18,42</sup>, The LifeLines Cohort Study<sup>156</sup>,  
49 The InterAct Consortium<sup>156</sup>, kConFab/AOCS Investigators<sup>156</sup>, Endometrial Cancer  
50 Association Consortium<sup>156</sup>, Ovarian Cancer Association Consortium<sup>156</sup>, PRACTICAL  
51 consortium<sup>156</sup>, Amanda B. Spurdle<sup>69</sup>, Unnur Thorsteindottir<sup>3,44</sup>, Katherine S. Pollard<sup>10,157</sup>,  
52 Douglas F. Easton<sup>2,79</sup>, Joyce Y. Tung<sup>158</sup>, Jenny Chang-Claude<sup>159,160</sup>, David Hinds<sup>158</sup>, Anna  
53 Murray<sup>9</sup>, Joanne M. Murabito<sup>30,161</sup>, Kari Stefansson<sup>\*3,44</sup>, Ken K. Ong<sup>\*1,162</sup> and John R.B  
54 Perry<sup>\*1</sup>

55 \* denotes equal contribution

## 56 Affiliations

- 57 1. MRC Epidemiology Unit, University of Cambridge School of Clinical Medicine, Box 285
- 58 Institute of Metabolic Science, Cambridge Biomedical Campus, Cambridge, CB2 0QQ, UK.
- 59 2. Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care,
- 60 University of Cambridge, CB1 8RN, UK.
- 61 3. deCODE genetics/Amgen, Inc., IS-101 Reykjavik, Iceland.
- 62 4. School of Engineering and Natural Sciences, University of Iceland, IS-101 Reykjavik,
- 63 Iceland,.
- 64 5. Division of Preventive Medicine, Brigham and Women's Hospital, Boston, MA 02215.
- 65 6. Harvard Medical School, Boston, MA 02115, USA.
- 66 7. Department of Epidemiology, Harvard School of Public Health, Boston, MA 02115, USA.
- 67 8. Department of Mathematics, Massachusetts Institute of Technology, Cambridge,
- 68 Massachusetts 02139-4307, USA.
- 69 9. Genetics of Complex Traits, University of Exeter Medical School, University of Exeter,
- 70 Exeter, EX2 5DW, UK.
- 71 10. Gladstone Institutes, San Francisco, California, 94158, USA.
- 72 11. Computer Science and Artificial Intelligence Lab, Massachusetts Institute of Technology,
- 73 Cambridge, MA, USA.
- 74 12. Broad Institute of the Massachusetts Institute of Technology and Harvard University,
- 75 140 Cambridge 02142, MA, USA.
- 76 13. Institute of Genetic Epidemiology, Helmholtz Zentrum München - German Research
- 77 Center for Environmental Health, 85764 Neuherberg, Germany.
- 78 14. Institute of Epidemiology II, Helmholtz Zentrum München - German Research Center for
- 79 Environmental Health, 85764 Neuherberg, Germany.
- 80 15. Research Unit of Molecular Epidemiology, Helmholtz Zentrum München - German
- 81 Research Center for Environmental Health, 85764 Neuherberg, Germany.
- 82 16. Department of Epidemiology, University of Groningen, University Medical Center
- 83 Groningen, Groningen, The Netherlands.
- 84 17. Genetics of Common Disorders Unit, IRCCS San Raffaele Scientific Institute and Vita-
- 85 Salute San Raffaele University, Milan, Italy.
- 86 18. Medical Research Council Human Genetics Unit, Institute of Genetics and Molecular
- 87 Medicine, University of Edinburgh, Edinburgh EH4 2XU, UK.
- 88 19. Medical Genetics Section, Centre for Genomic and Experimental Medicine, Institute of
- 89 Genetics and Molecular Medicine, University of Edinburgh, Edinburgh EH4 2XU, UK.
- 90 20. Division of Epidemiology & Community Health, University of Minnesota, Minneapolis
- 91 MN 55455.
- 92 21. Division of Epidemiology, Institute for Medicine and Public Health, Vanderbilt University,
- 93 Nashville, TN 37235, USA.
- 94 22. Vanderbilt Genetics Institute, Vanderbilt University, Nashville, TN.
- 95 23. Department of Epidemiology, Indiana University Richard M. Fairbanks School of Public
- 96 Health, Indianapolis, IN 46202, USA.
- 97 24. Indiana University Melvin and Bren Simon Cancer Center, Indianapolis, IN 46202, USA.

- 98 25. Department of Biological Psychology, VU University Amsterdam, van der  
99 Boechorststraat 1, 1081 BT, Amsterdam, The Netherlands.
- 100 26. Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, 17177  
101 Stockholm, Sweden.
- 102 27. Faculty of Medicine, University of Split, Split, Croatia.
- 103 28. Program in Medical and Population Genetics, Broad Institute, Cambridge, MA, USA.
- 104 29. Boston University School of Public Health, Department of Biostatistics. Boston,  
105 Massachusetts 02118, USA.
- 106 30. NHLBI's and Boston University's Framingham Heart Study, Framingham,  
107 Massachusetts 01702-5827, USA.
- 108 31. Department of Twin Research and Genetic Epidemiology, King's College London,  
109 London SE1 7EH, UK.
- 110 32. National Institute for Health Research (NIHR) Biomedical Research Centre at Guy's and  
111 St. Thomas' Foundation Trust, London, UK.
- 112 33. Department of Clinical Medical Sciences, Surgical and Health, University of Trieste,  
113 34149 Trieste, Italy.
- 114 34. School of Social and Community Medicine, University of Bristol, Bristol BS8 2BN, UK.
- 115 35. QIMR Berghofer Medical Research Institute, Brisbane, Queensland, Australia.
- 116 36. Department of Internal Medicine, Section Gerontology and Geriatrics, Leiden University  
117 Medical Center, Leiden, the Netherlands.
- 118 37. Institute of Genetics and Biophysics - CNR, via Pietro Castellino 111, 80131, Naples,  
119 Italy.
- 120 38. MRC Integrative Epidemiology Unit, University of Bristol, Bristol, UK.
- 121 39. Estonian Genome Center, University of Tartu, Tartu, 51010, Estonia.
- 122 40. Institute of Genetics and Biomedical Research, National Research Council, Cagliari,  
123 09042 Sardinia, Italy.
- 124 41. Centre for Cardiovascular Sciences, Queen's Medical Research Institute, University of  
125 Edinburgh, Royal Infirmary of Edinburgh, Little France Crescent, Edinburgh, EH16 4TJ,  
126 Scotland.
- 127 42. Centre for Global Health Research, Usher Institute of Population Health Sciences and  
128 Informatics, University of Edinburgh, Teviot Place, Edinburgh, EH8 9AG, Scotland.
- 129 43. Cancer Program, Broad Institute, Cambridge, MA, USA.
- 130 44. Faculty of Medicine, University of Iceland, IS-101 Reykjavik, Iceland.
- 131 45. Icelandic Heart Association, Kopavogur, Iceland.
- 132 46. Department of Internal Medicine, Erasmus MC, 3015GE Rotterdam, the Netherlands.
- 133 47. Institute for Community Medicine, University Medicine Greifswald, 17475 Greifswald,  
134 Germany.
- 135 48. Fred A. Litwin Center for Cancer Genetics, Lunenfeld-Tanenbaum Research Institute of  
136 Mount Sinai Hospital, Toronto, ON, Canada.
- 137 49. Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada.
- 138 50. Geriatric Unit, Azienda Sanitaria di Firenze, Florence, Italy.
- 139 51. Department of Gynaecology and Obstetrics, University Hospital Erlangen, Friedrich-  
140 Alexander University Erlangen-Nuremberg, Erlangen, Germany.
- 141 52. Human Genetics Group, Human Cancer Genetics Program, Spanish National Cancer  
142 Research Centre (CNIO), Madrid, Spain.
- 143 53. Centro de Investigación en Red de Enfermedades Raras (CIBERER), Valencia, Spain.
- 144 54. Swiss Institute of Bioinformatics, CH-1015, Lausanne, Switzerland.
- 145 55. Department of Computational Biology, University of Lausanne, Lausanne, Switzerland.
- 146 56. Institute of Social and Preventive Medicine, University Hospital of Lausanne, Lausanne,  
147 Switzerland.
- 148 57. Human Genetics Center, School of Public Health, The University of Texas Health  
149 Science Center at Houston, Houston, TX 77030, USA.
- 150 58. Copenhagen General Population Study, Herlev Hospital, Copenhagen University  
151 Hospital, University of Copenhagen, Copenhagen, Denmark.

- 152 59. Department of Clinical Biochemistry, Herlev Hospital, Copenhagen University Hospital,  
153 University of Copenhagen, Copenhagen, Denmark.
- 154 60. Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen,  
155 Denmark.
- 156 61. Dr. Margarete Fischer-Bosch-Institute of Clinical Pharmacology, Stuttgart, Germany.  
157 62. University of Tübingen, Tübingen, Germany.
- 158 63. German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ),  
159 Heidelberg, Germany.
- 160 64. Division of Clinical Epidemiology and Aging Research, German Cancer Research  
161 Center (DKFZ), Heidelberg, Germany.
- 162 65. Division of Preventive Oncology, German Cancer Research Center (DKFZ) and National  
163 Center for Tumor Diseases (NCT), Heidelberg, Germany.
- 164 66. Institute for Prevention and Occupational Medicine of the German Social Accident  
165 Insurance, Institute of the Ruhr University Bochum (IPA), Bochum, Germany.
- 166 67. Institute for Maternal and Child Health - IRCCS "Burlo Garofolo", 34137 Trieste, Italy.  
167 68. Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda,  
168 MD, USA.
- 169 69. Department of Genetics, QIMR Berghofer Medical Research Institute, Brisbane,  
170 Australia.
- 171 70. Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN, USA.  
172 71. Division of Genetics, Children's Hospital of Philadelphia, Philadelphia, PA, USA.  
173 72. Department of Genetics, University of Pennsylvania, Philadelphia, PA, USA.  
174 73. Academic Unit of Molecular Oncology, Department of Oncology and Metabolism,  
175 University of Sheffield, Sheffield, UK.
- 176 74. Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, the  
177 Netherlands.
- 178 75. Channing Division of Network Medicine, Department of Medicine, Brigham and  
179 Women's Hospital and Harvard Medical School, Boston, MA 02115, USA.
- 180 76. Department of Pathology, Leiden University Medical Center, Leiden, The Netherlands.  
181 77. Department of Human Genetics, Leiden University Medical Center, 2300 RC Leiden,  
182 The Netherlands.
- 183 78. Non-communicable Disease Epidemiology Department, London School of Hygiene and  
184 Tropical Medicine, London, UK.
- 185 79. Centre for Cancer Genetic Epidemiology, Department of Oncology, University of  
186 Cambridge, Cambridge, CB1 8RN, UK.
- 187 80. Department of General Practice and Primary health Care, University of Helsinki,  
188 Finland.
- 189 81. David Geffen School of Medicine, Department of Medicine Division of Hematology and  
190 Oncology, University of California at Los Angeles, CA, USA.
- 191 82. Department of Epidemiology, Gillings School of Global Public Health, University of North  
192 Carolina, Chapel Hill, NC 27514.
- 193 83. Longitudinal Studies Section, Translational Gerontology Branch, National Institute on  
194 Aging, Baltimore, Maryland 21224, United States of America.
- 195 84. Institute for Medical Biometrics and Epidemiology, University Clinic Hamburg-  
196 Eppendorf, Hamburg, Germany.
- 197 85. Department of Cancer Epidemiology/Clinical Cancer Registry, University Clinic  
198 Hamburg-Eppendorf, Hamburg, Germany.
- 199 86. Department of Genetics, University of Groningen, University Medical Centre Groningen,  
200 Groningen, The Netherlands.
- 201 87. Cancer Epidemiology Centre, Cancer Council Victoria, Melbourne, Australia.  
202 88. Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global  
203 Health, The University of Melbourne, Melbourne, Australia.
- 204 89. German Center for Diabetes Research, 85764 Neuherberg, Germany.  
205 90. Cancer & Environment Group, Center for Research in Epidemiology and Population  
206 Health (CESP), INSERM, University Paris-Sud, University Paris-Saclay, Villejuif, France.

- 207 91. Division of Epidemiology, Department of Health Sciences Research, Mayo Clinic,  
208 Rochester, Minnesota, USA.
- 209 92. Molecular Genetics of Breast Cancer, Deutsches Krebsforschungszentrum (DKFZ),  
210 Heidelberg, Germany.
- 211 93. Laboratory of Epidemiology and Population Sciences, National Institute on Aging,  
212 Intramural Research Program, National Institutes of Health, Bethesda, Maryland, 20892,  
213 USA.
- 214 94. Department of Psychiatry, University of Groningen, University Medical Center  
215 Groningen, Groningen, The Netherlands.
- 216 95. Department of Medical Oncology, Family Cancer Clinic, Erasmus MC Cancer Institute,  
217 Rotterdam, The Netherlands.
- 218 96. Department of Nutrition, Harvard School of Public Health, Boston, MA 02115, USA.
- 219 97. Department of Epidemiology, Erasmus MC, Rotterdam, the Netherlands.
- 220 98. Section of Genetic Medicine, Department of Medicine, University of Chicago, Chicago,  
221 IL, USA.
- 222 99. Department of Epidemiology and Biostatistics, MRC Health Protection Agency (HPA)  
223 Centre for Environment and Health, School of Public Health, Imperial College London, UK.
- 224 100. Biocenter Oulu, P.O.Box 5000, Aapistie 5A, FI-90014 University of Oulu, Finland.
- 225 101. Department of Children and Young People and Families, National Institute for Health  
226 and Welfare, Aapistie 1, Box 310, FI-90101 Oulu, Finland.
- 227 102. Institute of Health Sciences, P.O.Box 5000, FI-90014 University of Oulu, Finland.
- 228 103. Unit of Primary Care, Oulu University Hospital, Kajaanintie 50, P.O.Box 20, FI-90220  
229 Oulu, 90029 OYS, Finland.
- 230 104. Hebrew SeniorLife Institute for Aging Research, Boston, MA, 02131, USA.
- 231 105. Laboratory for Translational Genetics, Department of Oncology, University of Leuven,  
232 Leuven, Belgium.
- 233 106. Vesalius Research Center (VRC), VIB, Leuven, Belgium.
- 234 107. Division of Reproductive Medicine, Department of Obstetrics and Gynaecology,  
235 Erasmus MC, Rotterdam, The Netherlands.
- 236 108. Department of Epidemiology, School of Public Health, University of Washington,  
237 Seattle, WA 98195, USA.
- 238 109. Center for Human Genetics, Division of Public Health Sciences, Wake Forest School of  
239 Medicine.
- 240 110. Translational Cancer Research Area, University of Eastern Finland, Kuopio, Finland.
- 241 111. Institute of Clinical Medicine, Pathology and Forensic Medicine, University of Eastern  
242 Finland, Kuopio, Finland.
- 243 112. Imaging Center, Department of Clinical Pathology, Kuopio University Hospital, Kuopio,  
244 Finland.
- 245 113. NIHR Oxford Biomedical Research Centre, Churchill Hospital, OX3 7LE Oxford, UK.
- 246 114. Oxford Centre for Diabetes, Endocrinology, & Metabolism, University of Oxford,  
247 Churchill Hospital, OX3 7LJ Oxford, UK.
- 248 115. Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK.
- 249 116. Central Hospital of Augsburg, MONICA/KORA Myocardial Infarction Registry,  
250 Augsburg, Germany.
- 251 117. Institute of Human Genetics, Helmholtz Zentrum München, German Research Center  
252 for Environmental Health, Neuherberg, Germany.
- 253 118. Department of Electron Microscopy/Molecular Pathology, The Cyprus Institute of  
254 Neurology and Genetics, Nicosia, Cyprus.
- 255 119. Institute for Molecular Bioscience, The University of Queensland, Brisbane, Australia.
- 256 120. Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto,  
257 ON, Canada.
- 258 121. Laboratory Medicine Program, University Health Network, Toronto, ON, Canada.
- 259 122. Laboratory of Neurogenetics, National Institute on Aging, Bethesda, MD, USA.
- 260 123. Department of Obstetrics and Gynecology, Helsinki University Hospital, University of  
261 Helsinki, Helsinki, Finland.

262 124. Institute of Health and Biomedical Innovation, Queensland University of Technology,  
263 Australia.  
264 125. Interdisciplinary Center Psychopathology and Emotion Regulation, University of  
265 Groningen, University Medical Center Groningen, Groningen, The Netherlands.  
266 126. British Heart Foundation Glasgow Cardiovascular Research Centre, Institute of  
267 Cardiovascular and Medical Sciences, College of Medical, Veterinary and Life Sciences,  
268 University of Glasgow, Glasgow G12 8TA, UK.  
269 127. Psychiatric & Neurodevelopmental Genetics Unit, Department of Psychiatry,  
270 Massachusetts General Hospital, Boston, MA, USA.  
271 128. Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard,  
272 Cambridge, Massachusetts 02142, USA.  
273 129. Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, UK.  
274 130. Analytic and Translational Genetics Unit, Massachusetts General Hospital and Harvard  
275 Medical School, Boston, Massachusetts, USA.  
276 131. Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Finland.  
277 132. National Institute for Health and Welfare, Finland.  
278 133. Unit of Molecular Bases of Genetic Risk and Genetic Testing, Department of  
279 Preventive and Predictive Medicine, Fondazione IRCCS Istituto Nazionale dei Tumori (INT),  
280 Milan, Italy.  
281 134. Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden.  
282 135. Division of Molecular Pathology, The Netherlands Cancer Institute - Antoni van  
283 Leeuwenhoek Hospital, Amsterdam, The Netherlands.  
284 136. Division of Psychosocial Research and Epidemiology, The Netherlands Cancer  
285 Institute - Antoni van Leeuwenhoek hospital, Amsterdam, The Netherlands.  
286 137. Department of Pathology, The University of Melbourne, Melbourne, Australia.  
287 138. Department of Obstetrics and Gynaecology, University of Cambridge, Cambridge,  
288 United Kingdom.  
289 139. Institute of Medical Informatics, Biometry and Epidemiology, Chair of Genetic  
290 Epidemiology, Ludwig-Maximilians-Universität, 81377 Munich, Germany.  
291 140. Department of Public Health, University of Helsinki, Helsinki, Finland.  
292 141. Vanderbilt Epidemiology Center, Institute for Medicine and Public Health, Vanderbilt  
293 University, Nashville, TN, USA.  
294 142. Department of Obstetrics and Gynecology, Vanderbilt University School of Medicine,  
295 Nashville, TN, USA.  
296 143. Interfaculty Institute for Genetics and Functional Genomics, University Medicine  
297 Greifswald, 17475 Greifswald, Germany.  
298 144. University Hospital of Lausanne, Lausanne, Switzerland.  
299 145. Department of Internal Medicine, Division of Endocrinology, Leiden University Medical  
300 Center, Leiden, the Netherlands.  
301 146. Einthoven Laboratory for Experimental Vascular Medicine, Leiden University Medical  
302 Center, Leiden, the Netherlands.  
303 147. Laboratory of Cancer Genetics and Tumor Biology, Cancer and Translational Medicine  
304 Research Unit, Biocenter Oulu, University of Oulu, Oulu, Finland.  
305 148. Laboratory of Cancer Genetics and Tumor Biology, Northern Finland Laboratory  
306 Centre NordLab, Oulu, Finland.  
307 149. Department of Endocrinology, University of Groningen, University Medical Centre  
308 Groningen, Groningen, The Netherlands.  
309 150. Department of Obstetrics and Gynecology, University Medicine Greifswald, 17475  
310 Greifswald, Germany.  
311 151. University of Sassari, Department of Biomedical Sciences, Sassari, 07100 Sassari,  
312 Italy.  
313 152. Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, USA.  
314 153. Department of Public Health and Primary Care, Leiden University Medical Center,  
315 Leiden, the Netherlands.

- 316 154. Research Unit for Gynaecology and Obstetrics, Department of Clinical Research,  
317 University of Southern Denmark, Denmark.
- 318 155. Department of Obstetrics and Gynaecology, Campus Grosshadern, Ludwig-  
319 Maximilians-University, Munich, Germany.
- 320 156. Full consortium membership is displayed in the supplementary material.
- 321 157. Division of Biostatistics, Institute for Human Genetics, and Institute for Computational  
322 Health Sciences, University of California, San Francisco, California, 94158, USA.
- 323 158. 23andMe Inc., 899 W. Evelyn Avenue, Mountain View, California 94041, USA.
- 324 159. Division of Cancer Epidemiology, German Cancer Research Center (DKFZ),  
325 Heidelberg, Germany.
- 326 160. University Cancer Center Hamburg (UCCH), University Medical Center Hamburg-  
327 Eppendorf, Hamburg, Germany.
- 328 161. Boston University School of Medicine, Department of Medicine, Section of General  
329 Internal Medicine, Boston, MA 02118, USA.
- 330 162. Department of Paediatrics, University of Cambridge, Cambridge, CB2 0QQ, UK.
- 331 Correspondence to John R.B. Perry ([john.perry@mrc-epid.cam.ac.uk](mailto:john.perry@mrc-epid.cam.ac.uk)) and Ken K. Ong  
332 ([ken.ong@mrc-epid.cam.ac.uk](mailto:ken.ong@mrc-epid.cam.ac.uk)).

333

334

335

336

337

## 338 Abstract

339 The timing of puberty is a highly polygenic childhood trait that is epidemiologically associated  
340 with various adult diseases. Using 1000-Genome imputed genotype data in up to ~370,000  
341 women, we identify 389 independent signals ( $P < 5 \times 10^{-8}$ ) for age at menarche, a notable  
342 milestone in female pubertal development. In Icelandic data from deCODE, these signals  
343 explain ~7.4% of the population variance in age at menarche, corresponding to ~25% of the  
344 estimated heritability. We implicate ~250 genes via coding variation or associated  
345 expression, demonstrating significant enrichment in neural tissues. Rare variants near  
346 imprinted genes MKRN3 and DLK1 were identified, exhibiting large effects only when  
347 paternally inherited. Mendelian randomization analyses indicate causal inverse associations,  
348 independent of BMI, between puberty timing and risks for breast and endometrial cancers in  
349 women, and prostate cancer in men. In aggregate, our findings reveal new complexity in the  
350 genetic regulation of puberty timing and support causal links with cancer susceptibility.

## 351 Introduction

352 Puberty is the developmental stage of transition from childhood to physical and sexual  
353 maturity and its timing varies markedly between individuals<sup>1</sup>. This variation reflects the  
354 influence of genetic, nutritional and other environmental factors and is associated with the  
355 subsequent risks for several diseases in adult life<sup>2</sup>. Our previous large-scale genomic  
356 studies identified 113 independent regions associated with age at menarche (AAM), a well-  
357 recalled milestone of puberty in females<sup>3,4</sup>. The vast majority of those signals have  
358 concordant effects on the age at voice breaking (genome-wide genetic correlation between  
359 traits  $r_g = 0.74$ ), a corresponding milestone in males<sup>5</sup>. Those genetic findings implicated a  
360 diverse range of mechanisms involved in the regulation of puberty timing, identified  
361 significant enrichment of AAM-associated variants in/near genes disrupted in rare disorders  
362 of puberty, and highlighted shared aetiological factors between puberty timing and metabolic  
363 disease outcomes<sup>2,3</sup>.

364 However, those previous studies were based on genome-wide association data that were  
365 imputed to the relatively sparse HapMap2 reference panel or they used gene-centric arrays.  
366 Consequently, the reported genetic signals explained only a small fraction of the population  
367 variance, suggesting that several hundreds or thousands of signals are involved<sup>3,4</sup>. Here, we  
368 report an enlarged genomic analysis for AAM in a nearly 2-fold higher sample of women  
369 than previously<sup>3</sup>, and using more densely imputed genomic data. Our findings increase by  
370 more than 3-fold the number of independently associated signals and indicate likely causal  
371 effects of puberty timing on risks of various sex steroid sensitive cancers in men and women.

## 372 Results

373 Genome-wide array data, imputed to the 1000-Genome reference panel, were available in  
374 up to 329,345 women of European ancestry. These comprised 40 studies from the  
375 ReproGen consortium (N=179,117), in addition to the 23andMe, Inc. (N=76,831) and UK  
376 Biobank studies (N=73,397) (**Table S1**). The distribution of genome-wide test statistics  
377 demonstrated significant inflation ( $\lambda_{GC} = 1.75$ ), however LD score regression  
378 analyses confirmed that this inflation was solely due to polygenicity rather than population  
379 structure (LD score intercept = 1.00, s.e 0.02). In total, 37,925 variants were associated with  
380 AAM at  $P < 5 \times 10^{-8}$ , which were resolved to 389 statistically-independent signals (**Figure S1**,



381 **Table S2).** Per-allele effect sizes ranged from ~1 week to 5 months, 16 index variants were  
382 classed as low-frequency (minor allele frequency <5%; minimum observed 0.5%), and 26  
383 were insertion/deletion polymorphisms. Signals were distributed evenly across all 23  
384 chromosomes with respect to chromosome size (**Figure S2**). Of the previously reported 106  
385 autosomal, 5 exome-array and 2 X-chromosome signals for AAM, all remained associated at  
386 genome-wide significance, except for two common loci (reported as *SCRIB/PARP10*  
387 [ $P=5 \times 10^{-4}$ ] and *FUT8* [ $P=5.4 \times 10^{-7}$ ]) and one rare variant not captured by the 1000G  
388 reference panel (p.W275X, *TACR3*).

389 Independent replication in the deCODE study (N=39,543 women) showed that 367 (94.3%)  
390 of the 389 signals had directionally-concordant effects (187 at  $P < 0.05$ ) and 368 retained  
391 genome-wide significance in a combined meta-analysis (**Table S3**). In aggregate, the top  
392 389 index SNPs explained 7.4% of the trait variance in deCODE and 7.2% in UK Biobank  
393 (the latter estimate used weights derived from a meta-analysis excluding UK Biobank).  
394 These estimates are double that explained by the previously reported 106 signals<sup>3</sup> (3.7% in  
395 deCODE) and are equivalent to one quarter of the total chip-captured heritability  
396 ( $h^2_{\text{SNP}}=32\%$ ,  $se=1\%$ ) for AAM, estimated in UK Biobank.

397 Consistent with our previous reports, we found a strongly shared genetic architecture  
398 between AAM in women and age at voice breaking in men (considered as a continuous trait  
399 in 55,871 men in 23andMe, Inc.) (genetic correlation ( $r_g$ )=0.75  $P=1.2 \times 10^{-79}$ ). Of the 389 AAM  
400 signals, 327 demonstrated directionally-consistent trends or associations with age at voice  
401 breaking in men (binomial  $P=1.4 \times 10^{-44}$ ), and 18 signals reached a conservative multiple test-  
402 corrected significance threshold ( $P < 1 \times 10^{-4}$ ; i.e. 0.05 / 389) (**Table S4**). Similarly, in UK  
403 Biobank where age at voice breaking was recorded using only 3 categories, 277 and 297 of  
404 the 377 autosomal loci demonstrated directionally-consistent trends or associations with  
405 “relatively early voice breaking” (N=2,678 cases, N=55,763 controls, binomial  $P=2.4 \times 10^{-20}$ )  
406 and “relatively late voice breaking” (N=3,566 cases,  $P=1.9 \times 10^{-30}$ ), respectively (**Table S5**).

#### 407 **Implicated genes and tissues**

408 We used a number of analytical techniques to implicate genes in the regulation of AAM.  
409 These included: mapping of non-synonymous SNPs, gene expression QTLs and integration  
410 of Hi-C chromatin interaction data. Eight of the 389 lead variants were non-synonymous, and  
411 a further 24 genes were implicated by highly correlated non-synonymous variants ( $r^2 > 0.8$ )  
412 (**Table S6**). These include genes disrupted in rare disorders of puberty: aromatase  
413 (*CYP19A1*, #307), gonadotropin-releasing hormone (*GNRH1*, #178), kisspeptin (*KISS1*,  
414 signal #31); and the stop-gained variant in fucosyltransferase 2 (*FUT2*, #357) that confers  
415 blood group secretor status.

416 Two approaches were used to interrogate publicly available gene expression datasets, both  
417 of which use one or more SNPs (not restricted to lead SNPs) to infer patterns of gene  
418 expression based on imputation reference panels (see **methods**). Firstly, to maximise power  
419 we analysed data from the largest available eQTL dataset for any tissue (whole blood,  
420 N=5,311)<sup>6</sup>, under the assumption that some causal genes and regulatory mechanisms might  
421 be ubiquitously expressed or functionally involved in blood tissues. Systematic eQTL  
422 integration using the Summary Mendelian Randomization approach<sup>7</sup> prioritised 113  
423 transcripts, for 60 of which there was evidence for causal or pleiotropic effects, rather than  
424 coincidental overlap of signal (as indicated by HEIDI heterogeneity test  $P > 0.009$ ) (**Table S7**).

425 Secondly, we used LD score regression applied to specifically expressed genes (LDSC-  
426 SEG)<sup>8</sup> to identify AAM-relevant tissues and cell types that are enriched for AAM heritability.  
427 Five of the 46 GTEx tissues were positively enriched for AAM-associated variants (**Figure**  
428 **1**). Notably, all of these were central nervous system tissues, including the pituitary and,  
429 additionally, the hypothalamus was just below the significance threshold for enrichment  
430 ( $P=9.8\times 10^{-3}$ ), consistent with the key role of this central axis<sup>2</sup>. Targeted assessment of these  
431 six enriched brain tissues using MetaXcan identified 205 genes whose expression was  
432 regulated by AAM-associated variants (**Table S8**). Of note, later AAM was associated with  
433 higher transcript levels of *LIN28B* (#147) in the pituitary, *NCOA6* (Nuclear receptor  
434 coactivator 6; #365) in the cerebellum, and *HSD17B12* (encoding Hydroxysteroid (17-Beta)  
435 Dehydrogenase 12; #250) in various tissues.

436 To identify possible distal causal genes, we interrogated reported Hi-C data to assess if any  
437 of the AAM loci are located in regions of chromatin looping<sup>9</sup>. 335 of the 389 loci were located  
438 within a topologically associating domain (TAD) – a defined boundary region containing  
439 chromatin contact points, each of which contained on average ~5 genes (**Table S9**). These  
440 included 22 of the 31 gene desert regions (nearest protein-coding gene >300kb), where  
441 TADs contained notable distal candidate genes such as *INHBA* (#158), *BDNF* (#248),  
442 *JARID2* (#128) and several gamma-aminobutyric acid receptors (#91). We also observed  
443 several regions where multiple independent AAM signals all reside within one TAD  
444 containing the same single gene – *RORB* (signal #200 intronic, signal #199 ~200kb  
445 downstream, #198 ~1.2Mb downstream), *THRB* (#67 intronic, #68 ~180kb upstream) and  
446 *TACR3* (#96 5'UTR, #97 ~25kb upstream, #98 ~133kb upstream and #95 ~263Kb  
447 downstream).

448 66 AAM signals were located in a specific contact point (between 5-25kb in size) within the  
449 335 TADs, indicating a direct physical connection between these signals and a distal  
450 genomic region, on average ~320kb away. This included the previously reported example of  
451 the BMI-associated (and AAM-associated) *FTO* SNP and a distal *IRX3* promoter ~1Mb away  
452 (signal #326)<sup>10</sup>. The longest chromatin interaction observed was ~38.6Mb, where two distinct  
453 AAM signals located ~300kb apart (#206 and #207) were both in contact with the same  
454 distal genomic region ~38.6Mb away that contains only one gene: prostaglandin E synthase  
455 2 (*PTGES2*).

## 456 **Transcription factor binding enrichment**

457 To identify functional gene networks implicated in the regulation of AAM, we tested for  
458 enriched co-occurrence of AAM associations and predicted regulators within 226 enhancer  
459 modules combining DNaseI hypersensitive sites and chromatin states in 111 cell types and  
460 tissues. In total, we tested 2,382 transcription factor-enhancer module combinations. Sixteen  
461 transcription factor motifs were enriched for co-occurrence with AAM-associated variants  
462 within enhancer regions at study level significance ( $FDR<0.05$ ) (**Table S10**). Furthermore, 5  
463 of the 16 motif-associated transcription factors also mapped within 1Mb of an index AAM-  
464 associated SNP. These transcription factors included notable candidates; firstly, pituitary  
465 homeobox 1 (*PITX1*), is located within 50kb of genome-wide significant SNPs (~500kb from  
466 lead index #114). Secondly, *SMAD3*, a gene recently implicated in susceptibility to dizygous  
467 twinning<sup>11</sup>, is located within 600kb of an index SNP and its expression in several GTEx brain  
468 tissues is genetically correlated with AAM. Thirdly, *RXRβ* is located within ~500kb of a novel  
469 index SNP (signal #133), and it represents the fifth (out of nine) retinoid-related receptor

470 gene implicated by genome-wide significant AAM variants. This set now includes all three  
471 retinoid X receptor genes (*RXRA*, *RXRB* and *RXRG*), and retinoid-related receptor genes  
472 are the nearest gene to the index SNP at three AAM loci (*RXRA*, *RORA* and *RORB*).

### 473 **Pathway analyses**

474 To identify other mechanisms that regulate pubertal timing, we tested all SNPs genome-wide  
475 for enrichment of AAM associations with pre-defined biological pathway genes. Ten  
476 pathways reached study-wide significance (FDR<0.05). Five pathways were related to  
477 transcription factor binding, and the other pathways were: peptide hormone binding, PI3-  
478 kinase binding, angiotensin stimulated signalling, neuron development and gamma-  
479 aminobutyric acid (GABA) type B receptor signalling (**Table S11**).

480 All of our previously reported custom pathways (**Table S12**)<sup>3</sup> remained significant in this  
481 expanded dataset: nuclear hormone receptors ( $P=2.4\times 10^{-3}$ ); Mendelian pubertal disorder  
482 genes ( $P=1.9\times 10^{-3}$ ); and JmjC-domain-containing lysine-specific demethylases ( $P=1\times 10^{-4}$ ).  
483 Notably, new genome-wide significant signals mapped to lysine-specific demethylase genes:  
484 *JMJD1C* (signal #223), *PHF2* (#208), *KDM4B* (#347), *KDM6B* (#332), *JARID2* (#128), or to  
485 Mendelian pubertal disorder genes: *CYP19A1* (#307), *FGF8* (#230), *GNRH1* (#178) *KAL1*  
486 (#378), *KISS1* (#31), *NR5A1* (#215), and *NR0B1* (#379). The strongest AAM signal remains  
487 at *LIN28B*<sup>3,12,13</sup>, which encodes a key repressor of *let-7* miRNA biogenesis and cell  
488 pluripotency<sup>14</sup>. Transgenic *Lin28a/b* mice demonstrate both altered pubertal growth and  
489 glycaemic control<sup>15</sup>, suggesting that the *Lin28/let-7* axis could link puberty timing to type 2  
490 diabetes susceptibility in humans. *let-7* miRNA targets are reportedly enriched for variants  
491 associated with type 2 diabetes<sup>16</sup>. We tested the same set of computationally-predicted and  
492 experimentally-derived mRNA/protein *let-7* miRNA targets<sup>16</sup>, and observed significant  
493 enrichment of AAM-associated variants at miRNA targets that are down-regulated by *let-7b*  
494 overexpression in primary human fibroblasts (**Table S12**,  $P_{\min}=1\times 10^{-3}$ ).

### 495 **Imprinted genes and parent-of-origin effects**

496 We previously reported an excess of parent-of-origin specific associations for those AAM  
497 variants that map near imprinted genes, as defined primarily from animal studies<sup>3</sup>. Recent  
498 data from the GTEx consortium now allow a more systematic assessment of imprinted gene  
499 enrichment using genes defined from human transcriptome-wide analyses<sup>17</sup>. Consistent with  
500 our previous observations, imprinted genes were enriched for AAM-associated variants  
501 (MAGENTA  $P=4\times 10^{-3}$ ), with a concordant excess of parent-of-origin specific associations for  
502 the 389 index AAM variants (**Figure S3**, **Table S3**).

503 Systematic assessment of the 389 AAM gene regions in the Icelandic deCODE study  
504 revealed novel rare variants in two imprinted gene regions with robust parent-of-origin  
505 specific associations with AAM. Firstly, we identified a rare 5' UTR variant rs530324840  
506 (MAF=0.80% in Iceland) in *MKRN3* that is associated with AAM under the paternal  
507 ( $P=6.4\times 10^{-11}$ ,  $\beta= -0.52$  years) but not the maternal model ( $P=0.20$ ,  $\beta=0.098$ ,  $P_{\text{het}}=1.3\times 10^{-7}$ )  
508 (**Table 1 & S13**). rs530324840 is by far the most significant variant at the *MKRN3* locus and  
509 is uncorrelated with our previously reported common variant rs12148769 at the same locus  
510 ( $r^2 < 0.001$  in deCODE)<sup>3</sup> (**Figure S4**). We note that the rare 5' UTR variant rs184950120  
511 detected in the current GWAS meta-analysis also shows paternal-specific association in

512 deCODE and, despite their near location (235bp from rs530324840), is uncorrelated to  
513 rs530324840 ( $r^2 < 0.0001$  in deCODE).

514 The second novel robust parent-of-origin specific signal is indicated by a rare intergenic  
515 variant at the *DLK1* locus (rs138827001; MAF=0.36% in Iceland) that associates with AAM  
516 under the paternal model ( $P=4.7 \times 10^{-10}$ ,  $\beta = -0.70$  years) but not the maternal model ( $P=0.88$ ,  
517  $\beta = -0.018$  years,  $P_{\text{het}}=1.4 \times 10^{-4}$ ) (**Table 1, Figure S5**). rs138827001 is uncorrelated with the  
518 two previously reported common variants rs10144321 and rs7141210 at the *DLK1* locus ( $r^2$   
519  $< 0.01$  in Iceland) that both also showed paternal allele-specific associations<sup>3</sup>. At this locus,  
520 we observed a further common variant rs61992671 (MAF=48.5% in Iceland) 4.4kb upstream  
521 of the Maternally Expressed 9 (*MEG9*) gene (~300kb from *DLK1*) that was associated with  
522 AAM under the maternal model ( $P=6.0 \times 10^{-8}$ ,  $\beta = -0.077$  years) but not the paternal model  
523 ( $P=0.27$ ,  $\beta = 0.015$  years,  $P_{\text{het}}=1.9 \times 10^{-5}$ ). rs61992671 was uncorrelated ( $r^2 < 0.05$ ) with the two  
524 common signals identified in the meta-analysis (rs10144321 and rs7141210) and replicated  
525 with a consistent magnitude of effect in the our GWAS meta-analysis (additive model,  
526  $P=5.1 \times 10^{-6}$ ).

## 527 **Disproportionate genetic effects on early or late puberty timing**

528 Family-based studies in twins have suggested age-related differences in the impacts of  
529 genetic and environmental factors on AAM<sup>18</sup>. To test for asymmetry in the genetic effects on  
530 puberty timing, we defined two groups of women in the UK Biobank study based on  
531 approximated quintiles for AAM – “early” (8-11 years inclusive, N=14,922) and “late” (15-19,  
532 N=12,290). Each group was compared to the same median quintile AAM reference group  
533 (age 13, N=17,717). Estimated genome-wide heritability was higher for early AAM  
534 ( $h^2_{\text{SNP}}=28.8\%$ ; s.e 2.3%) than late AAM ( $h^2_{\text{SNP}}=21.5\%$ ; s.e. 2.5%,  $P_{\text{diff}}=0.03$ ). Accordingly,  
535 217/377 (57.7%) autosomal index SNPs had larger effect estimates on early than late AAM  
536 (binomial  $P=0.004$  vs. 50% expected), and the aggregated effect of the 377 SNPs also  
537 differed between strata ( $P=2.3 \times 10^{-4}$ ) (**Figure 2, Table S14**). These differences remained  
538 when matching the early and late AAM strata for sample size and phenotype ranges (**Table**  
539 **S15**).

540 In contrast, we observed the opposite pattern of disproportion in the genetic effects on male  
541 voice breaking in UK Biobank (“relatively early” N=2678, “relatively late” N=3566). Genome-  
542 wide heritability estimates tended to be higher for relatively late voice breaking (7.8%, s.e  
543 1.2%) than for relatively early (6.9%, s.e 1.3%), and 227/377 (60.2%) index SNPs had larger  
544 effect estimates on relatively late than relatively early voice breaking (binomial  $P=4.3 \times 10^{-5}$ ).

## 545 **BMI-independent effects of puberty timing on cancer risks**

546 Traditional (non-genetic) epidemiological studies have reported complex associations  
547 between puberty timing, body mass index (BMI) and adult cancer risks. For example, large  
548 studies using historical growth records identified lower adolescent BMI and earlier puberty  
549 timing (estimated by the age at peak adolescent growth) as predictors of higher breast  
550 cancer risk in women<sup>19,20</sup>. Conversely, BMI is positively associated with breast cancer risk in  
551 postmenopausal women<sup>21</sup>. Furthermore, the strong inter-relationship between puberty timing  
552 and BMI limits the ability to consider their distinct influences on disease risks in traditional  
553 observational studies. Consistent with our previous report<sup>5</sup>, we observed a strong inverse  
554 genetic correlation between AAM and BMI ( $rg = -0.35$ ,  $P=1.6 \times 10^{-72}$ ). 39 AAM loci overlapped

555 with reported loci for adult BMI<sup>22</sup>, yet even those AAM signals with weak individual  
556 associations with adult BMI still contributed to BMI when considered in aggregate: the 237  
557 AAM variants without a nominal individual association with adult BMI (all  $P > 0.05$ ) were  
558 collectively associated with adult BMI ( $P = 4.2 \times 10^{-9}$ ) (**Figure S6**). This finding precludes an  
559 absolute distinction between BMI-related and BMI-unrelated AAM variants.

560 In Mendelian randomisation analyses, we therefore included adjustment for genetically-  
561 predicted BMI (as predicted by the 375 autosomal AAM variants) in order to assess the likely  
562 direct (i.e. BMI-independent) effects of AAM on the risks for various sex steroid-sensitive  
563 cancers (see **methods**). In these BMI-adjusted models, increasing AAM was associated with  
564 lower risk for breast cancer (OR=0.935 per year, 95% confidence interval: 0.894-0.977;  
565  $P = 2.6 \times 10^{-3}$ ), and in particular with oestrogen receptor (ER)-positive but not ER-negative  
566 breast cancer ( $P$ -heterogeneity = 0.02) (**Figure 3, Table S16**). Similarly, increasing AAM  
567 adjusted for genetically-predicted BMI was associated with lower risks for: ovarian cancer  
568 (OR=0.930, 0.880-0.982;  $P = 9.3 \times 10^{-3}$ ), in particular serous ovarian cancer (OR=0.917,  
569 0.859-0.978;  $P = 8.9 \times 10^{-3}$ ); and endometrial cancer (OR=0.781, 0.699-0.872;  $P = 9.97 \times 10^{-6}$ ).  
570 Assuming an equivalent per-year effect of the current AAM variants on age at voice  
571 breaking, as we reported for the 106 previously identified AAM variants<sup>5</sup>, we could also infer  
572 a protective effect of later puberty timing, independent of BMI, on lower risk for prostate  
573 cancer in men (OR=0.925, 0.876-0.976;  $P = 4.4 \times 10^{-3}$ ).

574 These findings were supported by sensitivity tests using sub-groups of AAM signals stratified  
575 by their individual associations with adult BMI. The 'BMI-unrelated' variant score (comprising  
576 314 variants) supported a direct effect of AAM timing on breast cancer risk in women  
577 (OR=0.946, 0.904-0.988;  $P = 1.3 \times 10^{-2}$ ). In contrast, a score using only the 61 BMI-related  
578 AAM variants gave a significant result in the opposite direction (OR=1.15, 1.06-1.25;  
579  $P = 4.3 \times 10^{-4}$ ) (**Table S16**), consistent with the recently reported inverse association between  
580 genetically-predicted BMI and breast cancer risk<sup>23,24</sup>. Further sensitivity tests (heterogeneity  
581 and MR-Egger tests) using the 'BMI-unrelated' variant score suggested that additional sub-  
582 pathways might link AAM to risk of ovarian cancer (MR-Egger Intercept  $P = 0.036$ ), but  
583 reassuringly these tests indicated no further pleiotropy (i.e. beyond the effects of BMI) in our  
584 analyses of breast, endometrial and prostate cancers (for all: I-square < 23% and MR-Egger  
585 Intercept  $P > 0.1$ ) (**Table S16, Figure S7**).

## 586 Discussion

587 In a substantially enlarged genomic analysis using densely imputed genomic data, we have  
588 identified 389 independent, genome-wide significant signals for AAM. In aggregate, these  
589 signals explain ~7.4% of the population variance in AAM, corresponding to ~25% of the  
590 estimated heritability. While assigning possible causal genes to associated loci is an ongoing  
591 challenge for GWAS findings, we adopted a number of recently described methods to  
592 implicate the underlying genes and tissues. 33 genes were implicated by non-synonymous  
593 variants and >200 genes were implicated by transcriptome-wide association in the five  
594 neural tissues enriched for AAM-associated gene activation. Transcriptome-wide association  
595 analyses also enabled the estimation of direction of gene expression in relation to AAM,  
596 notably indicating the likely delaying effect of *LIN28B* gene expression on AAM, which is  
597 consistent with inhibitory effects of this gene on developmental timing in animal and cell  
598 models<sup>14,15</sup>.

599 Our findings add to the growing evidence for a significant role of imprinted genes in the  
600 regulation of puberty timing<sup>3</sup>. In a recent family study, rare coding mutations (two frameshift,  
601 one stop-gained and one missense) in *MKRN3* were shown to cause central precocious  
602 puberty when paternally inherited<sup>25</sup>. Taken together, three distinct types of variants at  
603 *MKRN3* appear to influence puberty timing when paternally inherited: (i) multiple rare loss-of-  
604 function mutations with large effects<sup>25</sup> (ii) a common intergenic variant (rs530324840) with  
605 small effect, and (iii) two 5' UTR variants (rs184950120 and rs12148769) with intermediate  
606 allele frequencies (1 in 95 Icelandic women) and effects (~0.5 years per allele). Similarly, we  
607 found allelic heterogeneity at the imprinted *DLK1* locus where, as at *MKRN3*, a low  
608 frequency paternally-inherited allele conferred a substantial decrease in the age of puberty  
609 timing. At the same locus, maternal allele-specific association with an unrelated variant near  
610 to the maternally-expressed gene *MEG9* is consistent with multiple imprinting control centres  
611 at this imprinted gene cluster<sup>26</sup>.

612 The strong collective influence of the identified loci on AAM allowed informative stratification  
613 of AAM-associated variants in causal analyses to distinguish between BMI-related and BMI-  
614 unrelated pathways linking puberty timing to risk of sex steroid sensitive cancers. These  
615 findings were supported in BMI-adjusted models and, except for ovarian cancer, by  
616 additional tests for pleiotropy, and indicate causal influences of both lower adolescent BMI  
617 and earlier AAM on later cancer risks. The association between BMI and breast cancer risk  
618 is complex; directionally-opposing associations have been reported with adolescent and  
619 adult BMI, and with differing associations with pre- and post-menopausal breast  
620 cancer<sup>19,20,21</sup>. Recent Mendelian randomisation studies report a consistent protective effect  
621 of higher BMI on pre- and post-menopausal breast cancer<sup>23,24</sup>. Some studies have reported  
622 on the association between later puberty timing and lower risk of prostate cancer in men, but  
623 such data on puberty timing in men is scarcely recorded<sup>27</sup>. The influences of earlier puberty  
624 timing, independent of BMI, on higher risks of breast, ovarian and endometrial cancers in  
625 women, and prostate cancer in men, could be mediated by a longer duration of exposure to  
626 sex steroids. Alternatively, mechanisms that confer earlier puberty timing might also promote  
627 higher levels of hypothalamic-pituitary-gonadal axis activity, as exemplified by a variant in  
628 *FSHB* that confers earlier AAM, higher circulating follicle stimulating hormone concentrations  
629 in women, and higher susceptibility to dizygous twinning<sup>11</sup>.

630 We identified disproportionate effects of AAM variants on early or late puberty timing in a  
631 sex-discordant pattern. In females, variant effect estimates and heritability were higher for  
632 early versus late puberty timing, but the opposite was seen in males. These findings are  
633 concordant with clinical observations of sex-dependent penetrance of abnormal early and  
634 late puberty timing, even when accounting for presentation bias. Girls are more susceptible  
635 than boys to start puberty at abnormally young ages<sup>28</sup>, whereas boys are more susceptible  
636 than girls to have delayed onset of puberty<sup>29</sup>. These findings suggest some, yet to be  
637 unidentified, sex-specific gene-environment interactions. Future studies should  
638 systematically explore the potential influence of AAM-associated variants on rare disorders  
639 of puberty. In summary, our findings suggest unprecedented genetic complexity in the  
640 regulation of puberty timing and support new causal links with susceptibility to sex steroid-  
641 sensitive cancers in women and men.

## 642 Online Methods

### 643 GWAS meta-analysis for age at menarche in women

644 Each individual study tested SNPs using a two tailed additive linear regression model for  
645 association with age at menarche (AAM), including age at study visit and other study specific  
646 covariates. Insertion/deletion polymorphisms were coded as “I” and “D” for data storage  
647 efficiency and to allow harmonisation across all studies. Genetic variants and individuals  
648 were filtered on the basis of study specific quality control metrics. Association statistics for  
649 each SNP were then uploaded by study analysts for central processing. Study level results  
650 files were assessed following standardised quality control pipeline<sup>30</sup>, and results for each  
651 SNP were meta-analysed across studies using an inverse variance weighted model using  
652 METAL<sup>31</sup> in a two stage process. Firstly, results from ReproGen consortium studies (**Table**  
653 **S1**) were combined and then filtered so that only those SNPs which appeared in over half of  
654 these studies were taken forward. Secondly, aggregated ReproGen consortium results were  
655 combined with data from the UK Biobank<sup>32,33</sup> and 23andMe, Inc. studies<sup>5</sup>. Variants were only  
656 included in the final results file if they had results from at least two of these three sources,  
657 and a combined minor allele frequency (MAF) > 0.1%. We assessed potential inflation of test  
658 statistics due to sample relatedness and population stratification using LD score  
659 regression<sup>34</sup>. Here, an intercept value not significantly different from 1 indicates no such  
660 inflation, with a value over 1 indicating inflation.

661 A final list of index variants was first defined using a distance based metric, by which any  
662 SNPs passing the two tailed threshold of significance ( $P < 5 \times 10^{-8}$ ) within 1Mb of another  
663 significant SNP were considered to be located in the same locus. This list of signals was  
664 then further augmented using approximate conditional analysis in GCTA, using an LD  
665 reference panel from the UK Biobank study. Only secondary signals that were uncorrelated  
666 ( $r^2 < 0.05$ ) were included in the final list.

### 667 Replication and parent-of-origin testing

668 Replication of identified hits was performed in an independent sample of 39,486 women of  
669 European ancestry from the deCODE study, Iceland. Main effects and parent-of-origin  
670 association testing was performed using the same methodology as previously reported<sup>3,4</sup>.  
671 The fraction of variance explained by a variant associating under the additive model was  
672 calculated using the formula  $2 f (1-f) \beta_a^2$ , where  $f$  denotes the minor allele frequency of the  
673 variant and  $\beta_a$  is the additive effect. For variants associating under the recessive model, the  
674 formula  $f_h (1-f_h) \beta_r^2$  was used, where  $f_h$  denotes the homozygous frequency of the variant  
675 and  $\beta_r$  denotes the recessive effect. For variants associating under parent-of-origin models,  
676 fraction of variance explained was computed using the formulas  $f (1-f) \beta_m^2$  for the maternal  
677 model and  $f (1-f) \beta_p^2$  for the paternal model, where  $f$  denotes the minor allele frequency of  
678 the variant,  $\beta_m$  denotes the effect under the maternal model and  $\beta_p$  denotes the effect under  
679 the paternal model. Variance explained across multiple SNPs was calculated by summing  
680 the individual variances for all uncorrelated variants. We also estimate variance explained for  
681 top hits in UK Biobank using a combined allele score of all 377 autosomal genetic variants.  
682 Each individual variant was weighted using effect estimates derived from a meta-analysis  
683 excluding UK Biobank.

684



## 685 **Age at voice breaking in men**

686 Data on male voice breaking were available from two sources. Firstly, the 23andMe, Inc.  
687 study recorded recalled age at voice breaking in a sample of 55,871 men, as previously  
688 described<sup>5</sup>. This was recorded as a quantitative trait into pre-defined 2-year age bins by  
689 online questionnaire in response to the question “How old were you when your voice began  
690 to crack/deepen?”<sup>5</sup>. Individual SNP effect estimates from the two year age bins were  
691 rescaled to 1 year estimates for both voice breaking and AAM as reported previously.

692 Age at voice breaking was also recalled in the UK Biobank study, as previously described<sup>33</sup>.  
693 This was recorded as a categorical trait: “younger than average”, “about average age”, “older  
694 than average”, “do not know” or “prefer not to answer” in response to the question “When did  
695 your voice break”. In separate models, the earlier or later voice breaking groups were  
696 compared to the average group (used as the reference group).

## 697 **Disproportionate effects on early or late puberty timing**

698 Disproportionate effects on early or late puberty timing of AAM-associated SNPs were tested  
699 for AAM in UK Biobank. The distribution of AAM was divided into approximate quintiles, as  
700 previously reported<sup>33</sup>. Odds ratios for being in the youngest quintile (range 8-11) or the  
701 oldest (range 15-19) were compared to the middle quintile (age 13) as the reference, for  
702 each AAM-associated SNP and also for a combined weighted AAM-increasing allele score,  
703 with weights derived from a meta-analysis of all other studies except for UK Biobank.  
704 Sensitivity tests were performed by dividing UK Biobank individuals into broad strata based  
705 on birth year (before or after 1950) and geographic location (attendance at a study  
706 assessment centre in the North or South of the UK, as indicated by a line joining Mersey-  
707 Humber).

## 708 **Genetic correlation and genome-wide variance analysis**

709 Genome-wide genetic correlations with adult BMI<sup>22</sup> and voice breaking<sup>5</sup> were estimated  
710 using LD score regression implemented in LDSC<sup>34</sup>. The total trait variance of all genotyped  
711 SNPs was calculated using Restricted Estimate Maximum Likelihood (REML) implemented  
712 in BOLT<sup>35</sup>. This was estimated using the same UK Biobank study sample in the discovery  
713 analysis, excluding any related individuals. The proportion of heritability explained by index  
714 SNPs was estimated by dividing the variance explained by the index SNPs, by the total  
715 variance explained by all genotyped SNPs genome-wide.

## 716 **Mendelian randomisation analyses**

717 Individual genotype data on cancer outcomes were available from the Breast Cancer  
718 Association Consortium (BCAC) and Endometrial Cancer Association Consortium (ECAC).  
719 In addition, summary level results for ovary and prostate cancer were made available from  
720 the Ovarian Cancer Association Consortium (OCAC) and the Prostate Cancer Association  
721 Group to Investigate Cancer Associated Alterations in the Genome (PRACTICAL)  
722 consortium, respectively. Total analysed numbers were: 47,800 breast cancer cases and  
723 40,302 controls, 4401 endometrial cancer cases and 28,758 controls, 18,175 ovarian cancer  
724 cases and 26,134 controls, and 20,219 prostate cancer cases and 20,440 controls (from the  
725 PRACTICAL iCOGS dataset).

726 We performed Mendelian randomisation analyses to assess the likely causal effects of  
727 puberty timing on the risks for various sex steroid-sensitive cancers. Hence, AAM was  
728 predicted by a weighted genetic risk score of all 375 autosomal AAM-associated SNPs, and  
729 genetically-predicted AAM was tested for association with each cancer in a logistic  
730 regression model. The individual SNP genotype dosages comprising this score were  
731 imputed using the 1000 Genomes reference panel (minimum imputation  $r^2=0.43$ , median  
732 0.95). To avoid potential confounding by effects of the AAM genetic risk score on BMI, we  
733 performed BMI-adjusted analyses by including in models as a covariate the same AAM  
734 genetic risk score, but weighting each SNP for its effect on BMI (rather than on AAM) in the  
735 same study sample. Hence, we estimated the effect of genetically-predicted AAM controlling  
736 for genetically-predicted BMI by the same SNPs. BMI weighting was based on the  
737 association between each SNP and adult BMI in this sample (childhood BMI measurements  
738 were not available but there is reportedly high genetic correlation between adult and  
739 childhood obesity ( $rg=0.73$ )<sup>36</sup>. We did not adjust for measured BMI because such  
740 measurements in prevalent cancer cases are likely to introduce bias. As sensitivity tests,  
741 three further genetic score associations were performed for each cancer outcome: firstly,  
742 AAM predicted by the 314 AAM-associated SNPs that *were not* also individually associated  
743 with BMI in the BCAC iCOGs sample (at a nominal level of  $p<0.05$ ); secondly, AAM  
744 predicted by the 61 AAM-associated SNPs that *were* also associated with BMI in this sample  
745 (i.e  $P<0.05$ ); finally, AAM predicted by all 375 autosomal AAM-associated SNPs (unadjusted  
746 for BMI). To further consider pleiotropy, we tested for presence of heterogeneity between  
747 AAM-associated SNPs and analysed MR-Egger regression models<sup>37</sup>.

#### 748 **Pathway analyses**

749 Meta-Analysis Gene-set Enrichment of variaNT Associations (MAGENTA) was used to  
750 explore pathway-based associations in the full GWAS dataset. MAGENTA implements a  
751 gene set enrichment analysis (GSEA) based approach, as previously described<sup>38</sup>. Briefly,  
752 each gene in the genome is mapped to a single index SNP with the lowest P-value within a  
753 110 kb upstream, 40 kb downstream window. This P-value, representing a gene score, is  
754 then corrected for confounding factors such as gene size, SNP density and LD-related  
755 properties in a regression model. Genes within the HLA-region were excluded from analysis  
756 due to difficulties in accounting for gene density and LD patterns. Each mapped gene in the  
757 genome is then ranked by its adjusted gene score. At a given significance threshold (95th  
758 and 75th percentiles of all gene scores), the observed number of gene scores in a given  
759 pathway, with a ranked score above the specified threshold percentile, is calculated. This  
760 observed statistic is then compared to 1,000,000 randomly permuted pathways of identical  
761 size. This generates an empirical GSEA P-value for each pathway. Significance was  
762 determined when an individual pathway reached a false discovery rate (FDR)  $<0.05$  in either  
763 analysis. In total, 3216 pathways from Gene Ontology, PANTHER, KEGG and Ingenuity  
764 were tested for enrichment of multiple modest associations with AAM. MAGENTA software  
765 was also used for enrichment testing of custom gene sets.

#### 766 **Gene expression data integration**

767 In order to identify which tissues and cell types were most relevant to genes involved in  
768 pubertal development, we used an applied LD score regression<sup>39</sup> to specifically expressed  
769 genes ("LDSC-SEG")<sup>8</sup>. For each tissue, we ranked genes by a t-statistic for differential  
770 expression, using sex and age as covariates, and excluding all samples in related tissues.

771 For example, we compared expression in hippocampus samples to expression in all non-  
772 brain samples. We then took the top 10% of genes by this ranking, formed a genome  
773 annotation including these genes (exons and introns) plus 100kb on either side, and used  
774 stratified LD score regression to estimate the contribution of this annotation to per-SNP AAM  
775 heritability, adjusting for all categories in the baseline model<sup>39</sup>. We computed significance  
776 using a block jackknife over SNPs, and corrected for 46 hypotheses tested at  $P=0.05$ .

777 To identify specific eQTL linked genes, we utilised two complementary approaches to  
778 systematically integrate publicly available gene expression data with our genome-wide  
779 dataset:

780 Summary Mendelian Randomization (SMR) uses summary-level gene expression data to  
781 map potentially functional genes to trait-associated SNPs<sup>7</sup>. We ran this approach against the  
782 publicly available whole-blood eQTL dataset published by Westra et al.<sup>6</sup>, giving association  
783 statistics for 5,950 transcripts. A conservative significance threshold was set at  $P<8.4\times 10^{-6}$ ,  
784 in addition to a heterogeneity in dependent instruments (HEIDI) test statistic  $P>0.009$  for any  
785 variants which surpass the main threshold.

786 MetaXcan, a meta-analysis extension of the PrediXcan method<sup>40</sup>, was used to infer the  
787 association between genetically predicted gene expression (GPGE) and AAM. PrediXcan is  
788 a novel gene-based data aggregation and integration method which incorporates information  
789 from gene-expression data and GWAS data to translate evidence of association with a  
790 phenotype from the SNP-level to the gene. Briefly, PrediXcan first imputes gene-expression  
791 at an individual level using prediction models trained on measured transcriptome datasets  
792 with genome-wide SNP data and then regresses the imputed transcriptome levels with  
793 phenotype of interest. MetaXcan extends its application to allow inference of the direction  
794 and magnitude of GPGE-phenotype associations with only summary GWAS statistics, which  
795 is advantageous when SNP-phenotype associations result from a meta-analysis setting and  
796 also when individual level data are not available. As input we utilized GWAS meta-analysis  
797 summary statistics for AAM, LD matrix from the 1000 Genomes project, and as weights,  
798 gene-expression regression coefficients for SNPs from models trained with transcriptome  
799 data (V6p) from the GTEx Project<sup>41</sup>. GTEx is a large-scale collaborative effort where DNA  
800 and RNA from multiple tissues were sequenced from almost 1,000 deceased individuals of  
801 European, African, and Asian ancestries. MetaXcan analyses were targeted to those tissue  
802 types with prior evidence of association with AAM (based on the GTEx enrichment analyses  
803 described above). The threshold for statistical significance was estimated using the  
804 Bonferroni method for multiple testing correction across all tested tissues ( $P<2.57\times 10^{-6}$ ).

## 805 **Motif enrichment testing**

806 We identified transcription factors whose binding could be disrupted by AAM associated  
807 variants in enhancer regions by combining predicted enhancer regions across 111 human  
808 cell types and tissues with predicted motif instances of 651 transcription factor families as  
809 previously described<sup>42</sup>.

810 Briefly, we defined enhancer regions by first applying ChromHMM<sup>43</sup>, training a 15-state  
811 model for each reference epigenome on 5 histone modifications: H3K4me1, H3K4me3,  
812 H3K36me3, H3K9me3, and H3K27me3. We then produced a higher confidence set of  
813 predicted enhancer regions in each reference epigenome by intersecting DNaseI

814 hypersensitive sites (taking the union over 53 reference epigenomes for which DNase-Seq  
815 was performed) with enhancer-like chromatin states predicted in that reference  
816 epigenomes<sup>42</sup>. We defined 226 disjoint enhancer modules with distinct patterns of activity by  
817 hierarchically clustering the high confidence regions according to their patterns of activity  
818 (presence/absence) across the 111 reference epigenomes.

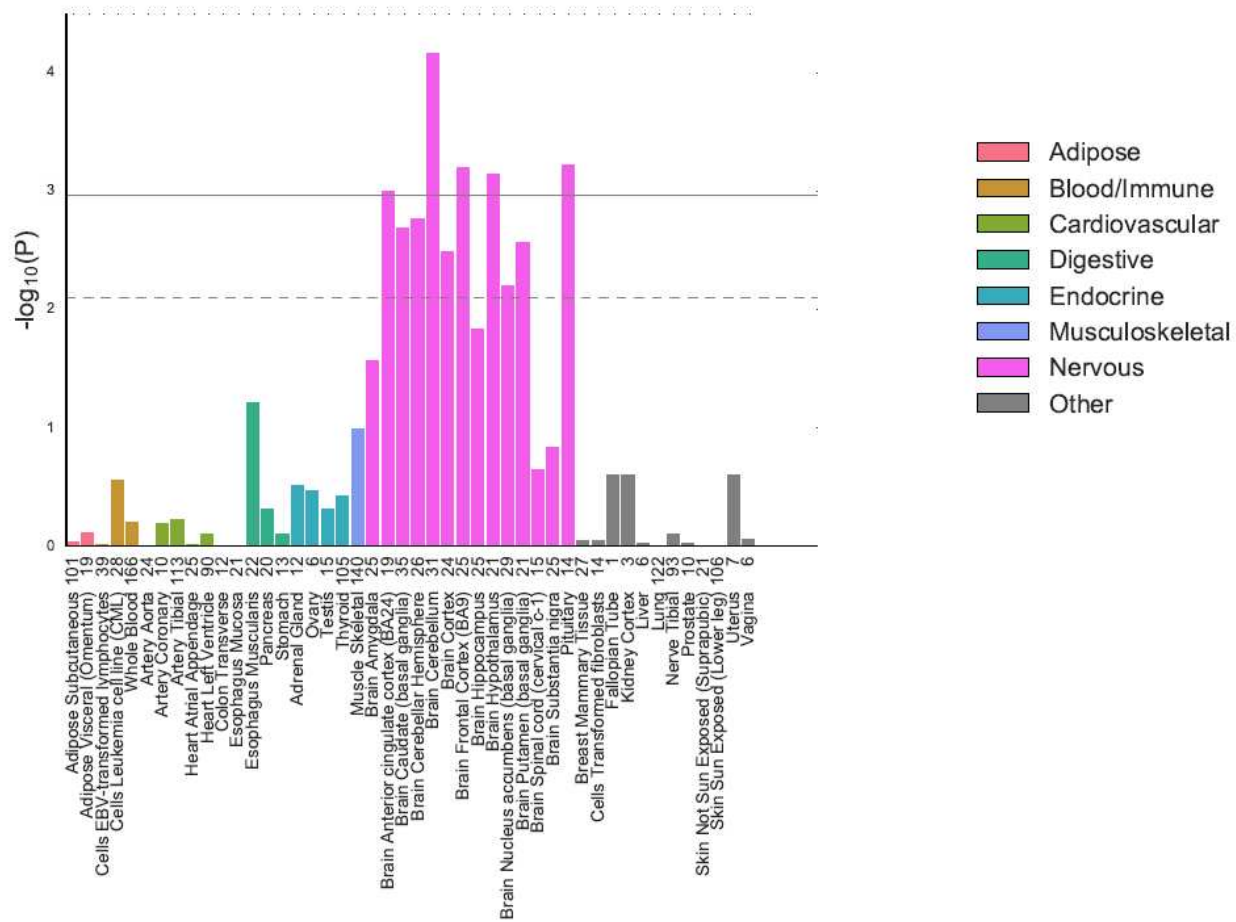
819 We predicted motif instances by first building a database of position weight matrices (PWMs)  
820 combining known motifs from Transfac and Jaspar with de novo discovered motifs in 427  
821 ChIP-Seq experiments for 123 transcription factors from ENCODE<sup>44</sup>. We predicted active  
822 regulators in each enhancer module by computing the enrichment of true PWM matches in  
823 the set of regions assigned to that module against the background of shuffled PWM  
824 matches. We only considered PWMs with conservation score at least 0.3, and used log2-fold  
825 enrichment > 1.5 as the significance cutoff.

826 We used the full set of AAM association summary statistics, excluding the 23andMe  
827 component, to identify a heuristic p-value threshold<sup>42</sup>. Briefly, we pruned a set of 8,094,080  
828 variants to 432,550 independent loci (pairwise  $r^2 < 0.1$ ). We scored each locus as the  
829 proportion of variants in the locus overlapping a predicted enhancer region, ranked loci by  
830 the best p-value in the locus, and then plotted enrichment curves comparing the cumulative  
831 score every 100 loci against the expected score for that total number of loci under the null  
832 where the score increases uniformly to the genome-wide value. We defined the right-most  
833 elbow point (inflection point) among all the enrichment curves as the heuristic p-value cutoff.

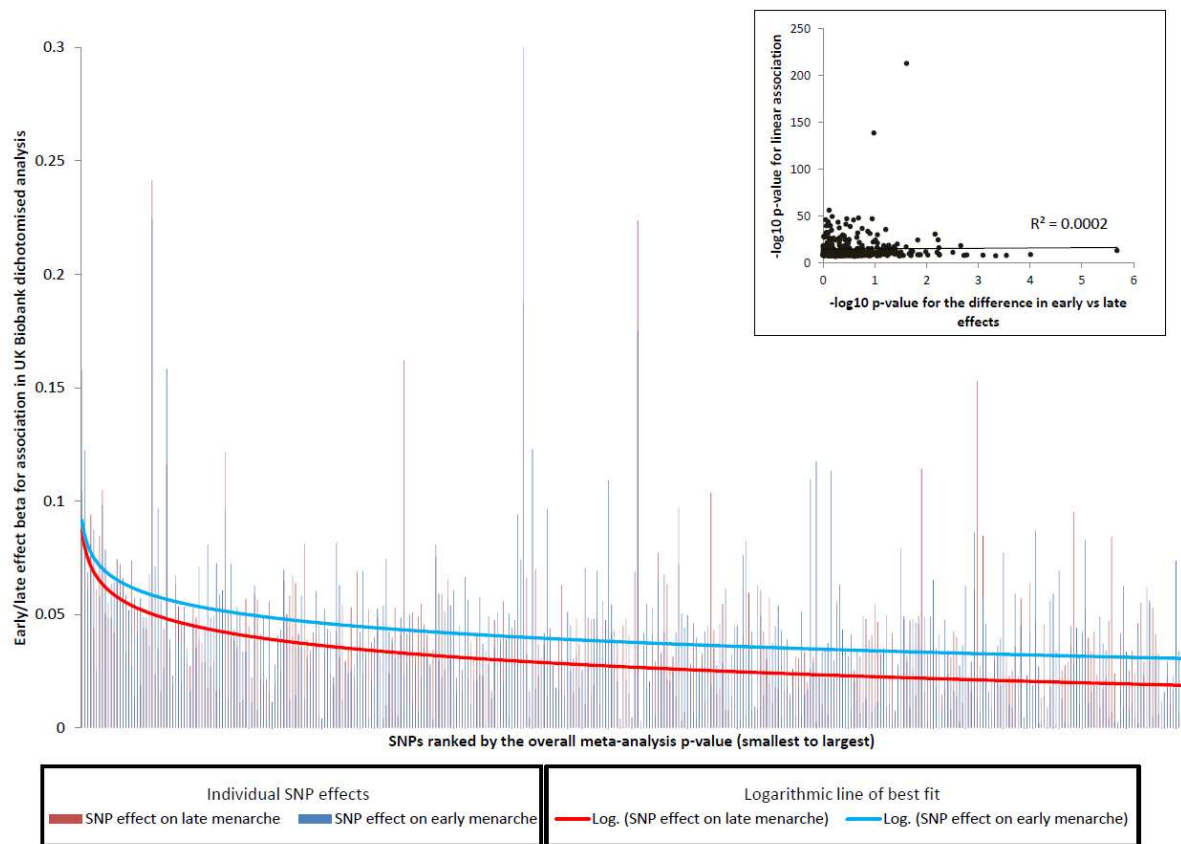
834 For each combination of enhancer module and predicted regulator, we constructed a 2x2  
835 contingency table counting enhancer regions in that module partitioned by presence of that  
836 motif and orthogonally by presence of an AAM association (based on the heuristic p-value  
837 cutoff described above). We restricted the set of regions to the domain on which motifs were  
838 discovered (excluding coding regions, 3' UTRs, transposons, and repetitive regions) and  
839 additionally to the subset of regions which harbor an imputed SNP for the disease. We  
840 computed one-sided p-values using Fisher's exact test.

## 841 **Hi-C integration**

842 Significant Hi-C interactions and contact domains were obtained from Rao et al. (GSE63525)  
843 for 6 ENCODE cell lines: K562, GM12878, HeLa-S3, IMR90, NHEK, and HUVEC. Their  
844 Juicer pipeline assigns statistical significance to each Hi-C interaction at resolutions ranging  
845 from 5kb-25kb, depending on coverage, at a 10% False Discovery Rate (FDR). Contact  
846 domains are genomic regions enriched for regulatory interactions and are more conserved  
847 across cell types than are specific interactions. They are conceptually similar to  
848 Topologically Associating Domains (TADs, Dixon et al. 2012) but with improved resolution  
849 (185kb median length vs. 880kb). We used the intersect command of bedtools to produce a  
850 list of significantly interacting Hi-C fragments containing one or more of our identified SNPs  
851 in either fragment from any of the six cell lines. For each SNP-containing fragment, genes  
852 present in the corresponding interacting fragment were identified as potential regulatory  
853 targets. As a second approach, we also scored genes based on the number of ENCODE cell  
854 types in which they were in the same contact domain as a SNP.

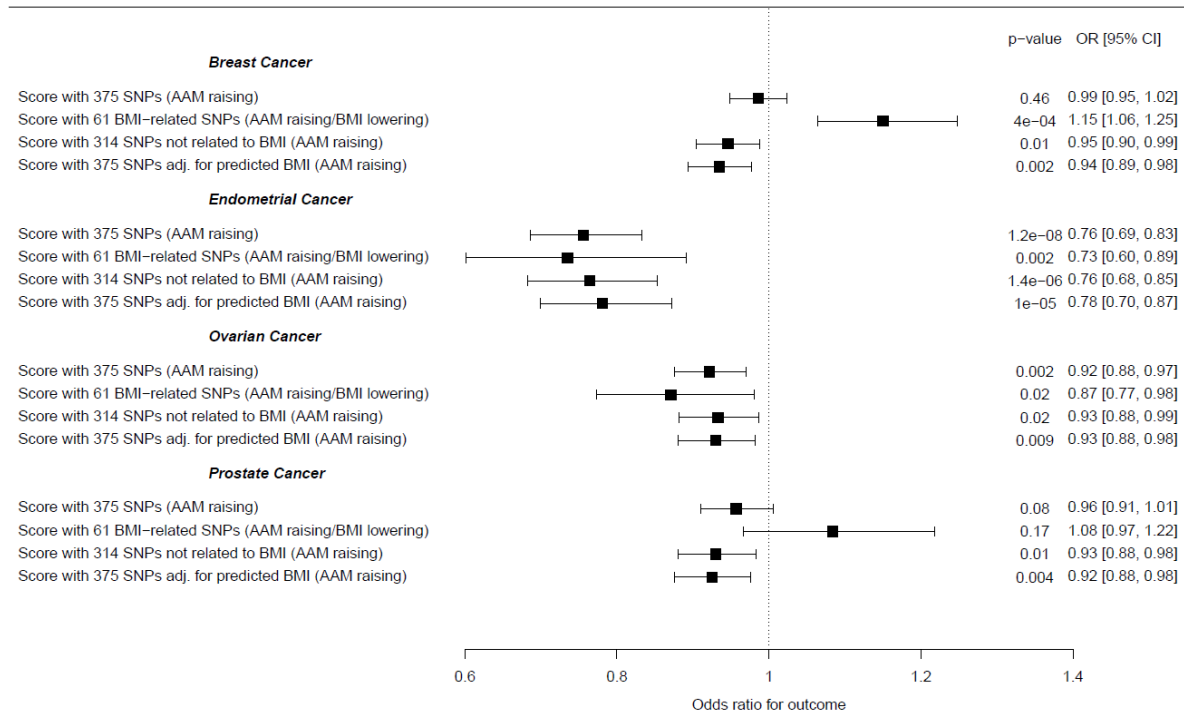


**Figure 1. GTEx tissue enrichment using LD score regression.** Numbers on the X-axis show sample number for each tissue. Dotted line represents significance at FDR < 5%, solid horizontal line represents Bonferroni-corrected significance for number of tissues tested.



**Figure 2. Stronger effects of age at menarche-associated signals on early menarche (blue) than late menarche (red) in women.** The 377 index menarche-associated SNPs are ordered from smallest to largest p-value for their continuous associations with age at menarche. The Y-axis indicates the log-odds ratio for each SNP on early menarche (blue; ages 8–11 years inclusive) or late menarche (red; 15–19 years inclusive). The reference group are women with menarche at 13 years. **Insert** shows the  $-\log_{10}$  p-values for the heterogeneity (based on Cochran’s Q) between the early and late menarche associations plotted against the  $-\log_{10}$  p-value for the continuous age at menarche association.

Associations between genetic scores and a range of cancers



**Figure 3. Effects and 95% confidence intervals of genetically-predicted age at menarche (AAM) on risks for various sex steroid-sensitive cancers, adjusted for the effects of the same AAM variants on BMI.** AAM was predicted by all 375 autosomal AAM-associated SNPs, and models were adjusted for the genetic effects of the same AAM variants on BMI. Three further genetic score associations are shown as sensitivity analyses for each outcome: firstly, AAM predicted by the 314 AAM-associated SNPs that *were not* also associated with BMI in the BCAC iCOGs sample (at a nominal level of  $p < 0.05$ ); secondly, AAM predicted by the 61 AAM-associated SNPs that *were* also associated with BMI in this sample; finally, AAM predicted by all 375 autosomal AAM-associated SNPs (unadjusted for BMI).

**Table 1: Parent-of-origin specific associations between sequence variants at *MKRN3*, *DLK1* and *MEG9* with age at menarche in Iceland (N=39,543).**

Marker	Position (hg38)	Allele		Freq. A1 (%)	Region	Additive		Maternal		Paternal		$P_{\text{mat vs. pat}}^2$
		A1	A2			$P$	$\beta^1$	$P$	$\beta^1$	$P$	$\beta^1$	
rs530324840 <sup>3</sup>	15:23,565,461	A	C	0.80	<i>MKRN3</i>	$4.4 \times 10^{-4}$	-0.206	$2.0 \times 10^{-1}$	0.098	$6.4 \times 10^{-11}$	-0.523	$1.3 \times 10^{-7}$
rs184950120 <sup>3</sup>	15:23,565,696	T	C	0.26	<i>MKRN3</i>	$1.0 \times 10^{-2}$	-0.265	$9.8 \times 10^{-1}$	0.003	$1.5 \times 10^{-4}$	-0.502	$4.9 \times 10^{-2}$
rs12148769 <sup>3</sup>	15:23,906,947	A	G	10.1	<i>MKRN3</i>	$5.8 \times 10^{-6}$	-0.078	$3.4 \times 10^{-1}$	-0.022	$9.2 \times 10^{-8}$	-0.120	$2.3 \times 10^{-3}$
rs138827001 <sup>4</sup>	14:100,771,634	T	C	0.36	<i>DLK1</i>	$6.8 \times 10^{-6}$	-0.387	$8.8 \times 10^{-1}$	-0.018	$4.7 \times 10^{-10}$	-0.704	$1.4 \times 10^{-4}$
rs10144321 <sup>4</sup>	14:100,416,068	G	A	23.0	<i>DLK1</i>	$5.6 \times 10^{-6}$	-0.056	$4.0 \times 10^{-1}$	-0.014	$1.9 \times 10^{-7}$	-0.084	$9.7 \times 10^{-3}$
rs7141210 <sup>4</sup>	14:100,716,133	T	C	38.2	<i>DLK1</i>	$4.5 \times 10^{-2}$	0.021	$1.5 \times 10^{-1}$	-0.021	$2.3 \times 10^{-5}$	0.059	$4.0 \times 10^{-4}$
rs61992671 <sup>5</sup>	14:101,065,517	A	G	48.5	<i>MEG9</i>	$4.7 \times 10^{-3}$	-0.029	$6.0 \times 10^{-8}$	-0.077	$2.7 \times 10^{-1}$	0.015	$1.9 \times 10^{-5}$

1.  $\beta$  indicates the effect of allele A1 in years per allele.

2.  $P$ -value for heterogeneity between paternal and maternal allele associations.

3. rs530324840 is a novel variant identified by the parent-of-origin specific analysis. rs184950120 is the rare variant identified by the meta-analysis. rs12148769 is the previously reported intergenic common signal (Ref. 3).

4. rs138827001 is a novel variant identified by the parent-of-origin specific analysis. rs10144321 and rs7141210 are previously reported common variants (Ref. 3).

5. rs61992671 is a suggestive novel parent-of-origin specific association signal.



## References

1. Parent, A.S. *et al.* The timing of normal puberty and the age limits of sexual precocity: variations around the world, secular trends, and changes after migration. *Endocr Rev* **24**, 668-93 (2003).
2. Perry, J.R., Murray, A., Day, F.R. & Ong, K.K. Molecular insights into the aetiology of female reproductive ageing. *Nat Rev Endocrinol* **11**, 725-34 (2015).
3. Perry, J.R. *et al.* Parent-of-origin-specific allelic associations among 106 genomic loci for age at menarche. *Nature* **514**, 92-97 (2014).
4. Lunetta, K.L. *et al.* Rare coding variants and X-linked loci associated with age at menarche. *Nat Commun* **6**, 7756 (2015).
5. Day, F.R. *et al.* Genetic determinants of puberty timing in men and women: shared genetic aetiology between sexes and with health-related outcomes. *Nat Commun* **6**, 8842 (2015).
6. Westra, H.J. *et al.* Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat Genet* **45**, 1238-43 (2013).
7. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet* **48**, 481-7 (2016).
8. Finucane, H.K. *et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. Preprint at *bioRxiv* <https://doi.org/10.1101/103069> (2017).
9. Rao, S.S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665-80 (2014).
10. Smemo, S. *et al.* Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature* **507**, 371-5 (2014).
11. Mbarek, H. *et al.* Identification of Common Genetic Variants Influencing Spontaneous Dizygotic Twinning and Female Fertility. *Am J Hum Genet* **98**, 898-908 (2016).
12. Ong, K.K. *et al.* Genetic variation in LIN28B is associated with the timing of puberty. *Nat Genet* **41**, 729-733 (2009).
13. Perry, J.R. *et al.* Meta-analysis of genome-wide association data identifies two loci influencing age at menarche. *Nat Genet* **41**, 648-650 (2009).
14. Zhang, J. *et al.* LIN28 Regulates Stem Cell Metabolism and Conversion to Primed Pluripotency. *Cell Stem Cell* **19**, 66-80 (2016).
15. Zhu, H. *et al.* Lin28a transgenic mice manifest size and puberty phenotypes identified in human genetic association studies. *Nat Genet* **42**, 626-30 (2010).
16. Zhu, H. *et al.* The Lin28/let-7 axis regulates glucose metabolism. *Cell* **147**, 81-94 (2011).
17. Baran, Y. *et al.* The landscape of genomic imprinting across diverse adult human tissues. *Genome Res* **25**, 927-36 (2015).
18. van den Berg, S.M. & Boomsma, D.I. The familial clustering of age at menarche in extended twin families. *Behav Genet* **37**, 661-7 (2007).
19. Ahlgren, M., Melbye, M., Wohlfahrt, J. & Sorensen, T.I. Growth patterns and the risk of breast cancer in women. *N Engl J Med* **351**, 1619-26 (2004).
20. Collaborative Group on Hormonal Factors in Breast, C. Menarche, menopause, and breast cancer risk: individual participant meta-analysis, including 118 964 women with breast cancer from 117 epidemiological studies. *Lancet Oncol* **13**, 1141-51 (2012).

21. Bhaskaran, K. *et al.* Body-mass index and risk of 22 specific cancers: a population-based cohort study of 5.24 million UK adults. *Lancet* **384**, 755-65 (2014).
22. Locke, A.E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197-206 (2015).
23. Gao, C. *et al.* Mendelian randomization study of adiposity-related traits and risk of breast, ovarian, prostate, lung and colorectal cancer. *Int J Epidemiol* **45**, 896-908 (2016).
24. Guo, Y. *et al.* Genetically predicted body mass index and breast cancer risk: Mendelian randomization analyses of data from 145,000 women of European descent. *PLoS Med* **13**, 1002105 (2016).
25. Abreu, A.P. *et al.* Central precocious puberty caused by mutations in the imprinted gene MKRN3. *N Engl J Med* **368**, 2467-75 (2013).
26. da Rocha, S.T., Edwards, C.A., Ito, M., Ogata, T. & Ferguson-Smith, A.C. Genomic imprinting at the mammalian Dlk1-Dio3 domain. *Trends Genet* **24**, 306-16 (2008).
27. Giles, G.G. *et al.* Early growth, adult body size and prostate cancer risk. *Int J Cancer* **103**, 241-5 (2003).
28. de Vries, L., Kauschansky, A., Shohat, M. & Phillip, M. Familial central precocious puberty suggests autosomal dominant inheritance. *J Clin Endocrinol Metab* **89**, 1794-800 (2004).
29. Wehkalampi, K., Widen, E., Laine, T., Palotie, A. & Dunkel, L. Patterns of inheritance of constitutional delay of growth and puberty in families of adolescent girls and boys referred to specialist pediatric care. *J Clin Endocrinol Metab* **93**, 723-8 (2008).
30. Winkler, T.W. *et al.* Quality control and conduct of genome-wide association meta-analyses. *Nat Protoc* **9**, 1192-212 (2014).
31. Willer, C.J., Li, Y. & Abecasis, G.R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190-1 (2010).
32. Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* **12**, e1001779 (2015).
33. Day, F., Elks, C.E., Murray, A.M., Ong, K.K. & Perry, J.R. Puberty timing associated with diabetes, cardiovascular disease and also diverse health outcomes in men and women: the UK Biobank study. *Sci Rep* **5**, 11208 (2015).
34. Bulik-Sullivan, B.K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **47**, 291-5 (2015).
35. Loh, P.R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat Genet* **47**, 284-90 (2015).
36. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat Genet* **47**, 1236-41 (2015).
37. Bowden, J., Davey Smith, G. & Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int J Epidemiol* **44**, 512-25 (2015).
38. Segre, A.V. *et al.* Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. *PLoS Genet* **6**, e1001058 (2010).
39. Finucane, H.K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet* **47**, 1228-35 (2015).

40. Gamazon, E.R. *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet* **47**, 1091-8 (2015).
41. Consortium, G.T. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580-5 (2013).
42. Sarkar, A., Ward, L.D. & Kellis, M. Functional enrichments of disease variants across thousands of independent loci in eight diseases. Preprint at *bioRxiv* <http://dx.doi.org/10.1101/048066> (2016).
43. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43-9 (2011).
44. Roadmap Epigenomics, C. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-30 (2015).

## Acknowledgements

This research has been conducted using the UK Biobank Resource under application 5122 and 9797. Full study-specific acknowledgements can be found in the online supplement.

## Competing financial interests

The authors declare no competing financial interests

## Data availability statement

GWAS meta-analysis summary statistics from the ReproGen consortium are available to download from the ReproGen website ([www.reprogen.org](http://www.reprogen.org)).

## Author Contributions

All authors reviewed the original and revised manuscripts. Statistical analysis: F.R.D, D.J.T, H.H, D.I.C, H.F, P.S, K.S.R, S.W, A.Sa, E.Alb, E.Alt, M.A, C.M.B, T.Bo, A.Ca, E.D, A.G, C.He, J.J.H, R.K, I.K, P.L, K.L.L, M.M, B.M, G.M, S.E.M, I.M.N, R.N, T.N, L.P, N.Per, E.P, L.M.R, K.E.S, A.Se, A.V.S, L.S, A.T, J.R.B.P. Sample collection, genotyping and phenotyping: I.L.A, S.Ba, M.W.B, J.B, S.Be, M.B, E.B, S.E.B, M.K.B, J.S.B, H.Bra, H.Bre, L.B, T.Br, J.E.B, H.C, E.C, S.C, G.C, T.C, F.J.C, D.L.C, A.Co, L.C, K.C, G.D, E.J.C.N.d, R.d, I.DeV, J.D, P.D, I.D-S, A.M.D, J.G.E, P.A.F, L.F-R, L.Fe, D.F, L.Fr, M.G, I.G, G.G.G, H.G, D.F.G, P.G, P.H, E.H, U.H, T.B.H, C.A.H, G.H, M.J.H, J.L.H, F.H, D.Hu, A.I, H.I, M.J, P.K.J, D.K, Z.K, G.L, D.L, C.L, L.J.L, J.S.E.L, S.Le, J.Li, P.A.L, S.Li, Y.L, J.Lu, R.M, A.Ma, H.M, M.I.M, C.Mei, T.M, C.Men, A.Me, K.M, L.M, R.L.M, G.W.M, A.M.M, M.A.N, P.N, H.N, D.R.N, A.J.O, T.A.O, S.P, A.Pa, N.Ped, A.Pe, J.P, P.D.P.P, A.Po, P.R, I.Ra, S.M.R, A.R, F.R.R, I.Ru, R.R, D.R, C.F.S, M.K.S, R.A.S, M.Sh, R.S, M.C.S, U.S, M.Sta, M.Ste, K.Str, T.Ta, E.T, N.J.T, M.T, T.Tr, J.P.T, A.G.U, D.R.V, V.V, U.V, P.V, Q.W, E.W, K.W, G.W, R.W, B.H.RW, J.Z, M.Zo, M.Zy. Individual study principal investigators: B.Z.A, D.I.B, M.C, F.C, T.E, N.F, C.G, V.G, C.Ha, P.K, D.A.L, P.K.EM, N.G.M, D.O.M, E.A.N, O.P, D.P, A.L.P, P.M.R, H.S, T.D.S, D.S, D.T, S.U, J.A.V, H.V, N.J.W, J.F.W, A.B.S, U.T, K.P, D.F.E, J.Y.T, J.C, D.Hi, A.Mu, J.M.M, K.Ste, K.K.O, J.R.B.P. Working group: F.R.D, D.J.T, H.H, D.I.C, H.F, P.S, K.S.R, S.W, A.Sa, A.B.S, U.T, K.P, D.F.E, J.Y.T, J.C, D.Hi, A.Mu, J.M.M, K.Ste, K.K.O, J.R.B.P.