bioRxiv preprint doi: https://doi.org/10.1101/144410; this version posted May 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Refining the accuracy of validated target identification through coding variant fine-mapping in type 2 diabetes

Anubha Mahajan[1], Jennifer Wessel[2], Sara M Willems[3], Wei Zhao[4], Neil R Robertson[1,5], Audrey Y Chu[6,7], Wei Gan[1], Hidetoshi Kitajima[1], Daniel Taliun[8], N William Rayner[1,5,9], Xiuqing Guo[10], Yingchang Lu[11], Man Li[12,13], Richard A Jensen[14], Yao Hu[15], Shaofeng Huo[15], Kurt K Lohman[16], Weihua Zhang[17,18], James P Cook[19], Bram Prins[9], Jason Flannick[20,21], Niels Grarup[22], Vassily Vladimirovich Trubetskoy[8], Jasmina Kravic[23], Young Jin Kim[24], Denis V Rybin[25], Hanieh Yaghootkar[26], Martina Müller-Nurasyid[27,28,29], Karina Meidtner[30,31], Ruifang Li-Gao[32,33], Tibor V Varga[34], Jonathan Marten[35], Jin Li[36], Albert Vernon Smith[37,38], Ping An[39], Symen Ligthart[40], Stefan Gustafsson[41], Giovanni Malerba[42], Ayse Demirkan[40,43], Juan Fernandez Tajes[1], Valgerdur Steinthorsdottir[44], Matthias Wuttke[45], Cécile Lecoeur[46], Michael Preuss[11], Lawrence F Bielak[47], Marielisa Graff[48], Heather M Highland[49], Anne E Justice[48], Dajiang J Liu[50], Eirini Marouli[51], Gina Marie Peloso[20,25], Helen R Warren[51,52], ExomeBP Consortium, MAGIC Consortium, GIANT consortium, Saima Afaq[17], Shoaib Afzal[53,54,55], Emma Ahlqvist[23], Peter Almgren[56], Najaf Amin[40], Lia B Bang[57], Alain G Bertoni[58], Cristina Bombieri[42], Jette Bork-Jensen[22], Ivan Brandslund[59,60], Jennifer A Brody[14], Noël P Burtt[20], Mickaël Canouil[46], Yii-Der Ida Chen[10], Yoon Shin Cho[61], Cramer Christensen[62], Sophie V Eastwood[63], Kai-Uwe Eckardt[64], Krista Fischer[65], Giovanni Gambaro[66], Vilmantas Giedraitis[67], Megan L Grove[68], Hugoline G de Haan[33], Sophie Hackinger[9], Yang Hai[10], Sohee Han[24], Anne Tybjærg-Hansen[54,55,69], Marie-France Hivert[70,71,72], Bo Isomaa[73,74], Susanne Jäger[30,31], Marit E Jørgensen[75,76], Torben Jørgensen[55,77,78], Annemari Käräjämäki[79,80], Bong-Jo Kim[24], Sung Soo Kim[24], Heikki A Koistinen[81,82,83,84], Peter Kovacs[85], Jennifer Kriebel[31,86], Florian Kronenberg[87], Kristi Läll[65,88], Leslie A Lange[89], Jung-Jin Lee[4], Benjamin Lehne[17], Huaixing Li[15], Keng-Hung Lin[90], Allan Linneberg[77,91,92], Ching-Ti Liu[25], Jun Liu[40], Marie Loh[17,93,94], Reedik Mägi[65], Vasiliki Mamakou[95], Roberta McKean-Cowdin[96], Girish Nadkarni[97], Matt Neville[5,98], Sune F Nielsen[53,54,55], Ioanna Ntalla[51], Patricia A Peyser[99], Wolfgang Rathmann[31,100], Kenneth Rice[101], Stephen S Rich[102], Line Rode[53,54], Olov Rolandsson[103], Sebastian Schönherr[87],

bioRxiv preprint doi: https://doi.org/10.1101/144410; this version posted May 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Elizabeth Selvin[12], Kerrin S Small[104], Alena Stančáková[105], Praveen Surendran[106], Kent D Taylor[10], Tanya M Teslovich[8], Barbara Thorand[31,107], Gudmar Thorleifsson[44], Adrienne Tin[108], Anke Tönjes[109], Anette Varbo[53,54,55,69], Daniel R Witte[110,111], Andrew R Wood[26], Pranav Yajnik[8], Jie Yao[10], Loïc Yengo[46], Robin Young[106,112], Philippe Amouyel[113], Heiner Boeing[114], Eric Boerwinkle[68,115], Erwin P Bottinger[11], Rajiv Chowdhury[116], Francis S Collins[117], George Dedoussis[118], Abbas Dehghan[40,119], Panos Deloukas[51,120], Marco M Ferrario[121], Jean Ferrières[122,123], Jose C Florez[70,124,125,126], Philippe Frossard[127], Vilmundur Gudnason[37,38], Tamara B Harris[128], Susan R Heckbert[129], Joanna M M Howson[116], Martin Ingelsson[67], Sekar Kathiresan[20,126,130,131], Frank Kee[132], Johanna Kuusisto[105], Claudia Langenberg[3], Lenore J Launer[128], Cecilia M Lindgren[1,20,133], Satu Männistö[134], Thomas Meitinger[135,136], Olle Melander[56], Karen L Mohlke[137], Marie Moitry[138,139], Andrew D Morris[140,141], Alison D Murray[142], Renée de Mutsert[33], Marju Orho-Melander[143], Katharine R Owen[5,98], Markus Perola[134,144], Annette Peters[29,31,107], Michael A Province[39], Asif Rasheed[127], Paul M Ridker[7,126], Fernando Rivadineira[40,145], Frits R Rosendaal[33], Anders H Rosengren[23], Veikko Salomaa[134], Wayne H -H Sheu[146], Rob Sladek[147,148,149], Blair H Smith[150], Konstantin Strauch[27,151], André G Uitterlinden[40,145], Rohit Varma[152], Cristen J Willer[153,154,155], Matthias Blüher[85,109], Adam S Butterworth[106,156], John Campbell Chambers[17,18,157], Daniel I Chasman[7,126], John Danesh[106,156,158,159], Cornelia van Duijn[40], Josee Dupuis[6,25], Oscar H Franco[40], Paul W Franks[34,103,160], Philippe Froguel[46,161], Harald Grallert[31,86,162,163], Leif Groop[23,144], Bok-Ghee Han[24], Torben Hansen[22,164], Andrew T Hattersley[165], Caroline Hayward[35], Erik Ingelsson[41,166], Sharon LR Kardia[167], Fredrik Karpe[5,98], Jaspal Singh Kooner[18,157,168], Anna Köttgen[45], Kari Kuulasmaa[134], Markku Laakso[105], Xu Lin[15], Lars Lind[169], Yongmei Liu[58], Ruth J F Loos[11,170], Jonathan Marchini[1,171], Andres Metspalu[65], Dennis Mook-Kanamori[33,172], Børge G Nordestgaard[53,54,55], Colin N A Palmer[173], James S Pankow[174], Oluf Pedersen[22], Bruce M Psaty[175,176], Rainer Rauramaa[177], Naveed Sattar[178], Matthias B Schulze[30,31], Nicole Soranzo[9,156,179], Timothy D Spector[104], Kari Stefansson[38,44], Michael Stumvoll[180], Unnur Thorsteinsdottir[38,44], Tiinamaija Tuomi[74,82,144,181], Jaakko Tuomilehto[81,182,183,184], Nicholas J Wareham[3], James G Wilson[185], Eleftheria Zeggini[9], Robert A Scott[3], Inês Barroso[9,186], Timothy M Frayling[26], Mark O Goodarzi[187], James B Meigs[188], Michael Boehnke[8], Danish Saleheen[4,127], Andrew P Morris[1,19,65,*], Jerome I Rotter[189,*], Mark I McCarthy[1,5,98,*]

bioRxiv preprint doi: https://doi.org/10.1101/144410; this version posted May 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

1. Wellcome Trust Centre for Human Genetics, Nuffield Department of Medicine, University of Oxford, Oxford, OX3 7BN, UK.
2. Departments of Epidemiology and Medicine, Diabetes Translational Research Center, Indiana University, Indianapolis, IN, 46202-2872, USA.
3. MRC Epidemiology Unit, Institute of Metabolic Science, University of Cambridge, Cambridge, CB2 0QQ, UK.
4. Department of Biostatistics and Epidemiology, University of Pennsylvania, Philadelphia, Pennsylvania, 19104, USA.
5. Oxford Centre for Diabetes, Endocrinology and Metabolism, Radcliffe Department of Medicine, University of Oxford, Oxford, OX3 7LE, UK.
6. National Heart, Lung, and Blood Institute's Framingham Heart Study, Framingham, Massachusetts, 01702, USA.
7. Division of Preventive Medicine, Department of Medicine, Brigham and Women's Hospital, Boston, MA, 02215, USA.
8. Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, Michigan, 48109, USA.
9. Department of Human Genetics, Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire, CB10 1SA, UK.
10. Department of Pediatrics, The Institute for Translational Genomics and Population Sciences, LABioMed at Harbor-UCLA Medical Center, Torrance, California, 90502, US.
11. The Charles Bronfman Institute for Personalized Medicine, The Icahn School of Medicine at Mount Sinai, New York, 10029, USA.
12. Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, 21205, US.
13. Division of Nephrology and Hypertension, Department of Internal Medicine, University of Utah School of Medicine, Salt Lake City, Utah, 84132, US.
14. Cardiovascular Health Research Unit, Department of Medicine, University of Washington, Seattle, WA, 98101, USA.
15. Institute for Nutritional Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, University of the Chinese Academy of Sciences, Shanghai, People's Republic of China.

bioRxiv preprint doi: https://doi.org/10.1101/144410; this version posted May 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

16. Department of Biostatistical Sciences, Division of Public Health Sciences, Wake Forest University Health Sciences, Winston Salem, North Carolina, 27157, USA.
17. Department of Epidemiology and Biostatistics, Imperial College London, London, W2 1PG, UK.
18. Department of Cardiology, Ealing Hospital, London North West Healthcare NHS Trust, Middlesex, UB1 3HW, UK.
19. Department of Biostatistics, University of Liverpool, Liverpool, L69 3GA, UK.
20. Program in Medical and Population Genetics, Broad Institute, Cambridge, Massachusetts, 02142, USA.
21. Department of Molecular Biology, Massachusetts General Hospital, Boston, Massachusetts, 02114, USA.
22. The Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, 2100, Denmark.
23. Department of Clinical Sciences, Diabetes and Endocrinology, Lund University Diabetes Centre, Malmö, 20502, Sweden.
24. Center for Genome Science, Korea National Institute of Health, Chungcheongbuk-do, Republic of Korea.
25. Department of Biostatistics, Boston University School of Public Health, Boston, Massachusetts, 02118, USA.
26. Genetics of Complex Traits, University of Exeter Medical School, University of Exeter, Exeter, EX1 2LU, UK.
27. Institute of Genetic Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, 85764, Germany.
28. Department of Medicine I, University Hospital Grosshadern, Ludwig-Maximilians-Universität, Munich, 81377, Germany.
29. DZHK (German Centre for Cardiovascular Research), partner site Munich Heart Alliance, Munich, 81675, Germany.
30. Department of Molecular Epidemiology, German Institute of Human Nutrition Potsdam-Rehbruecke (DIfE), Nuthetal, 14558, Germany.
31. German Center for Diabetes Research (DZD), Neuherberg, 85764, Germany.
32. Department of Clinical Epidemiology, Leiden, 2300 RC, The Netherlands.

bioRxiv preprint doi: https://doi.org/10.1101/144410; this version posted May 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

33. Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, 2300 RC, The Netherlands.
34. Department of Clinical Sciences, Lund University Diabetes Centre, Genetic and Molecular Epidemiology Unit, Lund University, Malmö, SE-214 28, Sweden.
35. MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, EH4 2XU, UK.
36. Division of Cardiovascular Medicine, Department of Medicine, Stanford University School of Medicine, Palo Alto, CA, 94304, US.
37. Icelandic Heart Assocition, Kopavogur, 201, Iceland.
38. Faculty of Medicine, University of Iceland, Reykjavik, 101, Iceland.
39. Department of Genetics Division of Statistical Genomics, Washington University School of Medicine, St. Louis, Missouri, 63110, USA.
40. Department of Epidemiology, Erasmus University Medical Center, Rotterdam, 3015CN, The Netherlands.
41. Department of Medical Sciences, Molecular Epidemiology and Science for Life Laboratory, Uppsala University, Uppsala, 75185, Sweden.
42. Section of Biology and Genetics, Department of Neurosciences, Biomedicine and Movement sciences, University of Verona, Verona, 37134, Italy.
43. Department of Human Genetics, Leiden University Medical Center, Leiden, Netherlands.
44. deCODE Genetics, Amgen inc., Reykjavik, 101, Iceland.
45. Institute of Genetic Epidemiology, Medical Center – University of Freiburg, Faculty of Medicine, University of Freiburg, Freiburg, Freiburg, 79106, Germany.
46. CNRS-UMR8199, Lille University, Lille Pasteur Institute, Lille, 59000, France.
47. Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, Michigan, 48109, USA.
48. Department of Epidemiology, University of North Carolina, Chapel Hill, NC, 27514, USA.
49. Human Genetics Center, The University of Texas Graduate School of Biomedical Sciences at Houston, The University of Texas Health Science Center at Houston, Houston, Texas, 77030, USA.

bioRxiv preprint doi: https://doi.org/10.1101/144410; this version posted May 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

50. Department of Public Health Sciences, Institute of Personalized Medicine, Penn State College of Medicine, Hershey, PA, USA.
51. William Harvey Research Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London, UK.
52. National Institute for Health Research, Barts Cardiovascular Biomedical Research Unit, Queen Mary University of London, London, London, EC1M 6BQ, UK.
53. Department of Clinical Biochemistry, Herlev and Gentofte Hospital, Copenhagen University Hospital, Herlev, 2730, Denmark.
54. The Copenhagen General Population Study, Herlev and Gentofte Hospital, Copenhagen University Hospital, Copenhagen, DK-2730, Denmark.
55. Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark.
56. Department of Clinical Sciences, Hypertension and Cardiovascular Disease, Lund University, Malmö, 20502, Sweden.
57. Department of Cardiology, Rigshospitalet, Copenhagen University Hospital, Copenhagen, 2100, Denmark.
58. Department of Epidemiology & Prevention, Public Health Sciences, Wake Forest University Health Sciences, Winston-Salem, NC, 27157-1063, USA.
59. Institute of Regional Health Research, University of Southern Denmark, Odense, 5000, Denmark.
60. Department of Clinical Biochemistry, Vejle Hospital, Vejle, 7100, Denmark.
61. Department of Biomedical Science, Hallym University, Chuncheon, Republic of Korea.
62. Medical Department, Lillebælt Hospital Vejle, Vejle, Denmark.
63. Institute of Cardiovascular Science, University College London, London, WC1E 6BT.
64. Department of Nephrology and Medical Intensive Care Charité, University Medicine Berlin, Berlin, 10117, Germany.
65. Estonian Genome Center, University of Tartu, Tartu, 51010, Estonia.
66. Università Cattolica del Sacro Cuore, Roma, 00168, Italy.
67. Department of Public Health and Caring Sciences, Geriatrics, Uppsala University, Uppsala, SE-751 85, Sweden.

bioRxiv preprint doi: https://doi.org/10.1101/144410; this version posted May 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

68. Human Genetics Center, Department of Epidemiology, Human Genetics, and Environmental Sciences, School of Public Health, The University of Texas Health Science Center at Houston, Houston, Texas, USA.
69. Department of Clinical Biochemistry, Rigshospitalet, Copenhagen University Hospital, Copenhagen, 2100, Denmark.
70. Diabetes Research Center (Diabetes Unit), Department of Medicine, Massachusetts General Hospital, Boston, Massachusetts, 02114, USA.
71. Department of Population Medicine, Harvard Pilgrim Health Care Institute, Harvard Medical School, Boston, MA, 02215, USA.
72. Department of Medicine, Universite de Sherbrooke, Sherbrooke, QC, J1K 2R1, Canada.
73. Malmska Municipal Health Care Center and Hospital, Jakobstad, 68601, Finland.
74. Folkhälsan Research Centre, Helsinki, 00014, Finland.
75. Steno Diabetes Center Copenhagen, Gentofte, 2820, Denmark.
76. National Institute of Public Health, Southern Denmark University, Copenhagen, 1353, Denmark.
77. Research Centre for Prevention and Health, Capital Region of Denmark, Glostrup, 2600, Denmark.
78. Faculty of Medicine, Aalborg University, Aalborg, Denmark.
79. Department of Primary Health Care, Vaasa Central Hospital, Vaasa, Finland.
80. Diabetes Center, Vaasa Health Care Center, Vaasa, Finland.
81. Department of Health, National Institute for Health and Welfare, Helsinki, 00271, Finland.
82. Endocrinology, Abdominal Center, Helsinki University Hospital, Helsinki, Finland, 00029.
83. Minerva Foundation Institute for Medical Research, Helsinki, Finland.
84. Department of Medicine, University of Helsinki and Helsinki University Central Hospital, Helsinki, Finland.
85. Integrated Research and Treatment (IFB) Center AdiposityDiseases, University of Leipzig, Leipzig, 04103, Germany.
86. Research Unit of Molecular Epidemiology, Institute of Epidemiology II, Helmholtz Zentrum München Research Center for Environmental Health, Neuherberg, 85764, Germany.

bioRxiv preprint doi: https://doi.org/10.1101/144410; this version posted May 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

87. Division of Genetic Epidemiology, Department of Medical Genetics, Molecular and Clinical Pharmacology, Medical University of Innsbruck, Innsbruck, 6020, Austria.
88. Institute of Mathematical Statistics, University of Tartu, Tartu, Estonia.
89. Department of Medicine, Division of Bioinformatics and Personalized Medicine, University of Colorado Denver, Aurora, CO, USA, 80045.
90. Department of Ophthalmology, Taichung Veterans General Hospital, Taichung, 40705, Taiwan.
91. Department of Clinical Experimental Research, Rigshospitalet, Glostrup, Denmark.
92. Department of Clinical Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark.
93. Institute of Health Sciences, University of Oulu, Oulu, 90014, Finland.
94. Translational Laboratory in Genetic Medicine (TLGM), Agency for Science, Technology and Research (A*STAR), Singapore, 138648, Singapore.
95. Dromokaiteio Psychiatric Hospital, National and Kapodistrian University of Athens, Athens, Greece.
96. Department of Preventive Medicine, Keck School of Medicine of the University of Southern California, Los Angeles, California, 90007, US.
97. Division of Nephrology, Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, 10069, USA.
98. Oxford NIHR Biomedical Research Centre, Oxford University Hospitals Trust, Oxford, OX3 7LE, UK.
99. Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, Michigan, 48109, USA.
100. Institute for Biometrics and Epidemiology, German Diabetes Center, Leibniz Center for Diabetes Research at Heinrich Heine University Düsseldorf, Düsseldorf, Germany.
101. Department of Biostatistics, University of Washington, Seattle, WA, 98195-7232, USA.
102. Center for Public Health Genomics, Department Public Health Sciences, University of Virginia School of Medicine, Charlottesville, Virginia, 22908, US.
103. Department of Public Health and Clinical Medicine, Umeå University, Umeå, 90187, Sweden.
104. Department of Twin Research and Genetic Epidemiology, King's College London, London, SE1 7EH, UK.

bioRxiv preprint doi: https://doi.org/10.1101/144410; this version posted May 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

105. Institute of Clinical Medicine, Internal Medicine, University of Eastern Finland and Kuopio University Hospital, Kuopio, 70210, Finland.
106. MRC/BHF Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, CB1 8RN, UK.
107. Institute of Epidemiology II, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, 85764, Germany.
108. Welch Center for Prevention, Epidemiology, and Clinical Research, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA.
109. Department of Medicine, University of Leipzig, Leipzig, 04103, Germany.
110. Department of Public Health, Aarhus University, Aarhus, Denmark.
111. Danish Diabetes Academy, Odense, Denmark.
112. Robertson Centre for Biostatistics, University of Glasgow, Glasgow, UK.
113. Institut Pasteur de Lille, INSERM U1167, Université Lille Nord de France, Lille, F-59000, France.
114. Department of Epidemiology, German Institute of Human Nutrition Potsdam-Rehbruecke (DIfE), Nuthetal, 14558, Germany.
115. Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas, 77030, US.
116. Department of Public Health and Primary Care, University of Cambridge, Cambridge, CB1 8RN, UK.
117. Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland, 20892, USA.
118. Department of Nutrition and Dietetics, Harokopio University of Athens, Athens, 17671, Greece.
119. MRC-PHE Centre for Environment and Health, Imperial College London, London, W2 1PG, UK.
120. Princess Al-Jawhara Al-Brahim Centre of Excellence in Research of Hereditary Disorders (PACER-HD), King Abdulaziz University, Jeddah, 21589, Saudi Arabia.
121. Research Centre on Epidemiology and Preventive Medicine (EPIMED), Department of Medicine and Surgery, University of Insubria, Varese, 2100, Italy.
122. INSERM UMR 1027, Toulouse, 31000, France.

bioRxiv preprint doi: https://doi.org/10.1101/144410; this version posted May 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

123. Department of Cardiology, Toulouse University School of Medicine, Rangueil Hospital, Toulouse, 31059, France.
124. Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, 02114, USA.
125. Programs in Metabolism and Medical & Population Genetics, Broad Institute, Cambridge, MA, 02142, USA.
126. Department of Medicine, Harvard Medical School, Boston, Massachusetts, 02115, USA.
127. Center for Non-Communicable Diseases, Karachi, Pakistan.
128. Laboratory of Epidemiology and Population Sciences, National Institute on Aging, National Institutes of Health, Bethesda, MD, USA.
129. Department of Epidemiology, Cardiovascular Health Research Unit, University of Washington, Seattle, WA, 98195, USA.
130. Center for Genomic Medicine, Massachusetts General Hospital, USA.
131. Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA, USA.
132. UKCRC Centre of Excellence for Public Health (NI), Queens University of Belfast, Northern Ireland, BT7 1NN, UK.
133. Big Data Institute, Li Ka Shing Centre For Health Information and Discovery, University of Oxford, Oxford, OX37BN, UK.
134. National Institute for Health and Welfare, Helsinki, 00271, Finland.
135. Institute of Human Genetics, Technische Universität München, Munich, 81675, Germany.
136. Institute of Human Genetics, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, 85764, Germany.
137. Department of Genetics, University of North Carolina, Chapel Hill, North Carolina, 27599, USA.
138. Department of Epidemiology and Public Health, University of Strasbourg, Strasbourg, F-67085, France.
139. Department of Public Health, University Hospital of Strasbourg, Strasbourg, F-67081, France.
140. Clinical Research Centre, Centre for Molecular Medicine, Ninewells Hospital and Medical School, Dundee, DD1 9SY, UK.

bioRxiv preprint doi: https://doi.org/10.1101/144410; this version posted May 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

141. The Usher Institute to the Population Health Sciences and Informatics, University of Edinburgh, Edinburgh, EH16 4UX, UK.
142. Aberdeen Biomedical Imaging Centre, School of Medicine Medical Sciences and Nutrition, University of Aberdeen, Aberdeen, AB25 2ZD, UK.
143. Department of Clinical Sciences, Diabetes and Cardiovascular Disease, Genetic Epidemiology, Lund University, Malmö, 20502, Sweden.
144. Finnish Institute for Molecular Medicine (FIMM), University of Helsinki, Helsinki, Finland.
145. Department of Internal Medicine, Erasmus University Medical Center, Rotterdam, 3015CN, The Netherlands.
146. Department of Internal Medicine, Taichung Veterans General Hospital, Taichung Taiwan, National Yang-Ming University, School of Medicine, Taipei, Taiwan, National Defense Medical Center, School of Medicine, Taipei, Taiwan, Taichung, 40705, Taiwan.
147. McGill University and Génome Québec Innovation Centre, Montreal, Quebec, H3A 0G1, Canada.
148. Department of Human Genetics, McGill University, Montreal, Quebec, H3A 1B1, Canada.
149. Division of Endocrinology and Metabolism, Department of Medicine, McGill University, Montreal, Quebec, H3A 1A1, Canada.
150. Division of Population Health Sciences, Ninewells Hospital and Medical School, University of Dundee, Dundee, DD1 9SY, UK.
151. Institute of Medical Informatics, Biometry and Epidemiology, Chair of Genetic Epidemiology, Ludwig-Maximilians-Universität, Munich, 80802, Germany.
152. USC Roski Eye Institute, Department of Ophthalmology, Keck School of Medicine of the University of Southern California, Los Angeles, California, 90033, US.
153. Department of Internal Medicine, Division of Cardiovascular Medicine, University of Michigan, Ann Arbor, Michigan, 48109, USA.
154. Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan, 48109, USA.
155. Department of Human Genetics, University of Michigan, Ann Arbor, Michigan, 48109, USA.

bioRxiv preprint doi: https://doi.org/10.1101/144410; this version posted May 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

156. NIHR Blood and Transplant Research Unit in Donor Health and Genomics, Department of Public Health and Primary Care, University of Cambridge, Cambridge, CB1 8RN, UK.
157. Imperial College Healthcare NHS Trust, Imperial College London, London, W12 0HS, UK.
158. Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1RQ.
159. British Heart Foundation, Cambridge Centre of Excellence, Department of Medicine, University of Cambridge, Cambridge, CB2 0QQ, UK.
160. Department of Nutrition, Harvard School of Public Health, Boston, Massachusetts, 02115, USA.
161. Department of Genomics of Common Disease, School of Public Health, Imperial College London, London, W12 0NN, UK.
162. Clinical Cooparation Group Type 2 Diabetes, Helmholtz Zentrum München, Ludwig-Maximillians University Munich, Germany.
163. Clinical Cooparation Group Nutrigenomics and Type 2 Diabetes, Helmholtz Zentrum München, Technical University Munich, Germany.
164. Faculty of Health Sciences, University of Southern Denmark, Odense, 5000, Denmark.
165. University of Exeter Medical School, University of Exeter, Exeter, EX2 5DW, UK.
166. Department of Medicine, Division of Cardiovascular Medicine, Stanford University School of Medicine, Stanford, CA, 94305, US.
167. Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, Michigan, 48109, USA.
168. National Heart and Lung Institute, Cardiovascular Sciences, Hammersmith Campus, Imperial College London, London, W12 0NN, UK.
169. Department of Medical Sciences, Uppsala University, Uppsala, SE-751 85, Sweden.
170. Mindich Child Health and Development Institute, The Icahn School of Medicine at Mount Sinai, New York, NY, 10029, USA.
171. Department of Statistics, University of Oxford, Oxford, OX1 3TG, UK.
172. Department of Public Health and Primary Care, Leiden University Medical Center, Leiden, 2300 RC, The Netherlands.
173. Pat Macpherson Centre for Pharmacogenetics and Pharmacogenomics, Ninewells Hospital and Medical School, University of Dundee, Dundee, DD1 9SY, UK.

bioRxiv preprint doi: https://doi.org/10.1101/144410; this version posted May 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

174. Division of Epidemiology and Community Health, School of Public Health, University of Minnesota, Minneapolis, MN, 55454, US.
175. Cardiovascular Health Research Unit, Departments of Medicine, Epidemiology and Health Services, University of Washington, Seattle, WA, 98101-1448, USA.
176. Kaiser Permanent Washington Health Research Institute, Seattle, WA, 98101, USA.
177. Foundation for Research in Health, Exercise and Nutrition, Kuopio Research Institute of Exercise Medicine, Kuopio, Finland.
178. Institute of Cardiovascular and Medical Sciences, University of Glasgow, Glasgow, G12 8TA, UK.
179. Department of Hematology, School of Clinical Medicine, University of Cambridge, Cambridge, CB2 0AH.
180. Divisions of Endocrinology and Nephrology, University Hospital Leipzig, Liebigstgr. 18, Leipzig, 04103, Germany.
181. Research Programs Unit, Diabetes and Obesity, University of Helsinki, Helsinki, Finland.
182. Dasman Diabetes Institute, Dasman, 15462, Kuwait.
183. Department of Neuroscience and Preventive Medicine, Danube-University Krems, Krems, 3500, Austria.
184. Diabetes Research Group, King Abdulaziz University, Jeddah, 21589, Saudi Arabia.
185. Department of Physiology and Biophysics, University of Mississippi Medical Center, Jackson, Mississippi, 39216, USA.
186. Metabolic Research Laboratories, Institute of Metabolic Science, University of Cambridge, Cambridge, CB22 0QQ, UK.
187. Division of Endocrinology, Diabetes and Metabolism, Cedars-Sinai Medical Center, Los Angeles, CA, 90048.
188. General Medicine Division, Massachusetts General Hospital and Department of Medicine, Harvard Medical School, Boston, Massachusetts, 02114, USA.
189. Departments of Pediatrics and Medicine, The Institute for Translational Genomics and Population Sciences, LABioMed at Harbor-UCLA Medical Center, Torrance, California, 90502, US.

*These authors jointly directed this work.

bioRxiv preprint doi: https://doi.org/10.1101/144410; this version posted May 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Correspondence to:

Anubha Mahajan (anubha@well.ox.ac.uk)

Jerome I Rotter (jrotter@labiomed.org)

Mark I McCarthy (mark.mccarthy@drl.ox.ac.uk)

bioRxiv preprint doi: https://doi.org/10.1101/144410; this version posted May 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Identification of coding variant associations for complex diseases offers a direct route to biological insight, but is dependent on appropriate inference concerning the causal impact of those variants on disease risk. We aggregated exome-array and exome sequencing data for 81,412 type 2 diabetes (T2D) cases and 370,832 controls of diverse ancestry, identifying 40 distinct coding variant association signals (at 38 loci) reaching significance ($p<2.2\times10^{-7}$). Of these, 16 represent novel associations mapping outside known genome-wide association study (GWAS) signals. We make two important observations. First, despite a threefold increase in sample size over previous efforts, only five of the 40 signals are driven by variants with minor allele frequency <5%, and we find no evidence for low-frequency variants with allelic odds ratio >1.36. Second, we used GWAS data from 50,160 T2D cases and 465,272 controls to fine-map associated coding variants in their regional context, with and without additional weighting, to account for the global enrichment of complex trait association signals in coding exons. We demonstrate convincing support (posterior probability >80% under the "annotation-weighted" model) that coding variants are causal for the association at 16 of the 40 signals (including novel signals involving *POC5* p.His36Arg, *ANKH* p.Arg187Gln, *WSCD2* p.Thr113Ile, *PLCB3* p.Ser778Leu, and *PNPLA3* p.Ile148Met). However, one third of coding variant association signals represent "false leads" at which naïve analysis would have led to an erroneous inference regarding the effector transcript mediating the signal. Accurate identification of validated targets is dependent on correct specification of the contribution of coding and non-coding mediated mechanisms at associated loci.

Genome-wide association studies (GWAS) have identified many thousands of association signals influencing common, complex traits such as type 2 diabetes (T2D) and obesity[1-7]. Most of these significant association signals involve common variants that map to non-coding sequence and identification of their cognate effector transcripts has often proved challenging. The identification of coding variants causally implicated in trait predisposition offers a more direct route from association signal to biological inference.

The exome occupies only 1.5% of overall genome sequence, but modelling of complex trait architecture indicates that, for many common diseases, coding variants make a disproportionately large contribution to trait heritability[8,9]. This enrichment indicates that coding variant association signals have an enhanced probability of being causal when

bioRxiv preprint doi: https://doi.org/10.1101/144410; this version posted May 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

compared to those involving an otherwise equivalent non-coding variant. This does not, however, guarantee that all coding variant associations are causal. Alleles driving common-variant (minor allele frequency [MAF] ≥5%) GWAS signals typically reside on extended risk haplotypes that, due to linkage disequilibrium (LD), incorporate many common variants[10,11]. Consequently, the presence of a coding allele on the risk haplotype does not constitute sufficient evidence that it represents the causal variant at the locus, or that the gene within which it lies is mediating the association signal. Since much coding variant discovery has proceeded through exome-specific analyses via exome-array genotyping or exome sequencing, researchers have often been poorly-placed to position coding variant associations in the context of regional genetic variation, and it is unclear how often this may lead to incorrect assumptions regarding their causal role.

In our recent study of T2D predisposition[12], we surveyed the exomes of 34,809 T2D cases and 57,985 controls, of predominantly (>90%) European descent, and identified 13 distinct coding variant associations reaching genome-wide significance. Twelve of these associations involved common variants, but the data hinted at a substantial pool of near-significant lower-frequency coding variants of moderate impact (allelic odds ratio [OR] between 1.5 and 3.0) that might be amenable to detection in larger samples. We also reported that, whilst many of these signals fell within common variant loci previously identified by GWAS, it was often far from trivial to determine, using available data, whether those coding variants were causal or simply 'hitchhiking' on risk haplotypes.

Here, we report analyses that address these two key issues. First, we extended the scope of our exome-array genotyping study to include data from 81,412 T2D cases and 370,832 controls of diverse ancestry, substantially expanding our power to detect coding variant associations across the allele-frequency spectrum. Second, we undertook high-resolution fine-mapping of the detected signals in 50,160 T2D cases and 465,272 controls with genome-wide genotyping data, to understand the extent to which the identification of coding variant associations provides a reliable guide to causal mechanisms.

RESULTS

Study overview. We aggregated T2D association summary statistics from 54 studies in up to 452,244 individuals (effective sample size 228,825) across five ancestry groups

bioRxiv preprint doi: https://doi.org/10.1101/144410; this version posted May 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(**Supplementary Tables 1 and 2**): African American, East Asian, European, Hispanic/Latino, and South Asian. These included: (a) 58,425 cases and 188,032 controls genotyped with the exome array; (b) 14,608 cases and 174,322 controls from UK Biobank and GERA (Resource for Genetic Epidemiology on Adult Health and Aging) genotyped with GWAS arrays enriched for exome content and/or coverage of low-frequency variation across ethnic groups[13,14]; and (c) 8,379 cases and 8,478 controls with whole-exome sequence from the GoT2D/T2D-GENES[12] and SIGMA[15] studies. Overall, this represented a 3-fold increase in effective sample size over our previous study of T2D predisposition[12]. We performed European-specific (EUR, 60.9% of total effective sample size) and trans-ethnic (TE) meta-analyses of variants delineated by exome-array content (247,470 variants), with and without adjustment for body mass index (adjBMI).

We considered $p<2.2\text{x}10^{-7}$ as significant for protein truncating variants (PTVs) and moderate impact coding variants (including missense, in-frame indel and splice region variants) based on a weighted Bonferroni correction that accounts for the observed enrichment in complex trait association signals mapping to coding variation[16]. This threshold is close to that obtained through other approaches such as simple Bonferroni correction for the total number of coding variants on the array (**Methods**). Compared to our previous study[12], the expanded sample size substantially increased power to detect association for common variants of modest effect (e.g. from 14.4% to 97.9% for a variant with 20% MAF and OR=1.05) and lower-frequency variants with larger effects (e.g. from 11.8% to 97.5% for a variant with 1% MAF and OR=1.20) assuming homogenous allelic effects across ancestry groups (**Methods**).

Insights into coding variant association signals underlying T2D susceptibility. We detected significant associations at 69 coding variants under an additive genetic model, mapping to 38 loci (**Supplementary Fig. 1, Supplementary Table 3**). Of these, 52 (at 29 loci) were significant in the European-specific analysis, and 62 (at 35 loci) in the trans-ethnic analysis. Variants at *PLCB3, C17orf58,* and *ZHX3* were only significant in the European-specific analysis. We observed minimal evidence of heterogeneity in allelic OR between ancestry groups (**Supplementary Table 3**), and no compelling evidence for non-additive allelic effects, irrespective of allele frequency (**Supplementary Fig. 2, Supplementary Table 4**). Reciprocal conditional analyses (**Methods**) indicated that the 69 coding variants represented

bioRxiv preprint doi: https://doi.org/10.1101/144410; this version posted May 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

40 distinct association signals (conditional $p<2.2\times10^{-7}$) across the 38 loci, with two distinct signals each at *HNF1A* and *RREB1* (**Supplementary Table 5**). These 40 signals included the 13 associations reported in our earlier publication[12], all of which demonstrated more significant associations in this expanded trans-ethnic meta-analysis (**Supplementary Table 6**).

Sixteen of the 40 distinct association signals mapped outside regions previously implicated in T2D susceptibility, defined as >500kb from the reported lead GWAS SNPs (**Table 1**). These included signals involving missense variants in *POC5* (p.His36Arg, rs2307111, $p_{TE}=1.6\times10^{-15}$), *PNPLA3* (p.Ile148Met, rs738409, $p_{TEadjBMI}=2.8\times10^{-11}$), and *ZZEF1* (p.Ile2014Val, rs781831, $p_{TE}=8.3\times10^{-11}$).

Contribution of low-frequency and rare coding variation to T2D susceptibility. Despite increased power and good coverage of low-frequency variants on the exome array (>80% of coding variants with MAF >0.5% in European ancestry populations[12]), all but five of the 40 distinct coding variant association signals were common, with modest effects (allelic OR 1.02-1.36) (**Supplementary Fig. 3, Supplementary Table 3**). The five association signals attributable to lower-frequency variants were also of modest effect (allelic OR 1.09-1.29) (**Supplementary Fig. 3**). Two of these lower-frequency variant signals represented novel protective associations against T2D: *FAM63A* p.Tyr95Asn (rs140386498, MAF=1.2%, OR= 0.82 [0.77-0.88], $p_{EUR}=5.8\times10^{-8}$) and *ANKH* p.Arg187Gln (rs146886108, MAF=0.4%, OR=0.78 [0.69-87], $p_{EUR}=2.0\times10^{-7}$). Both of these variants were very rare or monomorphic in non-European individuals analysed in this study.

In Fuchsberger et al.[12], we observed a pool of 100 low-frequency coding variants with modest effects (estimated allelic ORs between 1.10 and 2.66) for which the association evidence was strong but not genome-wide significant. In this expanded analysis, only five of these variants, including the two novel associations at *FAM63A* p.Tyr95Asn and *ANKH* p.Arg187Gln, achieved significance. More precise effect size estimation with the larger sample size indicates that the OR estimates in the earlier study were subject to a substantial upwards bias (**Supplementary Fig. 3**).

To detect additional rare variant association signals, we performed gene-based analyses (burden and SKAT[17]) using previously-defined "strict" and "broad" masks, filtered for annotation and MAF[18] (**Methods**). We identified gene-based associations with T2D susceptibility ($p<2.5\times10^{-6}$, Bonferroni correction for 20,000 genes) for *FAM63A* (SKAT broad

bioRxiv preprint doi: https://doi.org/10.1101/144410; this version posted May 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

mask, 10 variants, combined MAF=1.90%, p_{EUR}=3.1x10^{-9}) and *PAM* (SKAT broad mask, 17 variants, combined MAF=4.67%, p_{TE}=8.2x10^{-9}). On conditional analysis (**Supplementary Table 7**), we found that the gene-based signal at *FAM63A* was entirely accounted for by the low-frequency p.Tyr95Asn allele described earlier (SKAT broad mask, conditional *p*=0.26). In these data, the gene-based signal for *PAM* is also attributable to a single low-frequency variant (p.Asp563Gly; SKAT broad mask, conditional *p*=0.15). However, a second, previously-described, low-frequency variant *PAM* p.Ser539Trp[19] is not represented on the exome array, and thus did not contribute to our analyses.

Fine-mapping of coding variant association signals with T2D susceptibility. The present study has identified 40 distinct coding variant associations with T2D, but this information is not sufficient to determine that the variants themselves are causal for the disease. To assess the role of these coding variants in the context of regional genetic variation at the locus, we fine-mapped the distinct association signals using a European ancestry GWAS meta-analysis including 50,160 T2D cases and 465,272 controls, aggregated from 24 studies by the DIAGRAM Consortium. Each component GWAS had been imputed using suitable high density reference panels (**Methods, Supplementary Table 8**): (i) 22 GWAS were imputed up to the Haplotype Reference Consortium[20]; (ii) the UK Biobank GWAS was imputed to a merged reference panel from the 1000 Genomes Project (multi-ethnic, phase 3, October 2014 release)[21] and the UK10K Project[9]; and (iii) the deCODE GWAS was imputed up to the deCODE Icelandic population-specific reference panel based on whole-genome sequence data[19] (**Methods, Supplementary Table 8**). Distinct association signals were delineated before fine-mapping using approximate conditional analyses (**Methods, Supplementary Table 5**). We excluded the locus in the major histocompatibility complex because of the extended and complex structure of LD across the region, which complicates fine-mapping efforts.

For each of the remaining 39 signals, we first constructed "functionally unweighted" credible variant sets which, at each locus, collectively account for 99% of the posterior probability (π_U) of driving the association, based exclusively on the meta-analysis summary statistics[22] (**Methods, Supplementary Table 9**). For each signal, we then calculated the total posterior probability attributed to coding variants (missense, in-frame indel, and splice region variants; **Figure 1, Supplementary Fig. 4 and 5**). Under this model, there were only

bioRxiv preprint doi: https://doi.org/10.1101/144410; this version posted May 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

two signals at which coding variants accounted for ≥80% of the posterior probability of association: *HNF4A* p.Thr139Ile (rs1800961, π_U>0.999) and *RREB1* p. Asp1171Asn (rs9379084, π_U=0.920). However, at other signals, including some, such as *GCKR* p.Pro446Leu and *SLC30A8* p.Arg276Trp, where robust empirical evidence has established their causal role[23,24], the genetic evidence supporting coding variant causation was weak. This is because coding variants were typically in high LD (r^2>0.9) with large numbers of non-coding variants, such that the posterior probabilities of association were distributed across many variants with broadly equivalent evidence for association.

These functionally unweighted sets are based on genetic fine-mapping data alone, and do not account for the knowledge that coding variants are disproportionately represented amongst GWAS associations with complex traits[8,9]. To accommodate this knowledge, we extended these fine-mapping analyses by incorporating an "annotation informed" prior model of causality. We derived the priors from estimates of enrichment of association signals by sequence annotation from an analysis conducted by deCODE across 96 quantitative and 123 binary phenotypes[16] (**Methods**). This model "boosts" the priors and hence the posterior probability (π_A) of coding variants. It also takes some account (in a tissue-non-specific way) of the GWAS enrichment of variants within enhancer elements (as assayed through DNase I hypersensitivity) as compared to non-coding variants mapping elsewhere. As expected, the annotation informed prior model generated smaller 99% credible sets across most signals, corresponding to fine-mapping at higher resolution (**Supplementary Table 9**).

As expected, the estimated contribution of coding variants was increased under the annotation informed model. At one signal, the East Asian specific *PAX4* p.Arg190His (rs2233580), the available fine-mapping data did not allow us to draw comprehensive conclusions on the contribution of coding variation to T2D susceptibility since the variant was not present in European GWAS. At the remaining 38 association signals, we could distinguish three broad patterns of causal relationships between coding variants and T2D risk.

Group 1: T2D association signal is driven by coding variants. At 16 of the 38 distinct signals (after excluding those at MHC and *PAX4*), coding variation accounted for >80% of the posterior probability of the association signal under the annotation informed model (**Figure**

bioRxiv preprint doi: https://doi.org/10.1101/144410; this version posted May 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

1, Table 2, Supplementary Table 9). This posterior probability was accounted for by a single coding variant at 12 signals and multiple coding variants at four. Reassuringly, group 1 signals provided confirmation of coding variant causation for several loci at which functional studies (involving manipulation of the variant and/or effector gene) have reinforced genetic association data, establishing the role of *GCKR*, *PAM*, *SLC30A8*, and three variants in strong LD with each other at the *KCNJ11-ABCC8* locus (**Table 2**). T2D association signals at the 12 remaining signals (**Fig. 1, Supplementary Table 9**) had not previously been established to be driven by coding variation, but our fine-mapping analyses pointed to high probability causal coding variants after incorporating the annotation informed priors: these included *HNF4A*, *RREB1* p. Asp1171Asn, *ANKH*, *WSCD2*, *POC5*, *TM6SF2*, *HNF1A* p.Ala146Val, *GIPR*, *HNF1A* p.Ile75Leu, *LPL*, *PLCB3*, and *PNPLA3* (**Table 2**). At several of these loci, independent evidence corroborates the causal role of the genes harbouring the associated coding variants with respect to T2D-risk. For example, rare coding mutations at *HNF1A* and *HNF4A* are causal for monogenic, early-onset forms of diabetes[25]; and at *TM6SF2* and *PNPLA3,* the coding variant concerned has been directly implicated in the development of non-alcoholic fatty liver disease (NAFLD)[26,27].

The use of priors capturing the enrichment of coding variants seems a reasonable model, genome-wide. However, at any given locus, strong priors (especially for PTVs) might elevate to apparent causality, variants, which, on the basis of genetic fine-mapping alone, would have been excluded from a causal role. However, comparison of the annotation informed and functionally unweighted credible sets, indicated that this was not the case for any of the 16 association signals in group 1. For 11 of the 16 (*GCKR*, *PAM*, *KCNJ11-ABCC8*, *HNF4A*, *RREB1* p. Asp1171Asn, *ANKH*, *POC5*, *TM6SF2*, *HNF1A* p.Ala146Val, *PLCB3*, *PNPLA3*) the coding variant was the lead SNP in the fine-mapping analysis (**Table 2**), with the highest posterior probability of association, even under the functionally unweighted model. At *SLC30A8, WSCD2*, and *GIPR*, the coding variants had similar posterior probabilities of association as the lead non-coding SNPs under the functionally unweighted prior: *SLC30A8* p.Arg276Trp (rs13266634, π_U=0.295) and (rs35859536, π_U=0.388); *WSCD2* p.Thr113Ile (rs3764002, π_U=0.281) and rs1426371 (π_U=0.475); and *GIPR* p.Glu138Gln (rs1800437, π_U=0.169) and rs10423928 (π_U=0.221). At these 14 signals therefore, the coding variants have either greater or equivalent posterior probabilities of association as the best flanking

bioRxiv preprint doi: https://doi.org/10.1101/144410; this version posted May 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

non-coding SNPs under the functionally unweighted model, but receive a boost in the posterior probability after accounting for variant annotation.

The situation is less clear at the *LPL* locus. Here, fine-mapping resolution is poor under the functionally unweighted prior, and the coding variant resides on an extended haplotype in strong LD with non-coding variants, some with higher posterior probabilities, such as rs74855321 (π_U=0.0481) (compared to *LPL* p.Ser474* [rs328, π_U=0.0231]). However, because *LPL* p.Ser474* is annotated as a PTV, it benefits from a substantially increased prior that is reflected in the annotation informed ranking. Ultimately, a decision on the causal role of any such variant must rest on the amalgamation of evidence from diverse sources including detailed functional evaluation of the coding variants, and of other variants with which it is in LD.

At the *HNF1A* p.Ile75Leu signal, the total posterior probability attributed to coding variants under the annotation informed prior was 0.894. However, the total probability was accounted for by p.Gly226Ala (rs56348580, π_A=0.894), a variant which is missing from the exome array. Conversely, the posterior probability attributed to the index coding variant, p.Ile75Leu (rs1169288) was <0.001, although this may reflect its absence from most commercial GWAS arrays and low-quality imputation. Fine-mapping analyses conducted using the Metabochip[10], on which *HNF1A* p.Ile75Leu, as well as many local non-coding variants, are directly typed, demonstrates that these two coding variants are likely to be driving distinct association signals at this locus, and are consistent with our observations from the exome array. Direct genotyping or sequencing will be required to fully disentangle the relationships between the various coding variants at this signal. However, the established role of rare coding variants in *HNF1A* with respect to monogenic forms of diabetes leaves the role of *HNF1A* as a causal transcript at this locus in little doubt.

Group2: T2D association signals are not attributable to coding variants. At 14 of the 38 distinct signals, coding variation accounted for <20% of the posterior probability of driving the association, even after applying the annotation informed prior model that boosts coding variant posterior probabilities. These signals are likely to be driven by local non-coding variation and mediated through regulation of gene expression. Six of these signals (*TPCN2*, *ZHX3*, *MLX*, *ZZEF1*, C17orf58, and *CEP68)* represent novel T2D-association signals identified in the exome-focused analysis. On the basis of the exome-array discoveries, it would have

bioRxiv preprint doi: https://doi.org/10.1101/144410; this version posted May 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

been natural to consider the named genes at these, and the other loci in this group, as strong candidates for mediation of their respective association signals. However, the fine-mapping analyses suggest that these coding variant signals are irrelevant to biological inference.

The coding variant association at the *CENTD2* (*ARAP1*) locus is a case-in-point. The association with the p.Gln802Glu variant in *ARAP1* (rs56200889, p_{TE}=4.8x10^{-8} but π_A<0.001 in the annotation informed analysis) is clearly seen in the fine-mapping analysis to be secondary to a substantially stronger non-coding association signal involving a cluster of variants including rs11603334 (p_{TE}=9.5x10^{-18}, π_A=0.0692) and rs1552224 (p_{TE}=2.5x10^{-17}, π_A=0.0941). The identity of the effector transcript at this locus has been the subject of considerable investigation, and some early studies used islet expression data to highlight *ARAP1* as the strongest candidate[28]. However, a more recent study integrating studies of human islet genomics and murine gene knockouts points firmly towards *STARD10* as the gene mediating the GWAS signal, consistent with the reassignment of the *ARAP1* coding variant association as irrelevant to biological inference at this locus[29].

Group 3: Fine-mapping data consistent with partial role for coding variants. At eight of the 38 distinct signals, the total posterior probability attributable to coding variation in the annotation informed analyses was between 20% and 80%. At these signals, the evidence is consistent with "partial" contributions from coding variants, although the precise inference is likely to be locus-specific, dependent on subtle variations in LD, imputation accuracy, and the extent to which the global priors accurately represent the functional impact of the specific variants concerned.

This group includes *PPARG* at which independent evidence corroborates the causal role of this specific effector transcript with respect to T2D risk. *PPARG* encodes the target of antidiabetic thiazolidinedione drugs and is known to harbour rare coding variants that are causal for lipodystrophy and insulin resistance, both conditions highly relevant to T2D. The common variant association signal at this locus has generally been attributed to the p.Pro12Ala coding variant (rs1801282) although confirmation that this variant has an empirical impact on PPARG function has been difficult to obtain[30-32]. In the functionally unweighted analysis, p.Pro12Ala had an unimpressive posterior probability of being causal (π_U=0.0238); after including annotation informed priors, the same variant emerged with the

bioRxiv preprint doi: https://doi.org/10.1101/144410; this version posted May 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

highest posterior probability (π_A=0.410), although the 99% credible set included 19 non-coding variants, spanning 67kb (**Supplementary Table 9**). These credible set variants included rs4684847 (π_A=0.00891), at which the T2D-associated allele has been reported to impact *PPARG2* expression and insulin sensitivity by altering binding of the homeobox transcription factor PRRX1[33]. These data are consistent with a model whereby regulatory variants contribute to the mechanisms through which the T2D GWAS signal impacts PPARG activity (in combination with or, potentially, to the exclusion of p.Pro12Ala). Future improvements in functional annotation for regulatory variants (gathered from relevant tissues and cell types) can be expected to provide increasingly granular priors that can be used to fine-tune assignment of causality at loci such as this.

Functional impact of coding alleles. In other contexts, the functional impact of coding alleles is correlated with: (i) variant-specific features, including measures of conservation and predicted impact on protein structure; and (ii) gene-specific features such as extreme selective constraints as quantified by the intolerance to functional variation[34]. To determine whether similar measures could capture information pertinent to T2D causation, we compared coding variants falling into the different fine-mapping groups for a variety of measures including MAF, Combined Annotation Dependent Depletion score (CADD-score)[35], and loss-of-function (LoF)-intolerance metric, pLI[34] (**Methods**, **Fig. 2**). As noted previously[35], CADD-score and MAF exhibit a negative correlation (Pearson's correlation r=-0.44, *p*=0.0033) across association signals. Variants from group 1 had significantly higher CADD-scores than those in group 2 (*p*=0.0017, by Kolmogorov-Smirnov test). With the exception of the variants at *KCNJ11-ABCC8* and *GCKR*, all group 1 coding variants considered likely to be driving T2D association signals have CADD-score ≥20 (i.e. the 1% most deleterious variants predicted across the human genome). On this basis, we would predict that the coding variant at *PAX4*, for which the fine-mapping data were inconclusive, is also likely causal for T2D.

Novel non-coding association signals for T2D susceptibility. Whilst the exome array primarily encompasses variation that alters protein function, it also incorporates previously reported non-coding lead SNPs from GWAS for a range of complex human phenotypes, including metabolic traits that may also impact T2D susceptibility. We detected novel

bioRxiv preprint doi: https://doi.org/10.1101/144410; this version posted May 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

significant ($p<5x10^{-8}$) associations with T2D status for 20 non-coding variants at 15 loci. Three of these (*POC5*, *LPL*, and *BPTF*) overlap with novel coding signals reported here (**Supplementary Table 10**). Lead SNPs at these loci had been included on the exome array as GWAS tags for association signals with metabolic traits including central and overall obesity, lipid levels, coronary heart disease, venous thromboembolism, and menarche, but here were demonstrated to be genome-wide significant for T2D as well (**Supplementary Table 10**). If instead of the conventional genome-wide significance threshold, we adopted a weighted Bonferroni threshold for all non-coding variants of $p<9.5x10^{-9}$, balancing the less stringent threshold, used for defining significance for coding variants, 10 of the 15 loci remained associated with T2D.

T2D loci and physiological classification. The development of T2D involves dysfunction of multiple mechanisms. Systematic analyses of the physiological effects of known T2D risk variants have provided improved understanding of the key intermediate processes involved in the disease and the mechanisms through which many of those variants exert their primary impact on disease risk[36]. We obtained association summary statistics for a range of metabolic traits and other outcomes for 94 T2D-associated index variants representing the 40 distinct coding signals and 54 distinct non-coding signals (including 12 novel and 42 previously reported non-coding GWAS lead SNPs from the exome array). We applied hierarchical clustering techniques (**Methods**) to generate, multi-trait association patterns, and were able to allocate 71 of the 94 loci to one of three categories (**Supplementary Fig. 6, Supplementary Table 11**). The first category, comprising of nine T2D-risk loci with strong BMI and dyslipidemia phenotypes, included three of the novel coding signals: *PNPLA3*, which showed strong association with dyslipidemia, and *POC5* and *BPTF* where the risk allele was associated with increased BMI (**Supplementary Fig. 6, Supplementary Table 11**). The T2D associations at both *POC5* and *BPTF* were substantially attenuated (at least by 2-fold decrease in $-\log_{10}p$) after adjusting for BMI (**Table 1, Supplementary Table 3, Supplementary Fig. 7**), indicating that the impact on T2D risk is likely mediated by a primary effect through increased adiposity. *PNPLA3* and *POC5* are established NAFLD[26] and BMI[6] loci, respectively. The second category included 39 loci at which the multi-trait profiles indicated a primary effect on insulin secretion. This category included four of the novel coding variant signals (*ANKH*, *ZZEF1*, *TTLL6*, and *ZHX3*). The third category encompassed 23

bioRxiv preprint doi: https://doi.org/10.1101/144410; this version posted May 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

loci with primary effects on insulin action including coding variant associations at *KIF9*, *PLCB3*, *CEP68*, *TPCN2*, *FAM63A*, and *PIM3*. For most variants in this category, the T2D-risk allele was associated with lower BMI, and T2D association signals were more pronounced after adjustment for BMI. At a subset of these loci, including *KIF9* and *PLCB3*, T2D-risk alleles were associated with markedly larger waist-hip ratio and lower body fat percentage, indicating that their mechanism of action likely reflects limitations in storage capacity of peripheral adipose tissue[37].

DISCUSSION

The present study adds to mounting evidence constraining the contribution of lower-frequency variants to T2D risk. Although the exome array interrogates only a subset of the universe of coding variants, it does capture the majority (~80%) of low-frequency (MAF>0.5%) variants in European populations (and 48-75% in other major ancestral groups)[12]. The substantial increase in sample size in the present study provides more robust enumeration of the effect size distribution in this low-frequency variant range, and indicates that previous analyses are likely to have, if anything, overestimated the contribution of low-frequency variants to T2D-risk.

The present study is less informative regarding rare variants. These are sparsely captured on the exome array. In addition, the combination of greater regional diversity in rare allele distribution and the enormous sample sizes required to detect many rare variant associations (which would require meta-analysis of data from multiple diverse populations) acts against their detection. We note that our complementary genome and exome sequence analyses have thus far failed to register strong evidence for a substantial rare variant component to T2D-risk[12].

Once a coding variant association is detected, it is natural to assume a causal connection between that variant, the gene in which it sits, and the disease or phenotype of interest. Whilst such assignments may be robust for many protein-truncating alleles, we demonstrate that this implicit assumption is often inaccurate, particularly for associations attributable to common, missense variants. A third of the coding variant associations we detected were, when assessed in terms of regional LD, highly unlikely to be causal. At these loci, the genes within which they reside are consequently deprived of their implied

bioRxiv preprint doi: https://doi.org/10.1101/144410; this version posted May 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

connection to disease risk, and attention redirected towards nearby non-coding variants and their impact on regional gene expression. As a group, coding variants we assign as causal are predicted to have a more deleterious impact on gene function than those that we exonerate, but, as in other settings, coding annotation methods lack both sensitivity and specificity. Besides, it is crucial to realise that empirical evidence that the associated coding allele is "functional" (i.e. that it can be shown to influence function of its cognate gene in some experimental assay) provides limited reassurance that the coding variant is responsible for the T2D association, unless that specific perturbation of gene function can itself be plausibly linked to the disease phenotype.

Our fine-mapping analyses make use of the observation that coding variants are globally enriched across GWAS signals[8,9,16] with greater prior probability of causality assigned to those with more severe impact on biological function. We assigned non-coding variants equivalently diminished priors, with lowest support for those mapping outside of DHS. The extent to which our findings corroborate previous assignments of causality (often backed up by detailed, disease appropriate functional assessment and orthogonal causal evidence) suggests that even these sparse annotations provide valuable information to guide target validation efforts. Nevertheless, we recognise that there are inevitable limits to the extrapolation of these broad-brush genome-wide enrichments to individual loci, and expect that improvements in functional annotation for both coding and regulatory variants, particularly when these are gathered from trait-relevant tissues and cell types, will provide more granular, trait-specific priors to fine-tune assignment of causality within associated regions. These will motivate target validation efforts that benefit from synthesis of both coding and regulatory routes to gene perturbation.

The term "smoking gun" has often been used to describe the potential of functional coding variants to provide causal inference with respect to pathogenetic mechanisms[38]: our study provides a timely reminder that, even when a suspect with a smoking gun is found at the scene of a crime, it should not automatically be assumed that they fired the fatal bullet.

bioRxiv preprint doi: https://doi.org/10.1101/144410; this version posted May 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

REFERENCES

1. Kooner, J.S. *et al.* Genome-wide association study in individuals of South Asian ancestry identifies six new type 2 diabetes susceptibility loci. *Nat Genet* **43**, 984-9 (2011).
2. Cho, Y.S. *et al.* Meta-analysis of genome-wide association studies identifies eight new loci for type 2 diabetes in east Asians. *Nat Genet* **44**, 67-72 (2011).
3. Morris, A.P. *et al.* Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat Genet* **44**, 981-90 (2012).
4. Mahajan, A. *et al.* Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nat Genet* **46**, 234-44 (2014).
5. Ng, M.C. *et al.* Meta-analysis of genome-wide association studies in African Americans provides insights into the genetic architecture of type 2 diabetes. *PLoS Genet* **10**, e1004517 (2014).
6. Locke, A.E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197-206 (2015).
7. Shungin, D. *et al.* New genetic loci link adipose and insulin biology to body fat distribution. *Nature* **518**, 187-96 (2015).
8. Gusev, A. *et al.* Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am J Hum Genet* **95**, 535-52 (2014).
9. Walter, K. *et al.* The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82-90 (2015).
10. Gaulton, K.J. *et al.* Genetic fine mapping and genomic annotation defines causal mechanisms at type 2 diabetes susceptibility loci. *Nat Genet* **47**, 1415-25 (2015).
11. Horikoshi, M. *et al.* Transancestral fine-mapping of four type 2 diabetes susceptibility loci highlights potential causal regulatory mechanisms. *Hum Mol Genet* **25**, 2070-2081 (2016).
12. Fuchsberger, C. *et al.* The genetic architecture of type 2 diabetes. *Nature* **536**, 41-7 (2016).

bioRxiv preprint doi: https://doi.org/10.1101/144410; this version posted May 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

13. 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).

14. Cook, J.P. & Morris, A.P. Multi-ethnic genome-wide association study identifies novel locus for type 2 diabetes susceptibility. *Eur J Hum Genet* **24**, 1175-80 (2016).

15. Estrada, K. *et al.* Association of a low-frequency variant in HNF1A with type 2 diabetes in a Latino population. *JAMA* **311**, 2305-14 (2014).

16. Sveinbjornsson, G. *et al.* Weighting sequence variants based on their annotation increases power of whole-genome association studies. *Nat Genet* **48**, 314-7 (2016).

17. Liu, D.J. *et al.* Meta-analysis of gene-level tests for rare variant association. *Nat Genet* **46**, 200-4 (2014).

18. Purcell, S.M. *et al.* A polygenic burden of rare disruptive mutations in schizophrenia. *Nature* **506**, 185-90 (2014).

19. Steinthorsdottir, V. *et al.* Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes. *Nat Genet* **46**, 294-8 (2014).

20. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* **48**, 1279-83 (2016).

21. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).

22. Maller, J.B. *et al.* Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat Genet* **44**, 1294-301 (2012).

23. Beer, N.L. *et al.* The P446L variant in GCKR associated with fasting plasma glucose and triglyceride levels exerts its effect through increased glucokinase activity in liver. *Hum Mol Genet* **18**, 4081-8 (2009).

24. Flannick, J. *et al.* Loss-of-function mutations in SLC30A8 protect against type 2 diabetes. *Nat Genet* **46**, 357-63 (2014).

25. Murphy, R., Ellard, S. & Hattersley, A.T. Clinical implications of a molecular genetic classification of monogenic beta-cell diabetes. *Nat Clin Pract Endocrinol Metab* **4**, 200-13 (2008).

26. Romeo, S. *et al.* Genetic variation in PNPLA3 confers susceptibility to nonalcoholic fatty liver disease. *Nat Genet* **40**, 1461-5 (2008).

bioRxiv preprint doi: https://doi.org/10.1101/144410; this version posted May 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

27. Kozlitina, J. *et al.* Exome-wide association study identifies a TM6SF2 variant that confers susceptibility to nonalcoholic fatty liver disease. *Nat Genet* **46**, 352-6 (2014).

28. Kulzer, J.R. *et al.* A common functional regulatory variant at a type 2 diabetes locus upregulates ARAP1 expression in the pancreatic beta cell. *Am J Hum Genet* **94**, 186-97 (2014).

29. Carrat, G.R. *et al.* Decreased STARD10 expression is associated with defective insulin secretion in humans and mice. *Am J Hum Genet* **100**, 238-256 (2017).

30. Deeb, S.S. *et al.* A Pro12Ala substitution in PPARgamma2 associated with decreased receptor activity, lower body mass index and improved insulin sensitivity. *Nat Genet* **20**, 284-7 (1998).

31. Majithia, A.R. *et al.* Rare variants in PPARG with decreased activity in adipocyte differentiation are associated with increased risk of type 2 diabetes. *Proc Natl Acad Sci U S A* **111**, 13127-32 (2014).

32. Majithia, A.R. *et al.* Prospective functional classification of all possible missense variants in PPARG. *Nat Genet* **48**, 1570-1575 (2016).

33. Claussnitzer, M. *et al.* Leveraging cross-species transcription factor binding site patterns: from diabetes risk loci to disease mechanisms. *Cell* **156**, 343-58 (2014).

34. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285-91 (2016).

35. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310-5 (2014).

36. Dimas, A.S. *et al.* Impact of type 2 diabetes susceptibility variants on quantitative glycemic traits reveals mechanistic heterogeneity. *Diabetes* **63**, 2158-71 (2014).

37. Lotta, L.A. *et al.* Integrative genomic analysis implicates limited peripheral adipose storage capacity in the pathogenesis of human insulin resistance. *Nat Genet* **49**, 17-26 (2017).

38. Altshuler, D. & Daly, M. Guilt beyond a reasonable doubt. *Nat Genet* **39**, 813-5 (2007).

bioRxiv preprint doi: https://doi.org/10.1101/144410; this version posted May 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

FIGURE LEGENDS

Figure 1 | Posterior probabilities for coding variants across loci with annotation-informed priors. Fine-mapping of 38 distinct association signals was performed using European ancestry GWAS meta-analysis including 50,160 T2D cases and 465,272 controls. For each signal, we constructed a credible set of variants accounting for 99% of the posterior probability of driving the association, incorporating an "annotation informed" prior model of causality which "boosts" the posterior probability of driving the association signal that is attributed to coding variants. Each bar here represents a signal with the total probability attributed to the coding variants within the 99% credible set plotted on the y axis. When the probability (bar) is split across multiple coding variants (at least 0.05 probability attributed to a variant) at a particular locus, these are indicated by blue, pink, yellow, and green colours. The combined probability of the remaining coding variants are highlighted in grey. *RREB1*(a): *RREB1* p. Asp1171Asn; *RREB1*(b): *RREB1* p.Ser1499Tyr; *HNF1A*(a): *HNF1A* p.Ala146Val; *HNF1A*(b): *HNF1A* p.Ile75Leu; *PPIP5K2*† : *PPIP5K2* p.Ser1207Gly; *MTMR3*†: *MTMR3* p.Asn960Ser; *IL17REL*†: *IL17REL* p.Gly70Arg; *NBEAL2*†: *NBEAL2* p.Arg511Gly, *KIF9*†: *KIF9* p.Arg638Trp.

Figure 2 | Plot of measures of variant-specific and gene-specific features of distinct coding signals to access the functional impact of coding alleles. Each point represents a coding variant with the minor allele frequency (MAF) plotted on the x axis and the Combined Annotation Dependent Depletion score (CADD-score) plotted on the y axis. Size of each point varies with the measure of intolerance of the gene to loss of function variants (pLI) and the colour represents the fine-mapping group each variant is assigned to. Group 1: Signal is driven by coding variant. Group 2: Signal attributable to non-coding variants. Group 3: Consistent with partial role for coding variants. Unclassified category includes *PAX4* and signal at *TCF19* within the MHC region where we did not perform fine-mapping. The distribution of CADD-score between different groups is shown in the inset.

Table 1 | Index coding variants for T2D association signals ($p<2.2x10^{-7}$) that map outside previously established susceptibility loci in trans-ethnic meta-analysis of up to 81,412 cases and 370,832 controls.

Locus	Index variant	Function	rs ID	Chr:Pos	Alleles	RAF	BMI unadjusted		BMI adjusted	
					Risk/Other		OR (95% CI)	*p*-value	OR (95% CI)	*p*-value
Trans-ethnic										
FAM63A	*FAM63A* p.Tyr95Asn	Missense	rs140386498	1: 150,972,959	A/T	0.988	1.23 (1.15-1.32)	$7.5x10^{-8}$	1.21 (1.14-1.29)	$6.7x10^{-7}$
CEP68	*CEP68* p.Gly74Ser	Missense	rs7572857	2: 65,296,798	G/A	0.849	1.05 (1.03-1.07)	$8.3x10^{-9}$	1.05 (1.03-1.07)	$6.6x10^{-7}$
KIF9	*KIF9* p.Arg638Trp	Missense	rs2276853	3: 47,282,303	A/G	0.587	1.02 (1.01-1.04)	$8.0x10^{-5}$	1.03 (1.02-1.04)	$5.3x10^{-8}$
ANKH	*ANKH* p.Arg187Gln	Missense	rs146886108	5: 14,751,305	C/T	0.996	1.20 (1.06-1.36)	$1.4x10^{-7}$	1.20 (1.07-1.34)	$3.5x10^{-7}$
POC5-ANKDD1B	*POC5* p.His36Arg	Missense	rs2307111	5: 75,003,678	T/C	0.551	1.04 (1.03-1.06)	$1.6x10^{-15}$	1.02 (1.01-1.04)	$2.1x10^{-5}$
LPL	*LPL* p.Ser474*	Stop gain	rs328	8: 19,819,724	C/G	0.905	1.05 (1.02-1.07)	$6.8x10^{-9}$	1.05 (1.02-1.07)	$2.3x10^{-7}$
TPCN2	*TPCN2* p.Val219Ile	Missense	rs72928978	11: 68,831,364	G/A	0.888	1.05 (1.03-1.08)	$5.2x10^{-7}$	1.05 (1.03-1.07)	$1.8x10^{-8}$
WSCD2	*WSCD2* p.Thr113Ile	Missense	rs3764002	12: 108,618,630	C/T	0.714	1.03 (1.01-1.04)	$3.3x10^{-8}$	1.03 (1.01-1.04)	$1.2x10^{-7}$
ZZEF1	*ZZEF1* p.Ile402Val	Missense	rs781831	17: 3,947,644	C/T	0.425	1.04 (1.03-1.06)	$8.3x10^{-11}$	1.03 (1.02-1.05)	$1.8x10^{-7}$
MLX	*MLX* p.Gln139Arg	Missense	rs665268	17: 40,722,029	G/A	0.301	1.03 (1.02-1.05)	$2.0x10^{-8}$	1.03 (1.01-1.04)	$1.1x10^{-5}$
TTLL6	*TTLL6* p.Glu712Asp	Missense	rs2032844	17: 46,847,364	C/A	0.760	1.03 (1.02-1.05)	$1.2x10^{-7}$	1.03 (1.01-1.04)	0.00098
PNPLA3	*PNPLA3* p.Ile148Met	Missense	rs738409	22: 44,324,727	G/C	0.241	1.04 (1.03-1.06)	$2.1x10^{-10}$	1.05 (1.03-1.06)	$2.8x10^{-11}$
PIM3	*PIM3* p.Val300Ala	Missense	rs4077129	22: 50,356,693	T/C	0.283	1.04 (1.02-1.06)	$1.9x10^{-7}$	1.04 (1.02-1.05)	$3.5x10^{-8}$
European-specific [a]										
PLCB3	*PLCB3* p.Ser778Leu	Missense	rs35169799	11: 64,031,241	T/C	0.061	1.03 (1.00-1.06)	0.00012	1.05 (1.02-1.07)	$1.7x10^{-6}$
C17orf58-BPTF	*C17orf58* p.Ile92Val	Missense	rs9891146	17: 65,988,049	T/C	0.346	1.03 (1.02-1.05)	$7.2x10^{-7}$	1.02 (1.00-1.03)	0.0027
ZHX3	*ZHX3* p.Asn310Ser	Missense	rs17265513	20: 39,832,628	C/T	0.172	1.04 (1.02-1.06)	$6.8x10^{-6}$	1.03 (1.02-1.05)	$3.2x10^{-5}$

Chr: chromosome. Pos: Position build 37. RAF: risk allele frequency. BMI: body mass index. OR: odds ratio. CI: confidence interval.

[a]Attained significance in European ancestry specific meta-analyses of up to 48,286 cases and 250,671 controls (**Supplementary Table 3**).

bioRxiv preprint doi: https://doi.org/10.1101/144410; this version posted May 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 2 | Posterior probabilities for coding variants within 99% credible set across loci with annotation informed and functionally unweighted prior.

Locus	Variant	rs ID	Chr	Position	Posterior probability		Cumulative posterior probability attributed to coding variants	
					π_U	π_A	π_U	π_A
MACF1	*MACF1* p.Ile39Val	rs16826069	1	39,797,055	0.0121	0.240	0.0317	0.628
	***MACF1* p.Met1424Val**	**rs2296172**	**1**	**39,835,817**	0.0113	0.224		
	MACF1 p.Lys1625Asn	rs41270807	1	39,801,815	0.00821	0.163		
FAM63A	*FAM63A* p.Tyr95Asn	rs140386498	1	150,972,959	0.00523	0.129	0.0122	0.303
GCKR	*GCKR* p. Pro 446Leu	rs1260326	2	27,730,940	0.773	0.995	0.773	0.995
THADA	***THADA* p.Cys845Tyr**	**rs35720761**	**2**	**43,519,977**	**<0.00100**	**0.0106**	0.00338	0.120
	THADA p.Thr897Ala	rs7578597	2	43,732,823	0.00300	0.1070		
CEP68	*CEP68* p.Gly74Ser	rs7572857	2	65,296,798	<0.00100	0.00431	<0.00100	0.00431
GRB14	*COBLL1* p.Asn901Asp	rs7607980	2	165,551,201	0.00583	0.160	0.00583	0.160
PPARG	*PPARG* p.Pro12Ala	rs1801282	3	12,393,125	0.0238	0.410	0.0238	0.410
KIF9	*SETD2* p.Pro1962Lys	rs4082155	3	47,125,385	0.00812	0.171	0.0183	0.384
	NBEAL2 p.Arg511Gly	rs11720139	3	47,036,756	0.00459	0.0967		
	***KIF9* p.Arg638Trp**	**rs2276853**	**3**	**47,282,303**	0.00278	0.0585		
IGF2BP2	*SENP2* p.Thr291Lys	rs6762208	3	185,331,165	<0.00100	<0.00100	<0.00100	<0.00100
WFS1	*WFS1* p.Val333Ile	rs1801212	4	6,302,519	<0.00100	0.00129	<0.00100	0.00433
ANKH	*ANKH* p.Arg187Gln	rs146886108	5	14,751,305	0.459	0.972	0.447	0.972
POC5	*POC5* p.His36Arg	rs2307111	5	75,003,678	0.697	0.954	0.702	0.986
PAM-PPIP5K2	*PAM* p.Asp336Gly	rs35658696	5	102,338,811	0.288	0.885	0.309	0.947
	PPIP5K2 p.Ser1207Gly	rs36046591	5	102,537,285	0.0204	0.0628		
RREB1 p.Asp1171Asn	*RREB1* p.Asp1171Asn	rs9379084	6	7,231,843	0.920	0.997	0.920	0.997
RREB1 p.Ser1499Tyr	*RREB1* p.Ser1499Tyr	rs35742417	6	7,247,344	<0.00100	0.0128	0.00495	0.111
LPL	*LPL* p.Ser474*	rs328	8	19,819,724	0.0231	0.832	0.0231	0.832
SLC30A8	*SLC30A8* p.Arg276Trp	rs13266634	8	118,184,783	0.295	0.823	0.295	0.823
GPSM1	*GPSM1* p.Ser391Leu	rs60980157	9	139,235,415	0.0310	0.557	0.0310	0.557
KCNJ11-ABCC8	*KCNJ11* p.Val250Ile	rs5215	11	17,408,630	0.208	0.412	0.481	0.951
	***KCNJ11* p.Lys29Glu**	**rs5219**	**11**	**17,409,572**	0.190	0.376		
	ABCC8 p.Ala1369Ser	rs757110	11	17,418,477	0.0826	0.163		
PLCB3	*PLCB3* p.Ser778Leu	rs35169799	11	64,031,241	0.113	0.720	0.130	0.830
TPCN2	*TPCN2* p.Val219Ile	rs72928978	11	68,831,364	<0.00100	0.00361	0.00601	0.140
CENTD2	*ARAP1* p.Gln802Glu	rs56200889	11	72,408,055	<0.00100	<0.00100	<0.00100	<0.00100
KLHDC5	*MRPS35* p.Gly43Arg	rs1127787	12	27,867,727	<0.00100	<0.00100	<0.00100	<0.00100
WSCD2	*WSCD2* p.Thr113Ile	rs3764002	12	108,618,630	0.281	0.955	0.282	0.958
HNF1A p.Ile75Leu	*HNF1A*_Gly226Ala	rs56348580	12	121,432,117	0.358	0.894	0.358	0.894
	***HNF1A* p.Ile75Leu**	**rs1169288**	**12**	**121,416,650**	<0.00100	<0.00100		
HNF1A p.Ala146Val	*HNF1A* p.Ala146Val	rs1800574	12	121,416,864	0.269	0.867	0.280	0.902
MPHOSPH9	*SBNO1* p.Ser729Asn	rs1060105	12	123,806,219	0.00166	0.0539	0.00176	0.0574
ZZEF1	*ZZEF1* p.Ile402Val	rs781831	17	3,947,644	<0.00100	0.00129	<0.00100	0.0183
MLX	*MLX* p.Gln139Arg	rs665268	17	40,722,029	0.00210	0.0382	0.00219	0.0398
TTLL6	***TTLL6* p.Glu712Asp**	**rs2032844**	**17**	**46,847,364**	<0.00100	<0.00100	0.0164	0.305
	CALCOCO2 p.Pro347Ala	rs10278	17	46,939,658	0.0100	0.187		
	SNF8 p.Arg155His	rs57901004	17	47,011,897	0.00493	0.0917		
C17orf58	*C17orf58* p.Ile92Val	rs9891146	17	65,988,049	<0.00100	0.00994	<0.00100	0.00995
CILP2	***TM6SF2* p.Glu167Lys**	**rs58542926**	**19**	**19,379,549**	0.211	0.732	0.263	0.913
	***TM6SF2* p.Leu156Pro**	rs187429064	19	19,380,513	0.0495	0.172		
GIPR	*GIPR* p.Glu318Gln	rs1800437	19	46,181,392	0.169	0.901	0.169	0.901
ZHX3	*ZHX3* p.Asn310Ser	rs17265513	20	39,832,628	<0.00100	0.00287	0.00316	0.110
HNF4A	*HNF4A* p.Thr139Ile	rs1800961	20	43,042,364	1.00	1.00	1.00	1.00
MTMR3-ASCC2	***ASCC2* p.Asp407His**	**rs28265**	**22**	**30,200,761**	0.0111	0.192	0.0278	0.481
	ASCC2 p.Pro423Ser	rs36571	22	30,200,713	0.00673	0.116		
	ASCC2 p.Val123Ile	rs11549795	22	30,221,120	0.00619	0.107		
	MTMR3 p.Asn960Ser	rs41278853	22	30,416,527	0.00377	0.0652		
PNPLA3	*PNPLA3* p.Ile148Met	rs738409	22	44,324,727	0.112	0.691	0.130	0.806
	PARVB p.Trp37Arg	rs1007863	22	44,395,451	0.0167	0.103		
PIM3	*IL17REL* p.Leu333Pro	rs5771069	22	50,435,480	0.0414	0.419	0.0470	0.475
	IL17REL p.Gly70Arg	rs9617090	22	50,439,194	0.00530	0.0536		
	***PIM3* p.Val300Ala**	**rs4077129**	**22**	**50,356,693**	<0.00100	0.00176		

Chr: chromosome. Pos: Position build 37. π_U: functionally unweighted prior; π_A: annotation informed prior. Index coding variants are highlighted in bold. Fine-mapping analysis included 50,160 T2D cases and 465,272 controls.

bioRxiv preprint doi: https://doi.org/10.1101/144410; this version posted May 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

DATA AVAILABILITY STATEMENT

Summary level data from the exome array component of this project will be made available at the DIAGRAM consortium website http://diagram-consortium.org/ and Accelerating Medicines Partnership T2D portal http://www.type2diabetesgenetics.org/.

ACKNOWLEDGMENTS

A full list of acknowledgments appears in the **Supplementary Information**. Part of this work was conducted using the UK Biobank resource.

AUTHOR CONTRIBUTIONS

Project co-ordination. A.Mahajan, A.P.M., J.I.R., M.I.M.

Core analyses and writing. A.Mahajan, J.W., S.M.W, W.Zhao, N.R.R., A.Y.C., W.G., H.K., R.A.S., I.Barroso, T.M.F., M.O.G., J.B.M., M.Boehnke, D.S., A.P.M., J.I.R., M.I.M.

Statistical Analysis in individual studies. A.Mahajan, J.W., S.M.W., W.Zhao, N.R.R., A.Y.C., W.G., H.K., D.T., N.W.R., X.G., Y.Lu, M.Li, R.A.J., Y.Hu, S.Huo, K.K.L., W.Zhang, J.P.C., B.P., J.Flannick, N.G., V.V.T., J.Kravic, Y.J.K., D.V.R., H.Y., M.M.-N., K.M., R.L.-G., T.V.V., J.Marten, J.Li, A.V.S., P.An, S.L., S.G., G.M., A.Demirkan, J.F.T., V.Steinthorsdottir, M.W., C.Lecoeur, M.Preuss, L.F.B., P.Almgren, J.B.-J., J.A.B., M.Canouil, K.-U.E., H.G.d.H., Y.Hai, S.Han, S.J., F.Kronenberg, K.L., L.A.L., J.-J.L., H.L., C.-T.L., J.Liu, R.M., K.R., S.S., P.S., T.M.T., G.T., A.Tin, A.R.W., P.Y., J.Y., L.Y., R.Y., J.C.C., D.I.C., C.v.D., J.Dupuis, P.W.F., A.Köttgen, D.M.-K., N.Soranzo, R.A.S., A.P.M.

Genotyping. A.Mahajan, N.R.R., A.Y.C., Y.Lu, Y.Hu, S.Huo, B.P., N.G., R.L.-G., P.An, G.M., E.A., N.A., C.B., N.P.B., Y.-D.I.C., Y.S.C., M.L.G., H.G.d.H., S.Hackinger, S.J., B.-J.K., P.K., J.Kriebel, F.Kronenberg, H.L., S.S.R., K.D.T., E.B., E.P.B., P.D., J.C.F., S.R.H., C.Langenberg, M.A.P., F.R., A.G.U., J.C.C., D.I.C., P.W.F., B.-G.H., C.H., E.I., S.L.K., J.S.K., Y.Liu, R.J.F.L., N.Soranzo, N.J.W., R.A.S., T.M.F., A.P.M., J.I.R., M.I.M.

Cross-trait lookups in unplublished data. S.M.W., A.Y.C., Y.Lu, M.Li, M.G., H.M.H., A.E.J., D.J.L., E.M., G.M.P., H.R.W., S.K., C.J.W.

Phenotyping. Y.Lu, Y.Hu, S.Huo, P.An, S.L., A.Demirkan, S.Afaq, S.Afzal, L.B.B., A.G.B., I.Brandslund, C.C., S.V.E., G.G., V.Giedraitis, A.T.-H., M.-F.H., B.I., M.E.J., T.J., A.Käräjämäki, S.S.K., H.A.K., P.K., F.Kronenberg, B.L., H.L., K.-H.L., A.L., J.Liu, M.Loh, V.M., R.M.-C., G.N., M.N., S.F.N., I.N., P.A.P., W.R., L.R., O.R., S.S., E.S., K.S.S., A.S., B.T., A.Tönjes, A.V., D.R.W.,

bioRxiv preprint doi: https://doi.org/10.1101/144410; this version posted May 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

H.B., E.P.B., A.Dehghan, J.C.F., S.R.H., C.Langenberg, A.D.Morris, R.d.M., M.A.P., A.R., P.M.R., F.R.R., V.Salomaa, W.H.-H.S., R.V., J.C.C., J.Dupuis, O.H.F., H.G., B.-G.H., T.H., A.T.H., C.H., S.L.K., J.S.K., A.Köttgen, L.L., Y.Liu, R.J.F.L., C.N.A.P., J.S.P., O.P., B.M.P., M.B.S., N.J.W., T.M.F., M.O.G.

Individual study design and principal investigators. N.G., P.An, B.-J.K., P.Amouyel, H.B., E.B., E.P.B., R.C., F.S.C., G.D., A.Dehghan, P.D., M.M.F., J.Ferrières, J.C.F., P.Frossard, V.Gudnason, T.B.H., S.R.H., J.M.M.H., M.I., F.Kee, J.Kuusisto, C.Langenberg, L.J.L., C.M.L., S.M., T.M., O.M., K.L.M., M.M., A.D.Morris, A.D.Murray, R.d.M., M.O.-M., K.R.O., M.Perola, A.P., M.A.P., P.M.R., F.R., F.R.R., A.H.R., V.Salomaa, W.H.-H.S., R.S., B.H.S., K.Strauch, A.G.U., R.V., M.Blüher, A.S.B., J.C.C., D.I.C., J.Danesh, C.v.D., O.H.F., P.W.F., P.Froguel, H.G., L.G., T.H., A.T.H., C.H., E.I., S.L.K., F.Karpe, J.S.K., A.Köttgen, K.K., M.Laakso, X.L., L.L., Y.Liu, R.J.F.L., J.Marchini, A.Metspalu, D.M.-K., B.G.N., C.N.A.P., J.S.P., O.P., B.M.P., R.R., N.Sattar, M.B.S., N.Soranzo, T.D.S., K.Stefansson, M.S., U.T., T.T., J.T., N.J.W., J.G.W., E.Z., I.Barroso, T.M.F., J.B.M., M.Boehnke, D.S., A.P.M., J.I.R., M.I.M.

MATERIALS & CORRESPONDENCE

Correspondence and requests for materials should be addressed to mark.mccarthy@drl.ox.ac.uk and anubha@well.ox.ac.uk. Reprints and permissions information is available at www.nature.com/reprints.

DISCLOSURES

Jose C Florez has received consulting honoraria from Merck and from Boehringer-Ingelheim. Daniel I Chasman received funding for exam chip genotyping in the WGHS from Amgen. Oscar H Franco works in ErasmusAGE, a center for aging research across the life course funded by Nestlé Nutrition (Nestec Ltd.), Metagenics Inc., and AXA. Nestlé Nutrition (Nestec Ltd.), Metagenics Inc., and AXA had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; and preparation, review or approval of the manuscript. Erik Ingelsson is an advisor and consultant for Precision Wellness, Inc., and advisor for Cellink for work unrelated to the present project. Bruce M Psaty serves on the DSMB for a clinical trial funded by the manufacturer (Zoll LifeCor) and on the Steering Committee of the Yale Open Data Access Project funded by Johnson & Johnson. Inês Barroso and spouse own stock in GlaxoSmithKline and Incyte Corporation.

bioRxiv preprint doi: https://doi.org/10.1101/144410; this version posted May 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Timothy Frayling has consulted for Boeringer IngelHeim and Sanofi on the genetics of diabetes. Danish Saleheen has received support from Pfizer, Regeneron, Genentech and Eli Lilly. Mark I McCarthy has served on advisory panels for NovoNordisk and Pfizer, and received honoraria from NovoNordisk, Pfizer, Sanofi-Aventis and Eli Lilly.

bioRxiv preprint doi: https://doi.org/10.1101/144410; this version posted May 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

ONLINE METHODS

Ethics statement. All human research was approved by the relevant institutional review boards, and conducted according to the Declaration of Helsinki. All participants provided written informed consent.

Derivation of significance thresholds. We considered five categories of annotation[16] of variants on the exome array in order of decreasing effect on biological function: (1) PTVs (stop-gain and stop-loss, frameshift indel, donor and acceptor splice-site, and initiator codon variants, n_1=8,388); (2) moderate-impact variants (missense, in-frame indel, and splice region variants, n_2=216,114); (3) low-impact variants (synonymous, 3' and 5' UTR, and upstream and downstream variants, n_3=8,829); (4) other variants mapping to DNase I hypersensitive sites in any of 217 cell types[8] (DHS, n_4=3,561); and (5) other variants not mapping to DHS (n_5=10,578). To account for the greater prior probability of causality for variants with greater effect on biological function, we determined a weighted Bonferroni-corrected significance threshold on the basis of reported enrichment[16], denoted w_i, in each annotation category, i: w_1=165; w_2=33; w_3=3; w_4=1.5; w_5=0.5. For coding variants (annotation categories 1 and 2):

$$\alpha = \frac{0.05\sum_{i=1}^{2} n_i w_i}{(\sum_{i=1}^{2} n_i)(\sum_{i=1}^{5} n_i w_i)} = 2.21\text{x}10^{-7}.$$

We note that this threshold is similar to a simple Bonferroni correction for the total number of coding variants on the array, which would yield:

$$\alpha = \frac{0.05}{224502} = 2.23\text{x}10^{-7}\ .$$

For non-coding variants (annotation categories 3, 4 and 5) the weighted Bonferroni-corrected significance threshold is:

$$\alpha = \frac{0.05\sum_{i=3}^{5} n_i w_i}{(\sum_{i=3}^{5} n_i)(\sum_{i=1}^{5} n_i w_i)} = 9.45\text{x}10^{-9}.$$

bioRxiv preprint doi: https://doi.org/10.1101/144410; this version posted May 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Exome-array study-level analyses. Within each study, genotype calling and quality control were undertaken according to protocols developed by the UK Exome Chip Consortium or the CHARGE central calling effort[39] (**Supplementary Table 1**). Within each study, variants were then excluded for the following reasons: (i) not mapping to autosomes or X chromosome; (ii) multi-allelic and/or insertion-deletion; (iii) monomorphic; (iv) call rate <99%; or (v) exact $p<10^{-4}$ for deviation from Hardy-Weinberg equilibrium (autosomes only).

We tested association of T2D with each variant in a linear mixed model, implemented in RareMetalWorker[17], using a genetic relationship matrix (GRM) to account for population structure and relatedness. For participants from family-based studies, known relationships were incorporated directly in the GRM. For founders and participants from population-based studies, the GRM was constructed from pair-wise identity by descent (IBD) estimates based on LD pruned ($r^2<0.05$) autosomal variants with MAF≥1%, after exclusion of those in high LD and complex regions[40,41], and those mapping to established T2D loci. We considered additive, dominant, and recessive models for the effect of the minor allele, adjusted for age and sex (where appropriate) and additional study-specific covariates (**Supplementary Table 2**). Analyses were also performed with and without adjustment for BMI (where available Supplementary Table 2).

For single-variant association analyses, variants with minor allele count ≤10 were excluded. Association summary statistics for each analysis were corrected for residual inflation by means of genomic control[42], calculated after excluding variants mapping to established T2D susceptibility loci. For gene-based analyses, we made no variant exclusions on the basis of minor allele count.

Exome-sequence analyses. We used summary statistics of T2D association analysis conducted on 8,321 T2D cases and 8,421 controls from across different ancestries, all genotyped using exome sequencing. Details of samples included, sequencing, and quality control are described elsewhere[12,15] (http://www.type2diabetesgenetics.org/). Samples were subdivided into 15 sub-groups according to ancestry and study of origin. Each sub-group was analysed independently, with sub-group specific principal components and genetic relatedness matrices. Association tests were performed with both a linear mixed model, as implemented in EMMAX[43], using covariates for sequencing batch, and the Firth test, using covariates for principal components and sequencing batch. Related samples were

bioRxiv preprint doi: https://doi.org/10.1101/144410; this version posted May 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

excluded from the Firth analysis but maintained in the EMMAX analysis. Variants were then filtered from each sub-group analysis, according to call rate, differential case-control missing-ness, or deviation from Hardy-Weinberg equilibrium (as computed separately for each sub-group). Association statistics were then combined via a fixed-effects inverse-variance weighted meta-analysis, at both the level of ancestry as well as across all samples. P-values were taken from the EMMAX analysis, while effect sizes estimates are taken from the Firth analysis. Analyses were performed with and without adjustment for BMI. From exome sequence summary statistics, we extracted variants passing quality control and present on the exome array.

GWAS analyses. The UK Biobank is a large detailed prospective study of more than 500,000 participants aged 40-69 years when recruited in 2006-2010[13]. Prevalent T2D status was defined using self-reported medical history and medication in UK Biobank participants[44]. Participants were genotyped with the UK Biobank Axiom Array or UK BiLEVE Axiom Array, and quality control and population structure analyses were performed centrally at UK Biobank. We defined a subset of "white European" ancestry samples (n=120,286) as those who both self-identified as white British and were confirmed as ancestrally "Caucasian" from the first two axes of genetic variation from principal components analysis. Imputation was also performed centrally at UK Biobank for the autosomes only, up to a merged reference panel from the 1000 Genomes Project (multi-ethnic, phase 3, October 2014 release)[21] and the UK10K Project[9]. We used SNPTESTv2.5[45] to test for association of T2D with each SNP in a logistic regression framework under an additive model, and after adjustment for age, sex, six axes of genetic variation, and genotyping array as covariates. Analyses were performed with and without adjustment for BMI, after removing related individuals.

GERA is a large multi-ethnic population-based cohort, created for investigating the genetic and environmental basis of age-related diseases [dbGaP phs000674.p1]. T2D status is based on ICD-9 codes in linked electronic medical health records, with all other participants defined as controls. Participants have previously been genotyped using one of four custom arrays, which have been designed to maximise coverage of common and low-frequency variants in non-Hispanic white, East Asian, African American, and Latino ethnicities[46,47]. Methods for quality control have been described previously[14]. Each of the

bioRxiv preprint doi: https://doi.org/10.1101/144410; this version posted May 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

four genotyping arrays were imputed separately, up to the 1000 Genomes Project reference panel (autosomes, phase 3, October 2014 release; X chromosome, phase 1, March 2012 release) using IMPUTEv2.3[48,49]. We used SNPTESTv2.5[45] to test for association of T2D with each SNP in a logistic regression framework under an additive model, and after adjustment for sex and nine axes of genetic variation from principal components analysis as covariates. BMI was not available for adjustment in GERA.

For UK Biobank and GERA, we extracted variants passing standard imputation quality control thresholds (IMPUTE info≥0.4)[50] and present on the exome array. Association summary statistics under an additive model were corrected for residual inflation by means of genomic control[42], calculated after excluding variants mapping to established T2D susceptibility loci: GERA (λ=1.097 for BMI unadjusted analysis) and UK Biobank (λ=1.043 for BMI unadjusted analysis, λ=1.056 for BMI adjusted analysis).

Single-variant meta-analysis. We aggregated association summary statistics under an additive model across studies, with and without adjustment for BMI, using METAL[51]: (i) effective sample size weighting of *Z*-scores to obtain *p*-values; and (ii) inverse variance weighting of log-odds ratios. For exome-array studies, allelic effect sizes and standard errors obtained from the RareMetalWorker linear mixed model were converted to the log-odds scale prior to meta-analysis to correct for case-control imbalance[52].

The European-specific meta-analyses aggregated association summary statistics from a total of 48,286 cases and 250,671 controls from: (i) 33 exome-array studies of European ancestry; (ii) exome-array sequence from individuals of European ancestry; and (iii) GWAS from UK Biobank. Note that non-coding variants represented on the exome array were not available in exome sequence. The European-specific meta-analyses were corrected for residual inflation by means of genomic control[42], calculated after excluding variants mapping to established T2D susceptibility loci: λ=1.091 for BMI unadjusted analysis and λ=1.080 for BMI adjusted analysis.

The trans-ethnic meta-analyses aggregated association summary statistics from a total of 81,412 cases and 370,832 controls across all studies (exome array, exome sequence, and GWAS), irrespective of ancestry. Note that non-coding variants represented on the exome array were not available in exome sequence. The trans-ethnic meta-analyses were corrected for residual inflation by means of genomic control[42], calculated after excluding

bioRxiv preprint doi: https://doi.org/10.1101/144410; this version posted May 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

variants mapping to established T2D susceptibility loci: λ=1.073 for BMI unadjusted analysis and λ=1.068 for BMI adjusted analysis. Heterogeneity in allelic effect sizes between exome-array studies contributing to the trans-ethnic meta-analysis was assessed by Cochran's *Q* statistic[53].

Detection of distinct association signals. Conditional analyses were undertaken to detect association signals by inclusion of index variants and/or tags for previously reported non-coding GWAS lead SNPs as covariates in the regression model at the study level. Within each exome-array study, approximate conditional analyses were undertaken under a linear mixed model using RareMetal[17], which uses score statistics and the variance-covariance matrix from the RareMetalWorker single-variant analysis to estimate the correlation in effect size estimates between variants due to LD. Study-level allelic effect sizes and standard errors obtained from the approximate conditional analyses were converted to the log-odds scale to correct for case-control imbalance[52]. Within each GWAS, exact conditional analyses were performed under a logistic regression model using SNPTESTv2.5[45]. GWAS variants passing standard imputation quality control thresholds (IMPUTE info≥0.4)[50] and present on the exome array were extracted for meta-analysis.

Association summary statistics were aggregated across studies, with and without adjustment for BMI, using METAL[51]: (i) effective sample size weighting of *Z*-scores to obtain *p*-values; and (ii) inverse variance weighting of log-odds ratios.

Non-additive association models. For exome-array studies only, we aggregated association summary statistics under recessive and dominant models across studies, with and without adjustment for BMI, using METAL[51]: (i) effective sample size weighting of *Z*-scores to obtain *p*-values; and (ii) inverse variance weighting of log-odds ratios. Allelic effect sizes and standard errors obtained from the RareMetalWorker linear mixed model were converted to the log-odds scale prior to meta-analysis to correct for case-control imbalance[52]. The European-specific meta-analyses aggregated association summary statistics from a total of 41,066 cases and 136,024 controls from 33 exome-array studies of European ancestry. The European-specific meta-analyses were corrected for residual inflation by means of genomic control[42], calculated after excluding variants mapping to established T2D susceptibility loci: λ=1.076 and λ=1.083 for BMI unadjusted analysis, under recessive and dominant models

bioRxiv preprint doi: https://doi.org/10.1101/144410; this version posted May 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

respectively, and λ=1.081 and λ=1.062 for BMI adjusted analysis, under recessive and dominant models respectively. The trans-ethnic meta-analyses aggregated association summary statistics from a total of 58,425 cases and 188,032 controls across all exome-array studies, irrespective of ancestry. The trans-ethnic meta-analyses were corrected for residual inflation by means of genomic control[42], calculated after excluding variants mapping to established T2D susceptibility loci: λ=1.041 and λ=1.071 for BMI unadjusted analysis, under recessive and dominant models respectively, and λ=1.031 and λ=1.063 for BMI adjusted analysis, under recessive and dominant models respectively.

Gene-based meta-analyses. For exome-array studies only, we aggregated association summary statistics under an additive model across studies, with and without adjustment for BMI, using RareMetal[17]. This approach uses score statistics and the variance-covariance matrix from the RareMetalWorker single-variant analysis to estimate the correlation in effect size estimates between variants due to LD. We performed gene-based analyses using a burden test (assuming all variants have same direction of effect on T2D susceptibility) and SKAT (allowing variants to have different directions of effect on T2D susceptibility). We used two previously defined filters for annotation and MAF[18] to define group files: (i) strict filter, including 44,666 variants; and (ii) broad filter, including all variants from the strict filter, and 97,187 additional variants.

We assessed the contribution of each variant to gene-based signals by performing approximate conditional analyses. We repeated RareMetal analyses for the gene, excluding each variant in turn from the group file, and compared the strength of the association signal.

Fine-mapping of coding variant association signals with T2D susceptibility. We defined a locus as mapping 500kb up- and down-stream of each index coding variant (**Supplementary Table 5**), excluding the MHC. Our fine-mapping analyses aggregated association summary statistics from 24 GWAS incorporating 41,284 T2D cases and 311,715 controls of European ancestry from the DIAGRAM Consortium (**Supplementary Table 8**). Each GWAS was imputed using miniMAC[12] or IMPUTEv2[48,49] up to reference panels from the Haplotype Reference Consortium[20], the 1000 Genomes Project (multi-ethnic, phase 3, October 2014 release)[21] and the UK10K Project[9], or population-specific whole-genome sequence data[19]

bioRxiv preprint doi: https://doi.org/10.1101/144410; this version posted May 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(**Supplementary Table 8**). Association with T2D susceptibility was tested for each remaining variant using logistic regression, adjusting for age, sex, and study-specific covariates, under an additive genetic model. Analyses were performed with and without adjustment for BMI. For each study, variants with minor allele count<5 or those with imputation quality r2-hat<0.3 (miniMAC) or proper-info<0.4 (IMPUTE2) were removed. Association summary statistics for each analysis were corrected for residual inflation by means of genomic control[42], calculated after excluding variants mapping to established T2D susceptibility loci.

We aggregated association summary statistics across studies, with and without adjustment for BMI, in a fixed-effects inverse variance weighted meta-analysis, using METAL[51]. The BMI unadjusted meta-analysis was corrected for residual inflation by means of genomic control (λ=1.012)[42], calculated after excluding variants mapping to established T2D susceptibility loci. No adjustment was required for BMI adjusted meta-analysis (λ=0.994). From the meta-analysis, variants were extracted that were present on the HRC panel and reported in at least 50% of total effective sample size.

To delineate distinct association signals in four regions, we undertook approximate conditional analyses, implemented in GCTA[54], to adjust for the index coding variants and non-coding lead GWAS SNPs: (i) *RREB1* p. Asp1171Asn, p.Ser1499Tyr, and rs9505118; (ii) *HNF1A* p.Ile75Leu and p.Ala146Val; (iii) *GIPR* p.Glu318Gln and rs8108269; and (iv) *HNF4A* p.Thr139Ile and rs4812831. We made use of summary statistics from the fixed-effects meta-analyses (BMI unadjusted for *RREB1, HNF1A*, and *HNF4A*, and BMI adjusted for *GIPR* as this signal was only seen in BMI adjusted analysis) and genotype data from 5,000 random individuals of European ancestry from the UK Biobank, as reference for LD between genetic variants across the region.

For each association signal, we first calculated an approximate Bayes' factor[55] in favour of association on the basis of allelic effect sizes and standard errors from the meta-analysis. Specifically, for the *j*th variant,

$$\Lambda_j = \sqrt{\frac{V_j}{V_j+\omega}} \exp\left[\frac{\omega\beta_j^2}{2V_j(V_j+\omega)}\right],$$

bioRxiv preprint doi: https://doi.org/10.1101/144410; this version posted May 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

where β_j and V_j denote the estimated allelic effect (log-OR) and corresponding variance from the meta-analysis. The parameter ω denotes the prior variance in allelic effects, taken here to be 0.04[55].

We then calculated the posterior probability that the *j*th variant drives the association signal, given by

$$\pi_j = \frac{\rho_j \Lambda_j}{\sum_k \rho_k \Lambda_k}.$$

In this expression, ρ_j denotes the prior probability that the *j*th variant drives the association signal, and the summation in the denominator is over all variants across the locus. We considered two prior models: (i) functionally unweighted, for which $\rho_j = 1$ for all variants; and (ii) annotation informed, for which ρ_j is determined by the functional severity of the variant. For the annotation informed prior, we considered five categories of variation[16], such that: (i) $\rho_j = 165$ for PTVs; (ii) $\rho_j = 33$ for moderate-impact variants; (iii) $\rho_j = 3$ for low-impact variants; (iv) $\rho_j = 1.5$ for other variants mapping to DHS; and (v) $\rho_j = 0.5$ for all other variants.

For each locus, the 99% credible set[22] under each prior was then constructed by: (i) ranking all variants according to their posterior probability of driving the association signal; and (ii) including ranked variants until their cumulative posterior probability of driving the association attained or exceeded 0.99.

Functional impact of coding alleles. We used CADD[35] to obtain scaled Combined Annotation Dependent Depletion score (CADD-scores) for each of the 40 significantly associated coding variants. The CADD method objectively integrates a range of different annotation metrics into a single measure (CADD-score), providing an estimate of deleteriousness for all known variants and an overall rank for this metric across the genome. We obtained the estimates of the intolerance of a gene to harbouring loss-of-function variants (pLI) from the ExAC data set[34]. We used the Kolmogorov-Smirnov test to determine whether fine-mapping groups 1 and 2 have the same statistical distribution for each of these parameters.

T2D loci and physiological classification. To explore the different patterns of association between T2D and other anthropometric/metabolic/endocrine traits and diseases, we

bioRxiv preprint doi: https://doi.org/10.1101/144410; this version posted May 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

performed hierarchical clustering analysis. We obtained association summary statistics for a range of metabolic traits and other outcomes for 94 coding and non-coding variants that were significantly associated with T2D through collaboration or by querying publically available GWAS meta-analysis datasets. The z-score (allelic effect/SE) was aligned to the T2D-risk allele. We obtained the distance matrix amongst z-score of the loci/traits using the Euclidean measure and performed clustering using the complete agglomeration method. Clustering was visualised it by constructing dendogram and a heatmap.

Data availability.

39. Grove, M.L. *et al.* Best practices and joint calling of the HumanExome BeadChip: the CHARGE Consortium. *PLoS One* **8**, e68095 (2013).
40. Price, A.L. *et al.* Long-range LD can confound genome scans in admixed populations. *Am J Hum Genet* **83**, 132-5; author reply 135-9 (2008).
41. Weale, M.E. Quality control for genome-wide association studies. *Methods Mol Biol* **628**, 341-72 (2010).
42. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997-1004 (1999).
43. Kang, H.M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nat Genet* **42**, 348-54 (2010).
44. Eastwood, S.V. *et al.* Algorithms for the Capture and Adjudication of Prevalent and Incident Diabetes in UK Biobank. *PLoS One* **11**, e0162388 (2016).
45. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat Rev Genet* **11**, 499-511 (2010).
46. Hoffmann, T.J. *et al.* Next generation genome-wide association tool: design and coverage of a high-throughput European-optimized SNP array. *Genomics* **98**, 79-89 (2011).
47. Hoffmann, T.J. *et al.* Design and coverage of high throughput genotyping arrays optimized for individuals of East Asian, African American, and Latino race/ethnicity using imputation and a novel hybrid SNP selection algorithm. *Genomics* **98**, 422-30 (2011).

bioRxiv preprint doi: https://doi.org/10.1101/144410; this version posted May 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

48. Howie, B.N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* **5**, e1000529 (2009).

49. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G.R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet* **44**, 955-9 (2012).

50. Winkler, T.W. *et al.* Quality control and conduct of genome-wide association meta-analyses. *Nat Protoc* **9**, 1192-212 (2014).

51. Willer, C.J., Li, Y. & Abecasis, G.R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190-1 (2010).

52. Cook, J.P., Mahajan, A. & Morris, A.P. Guidance for the utility of linear models in meta-analysis of genetic association studies of binary phenotypes. *Eur J Hum Genet* **25**, 240-245 (2017).

53. Ioannidis, J.P., Patsopoulos, N.A. & Evangelou, E. Heterogeneity in meta-analyses of genome-wide association investigations. *PLoS One* **2**, e841 (2007).

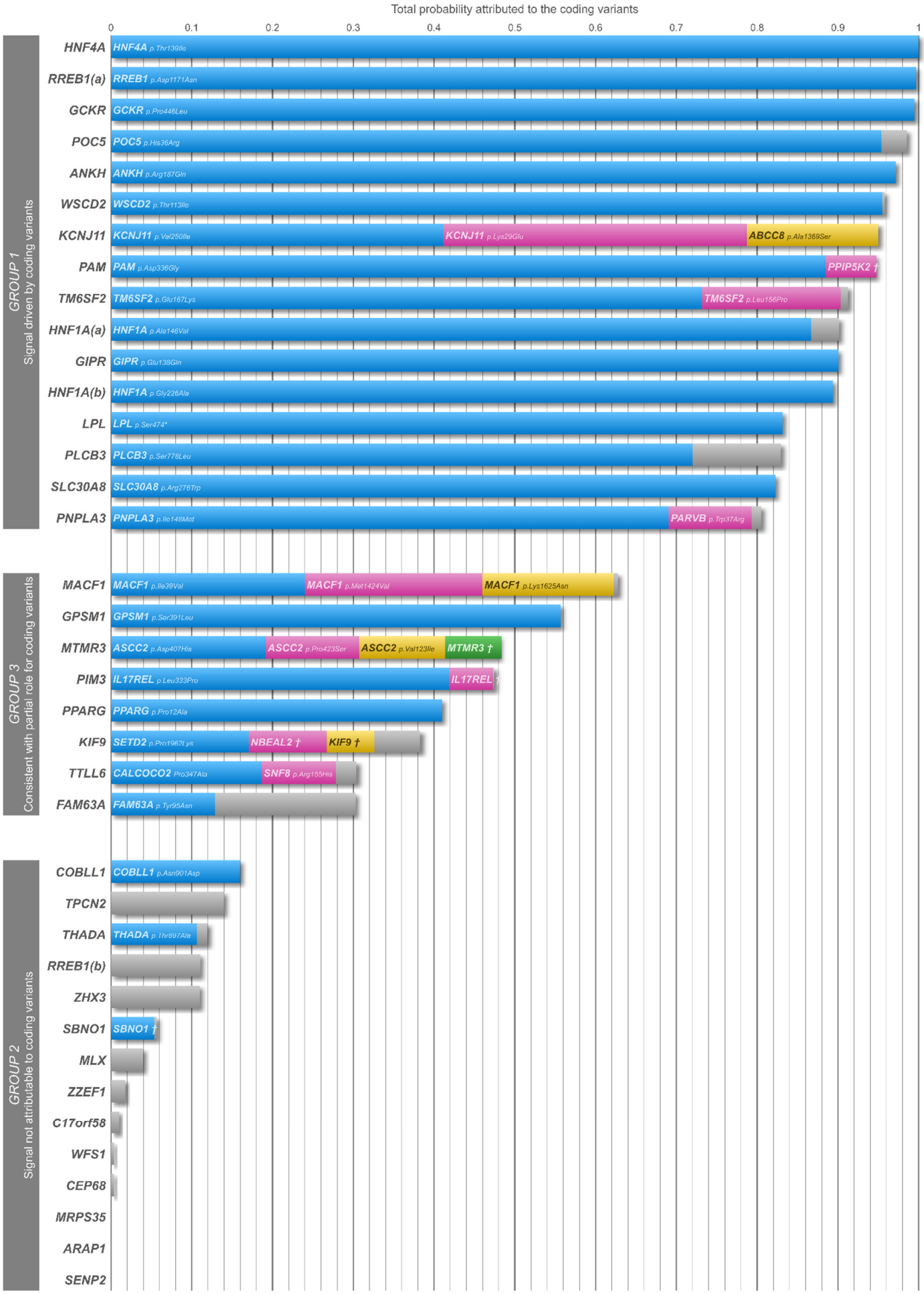
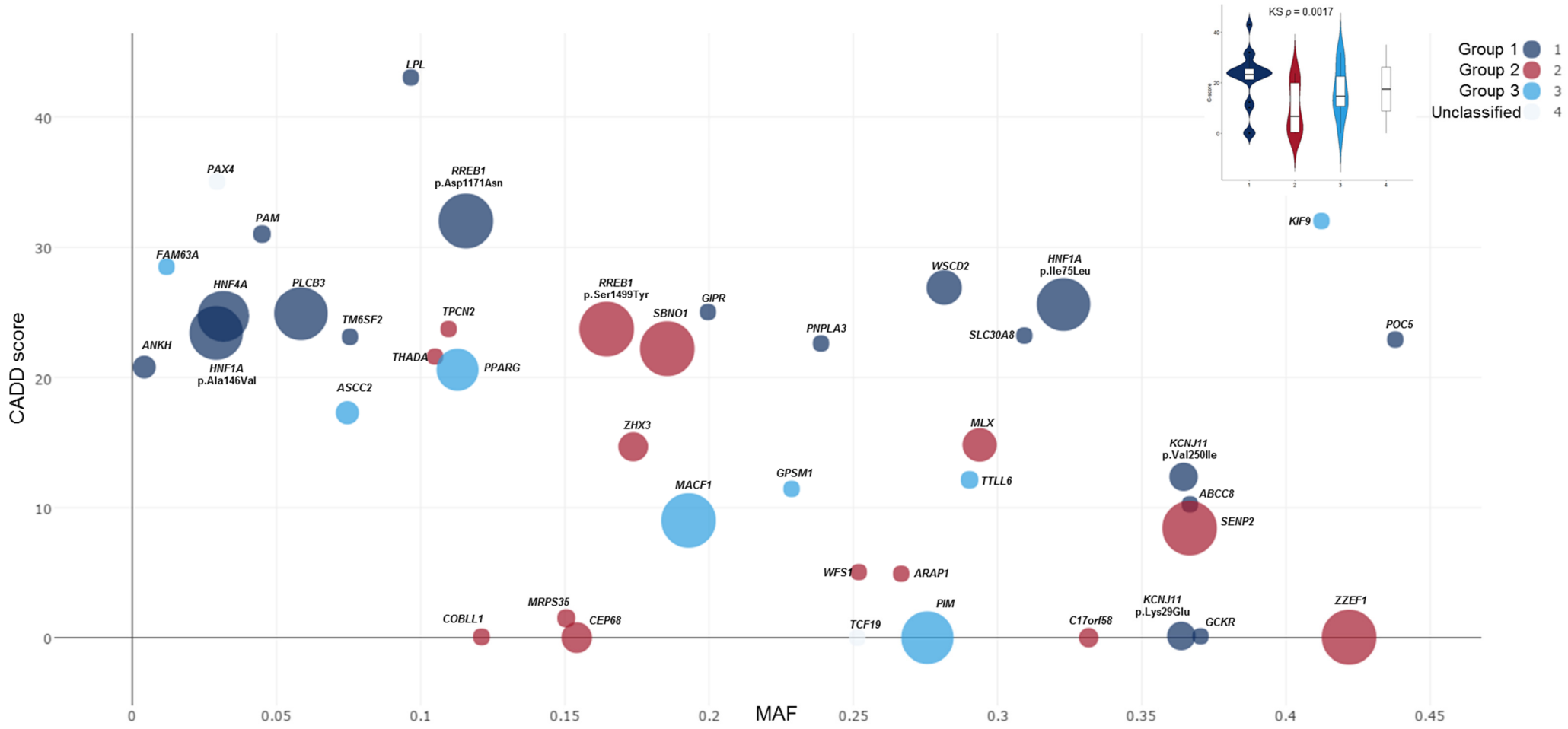
54. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat Genet* **44**, 369-75, s1-3 (2012).

55. Wakefield, J. A Bayesian measure of the probability of false discovery in genetic epidemiology studies. *Am J Hum Genet* **81**, 208-27 (2007).

URLs

Type 2 Diabetes Knowledge Portal: http://www.type2diabetesgenetics.org/

bioRxiv preprint doi: https://doi.org/10.1101/144410; this version posted May 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



bioRxiv preprint doi: https://doi.org/10.1101/144410; this version posted May 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.