

Genome-wide analyses identify 30 loci associated with obsessive–compulsive disorder

In the format provided by the
authors and unedited

Supplementary Notes and Figures

Supplementary Notes	2
Supplementary Note 1: Author Byline – Consortias	2
Supplementary Note 2: Ethics and sample descriptions.....	10
Supplementary Note 3: Definition of independent significant loci	47
Supplementary Note 4: X-chromosome analysis in 23andMe	48
Supplementary Note 5: Subgroup, MTAG & common-factor analyses (GenomicSEM)	48
Supplementary Note 6: Cross-trait genetic correlations.....	52
Supplementary Note 7: Gene-based results.	53
Supplementary Note 8: Sample size and number of GWS loci.....	57
Supplementary Note 9: Assumptions and Limitations of SMR (and TWAS).....	57
Supplementary Note 10: A note on liability scale heritability estimates and sample ascertainment.....	58
Supplementary Figures.....	60
Regional association plots and forest plots of the 30 significant SNPs.....	61
Manhattan-plots and QQ-plots of sub-analyses	92
GenomicSEM one factor model of the four OCD ascertainment sub-groups	98
Cross-trait genetic correlations of the ascertainment-specific subgroup analyses	100
Consistency of SNP findings across OCD GWAS versions.....	102
References	103

Supplementary Notes

Supplementary Note 1: Author Byline – Consortias

23andMe

Stella Aslibekyan, Adam Auton, Elizabeth Babalola, Robert K. Bell, Jessica Bielenberg, Jonathan Bowes, Katarzyna Bryc, Ninad S. Chaudhary, Daniella Coker, Sayantan Das, Emily DelloRusso, Sarah L. Elson, Nicholas Eriksson, Teresa Filstein, Pierre Fontanillas, Will Freyman, Zach Fuller, Chris German, Julie M. Granka, Karl Heilbron, Alejandro Hernandez, Barry Hicks, David A. Hinds, Ethan M. Jewett, Yunxuan Jiang, Katelyn Kukar, Alan Kwong, Yanyu Liang, Keng-Han Lin, Bianca A. Llamas, Matthew H. McIntyre, Steven J. Micheletti, Meghan E. Moreno, Priyanka Nandakumar, Dominique T. Nguyen, Jared O'Connell, Aaron A. Petrakovitz, G. David Poznik, Alexandra Reynoso, Shubham Saini, Morgan Schumacher, Leah Selcer, Anjali J. Shastri, Janie F. Shelton, Jingchunzi Shi, Suyash Shringarpure, Qiaojuan Jane Su, Susana A. Tat, Vinh Tran, Joyce Y. Tung, Xin Wang, Wei Wang, Catherine H. Weldon, Peter Wilton, Corinna D. Wong.

COGA authors

The Collaborative Study on the Genetics of Alcoholism (COGA), Principal Investigators B. Porjesz, V. Hesselbrock, T. Foroud; Scientific Director, A. Agrawal; Translational Director, D. Dick, includes ten different centers: University of Connecticut (V. Hesselbrock); Indiana University (H.J. Edenberg, T. Foroud, Y. Liu, M.H. Plawecki); University of Iowa Carver College of Medicine (S. Kuperman, J. Kramer); SUNY Downstate Health Sciences University (B. Porjesz, J. Meyers, C. Kamarajan, A. Pandey); Washington University in St. Louis (L. Bierut, J. Rice, K. Bucholz, A. Agrawal); University of California at San Diego (M. Schuckit); Rutgers University (J. Tischfield, D. Dick, R. Hart, J. Salvatore); The Children's Hospital of Philadelphia, University of Pennsylvania (L. Almasy); Icahn School of Medicine at Mount Sinai (A. Goate, P. Slesinger); and Howard University (D. Scott). Other COGA collaborators include: L. Bauer (University of Connecticut); J. Nurnberger Jr., L. Wetherill, X. Xuei, D. Lai, S. O'Connor, (Indiana University); G. Chan (University of Iowa; University of Connecticut); D.B. Chorlian, J. Zhang, P. Barr, S. Kinreich, G. Pandey (SUNY Downstate); N. Mullins (Icahn School of Medicine at Mount Sinai); A. Anokhin, S. Hartz, E. Johnson, V. McCutcheon, S. Saccone (Washington University); J. Moore, F. Aliev, Z. Pang, S. Kuo (Rutgers University); A. Merikangas (The Children's Hospital of Philadelphia and University of Pennsylvania); H. Chin and A. Parsian are the NIAAA Staff Collaborators. We continue to be inspired by our memories of Henri Begleiter and Theodore Reich, founding PI and Co-PI of COGA, and also owe a debt of gratitude to other past organizers of COGA, including Ting- Kai Li, P. Michael Conneally, Raymond Crowe, and Wendy Reich, for their critical contributions. This national collaborative study is supported by NIH Grant U10AA008401 from the National Institute on Alcohol Abuse and Alcoholism (NIAAA) and the National Institute on Drug Abuse (NIDA).

EstBB authors

Andres Metspalu¹, Tõnu Esko¹, Reedik Mägi¹, Mari Nelis¹, Georgi Hudjashov¹

1: Estonian Genome Centre, Institute of Genomics, University of Tartu

Hunt authors

Laurent F. Thomas^{1,2,3,4}, Anne H. Skogholt¹, Bendik Winsvold^{1,5,6}, Maiken Elvestad Gabrielsen¹, Sigrid Børte^{1,7,8}, Ole A Andreassen MD, PhD^{9,10}, Ottar Bjerkeset MD, PhD^{11,12}, Ole Kristian Drange MD, PhD^{11,13}, Grete Dyb MD, PhD¹⁴, Katrine Kveli Fjukstad MD, PhD^{15,16}, Lars G Fritsche PhD¹⁷, Ingrid Heuch MD, PhD⁵, Marit Sæbø Indredavik MD, PhD², Håvard Kallestad MD^{11,13}, Amy E Martinsen MSc^{1,2,5}, Gunnar Morken MD, PhD^{11,13}, Jonas Bille Nielsen PhD^{1,18}, Torunn Stene Nøvik MD, PhD¹⁹, Linda M Pedersen PhD⁵, Sara Germans Selvik MD, PhD²⁰, Børge Sivertsen PhD^{11,21,22}, Marit Skrove MD, PhD²³, Synne Øien Stensland MD, PhD^{7,14}, Eystein Stordal MD, PhD^{11,20}, Cristen J Willer PhD¹⁸, Wei Zhou PhD^{24,25}, Kristian Hveem^{1,26,27}, John-Anker Zwart^{1,5,8}

1: K. G. Jebsen Center for Genetic Epidemiology, Department of Public Health and Nursing, Faculty of Medicine and Health Sciences, Norwegian University of Science and Technology (NTNU), Trondheim, Norway.

2: Department of Clinical and Molecular Medicine, Faculty of Medicine and Health Sciences, Norwegian University of Science and Technology (NTNU), Trondheim, Norway.

3: BioCore - Bioinformatics Core Facility, Norwegian University of Science and Technology, Trondheim, Norway.

4: Clinic of Laboratory Medicine, St.Olavs Hospital, Trondheim University Hospital, Trondheim, Norway.

5: Department of Research and Innovation and Education, Division of Clinical Neuroscience, Oslo University Hospital, Oslo, Norway.

6: Department of Neurology, Oslo University Hospital, Oslo, Norway.

7: Research and Communication Unit for Musculoskeletal Health (FORMI), Department of Research and Innovation, Division of Clinical Neuroscience, Oslo University Hospital, Oslo, Norway.

8: Institute of Clinical Medicine, Faculty of Medicine, University of Oslo, Oslo, Norway.

9: Division of Mental Health and Addiction, Oslo University Hospital, Oslo, Norway.

10: NORMENT, University of Oslo, Oslo, Norway.

11: Department of Mental Health, Faculty of Medicine and Health Sciences, Norwegian University of Science and Technology (NTNU), Trondheim, Norway.

12: Faculty of Nursing and Health Sciences, NORD University, Levanger, Norway.

13: Division of Mental Health Care, St. Olavs Hospital, Trondheim University Hospital, Trondheim, Norway.

14: Norwegian Centre for Violence and Traumatic Stress Studies, Oslo, Norway.

15: Department of Psychiatry, Nord-Trøndelag Hospital Trust, Levanger Hospital, Levanger, Norway.

16: Department of Laboratory Medicine, Children's and Women's Health, Norwegian University of Science and Technology (NTNU), Trondheim, Norway.

17: Center for Statistical Genetics, Department of Biostatistics, University of Michigan, Ann Arbor, MI, 48109, USA.

- 18: Department of Internal Medicine, Division of Cardiovascular Medicine, University of Michigan, Ann Arbor, MI, 48109, USA.
- 19: Department of Child and Adolescent Psychiatry, St. Olavs Hospital, Trondheim University Hospital, Trondheim, Norway.
- 20: Department of Psychiatry, Hospital Namsos, Nord-Trøndelag Health Trust, Namsos, Norway.
- 21: Department of Health Promotion, Norwegian Institute of Public Health, Bergen, Norway.
- 22: Department of Research and Innovation, Helse-Fonna HF, Haugesund, Norway.
- 23: Regional Centre for Child and Youth Mental Health and Child Welfare, Department of Mental Health, Faculty of Medicine and Health Sciences, Norwegian University of Science and Technology (NTNU), Trondheim, Norway.
- 24: Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, 48109, USA.
- 25: Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA.
- 26: HUNT Research Center, Department of Public Health and Nursing, Faculty of Medicine and Health Sciences, Norwegian University of Science and Technology, Trondheim, Norway.
- 27: Department of Research, Innovation and Education, St. Olavs Hospital, Trondheim University Hospital, Trondheim, Norway.

VA Million Veteran Program

MVP Program Office

- Sumitra Muralidhar, Ph.D., Program Director
US Department of Veterans Affairs, 810 Vermont Avenue NW, Washington, DC 20420
- Jennifer Moser, Ph.D., Associate Director, Scientific Programs
US Department of Veterans Affairs, 810 Vermont Avenue NW, Washington, DC 20420
- Jennifer E. Deen, B.S., Associate Director, Cohort & Public Relations
US Department of Veterans Affairs, 810 Vermont Avenue NW, Washington, DC 20420

MVP Executive Committee

- Co-Chair: Philip S. Tsao, Ph.D.
VA Palo Alto Health Care System, 3801 Miranda Avenue, Palo Alto, CA 94304
- Co-Chair: Sumitra Muralidhar, Ph.D.
US Department of Veterans Affairs, 810 Vermont Avenue NW, Washington, DC 20420
- J. Michael Gaziano, M.D., M.P.H.
VA Boston Healthcare System, 150 S. Huntington Avenue, Boston, MA 02130
- Elizabeth Hauser, Ph.D.
Durham VA Medical Center, 508 Fulton Street, Durham, NC 27705

- Amy Kilbourne, Ph.D., M.P.H.
VA HSR&D, 2215 Fuller Road, Ann Arbor, MI 48105
- Shih-Wen Luoh, M.D., Ph.D.
VA Portland Health Care System, 3710 SW US Veterans Hospital Rd, Portland, OR 97239
- Michael Matheny, M.D., M.S., M.P.H.
VA Tennessee Valley Healthcare System, 1310 24th Ave. South, Nashville, TN 37212
- Dave Oslin, M.D.
Philadelphia VA Medical Center, 3900 Woodland Avenue, Philadelphia, PA 19104

MVP Co-Principal Investigators

- J. Michael Gaziano, M.D., M.P.H.
VA Boston Healthcare System, 150 S. Huntington Avenue, Boston, MA 02130
- Philip S. Tsao, Ph.D.
VA Palo Alto Health Care System, 3801 Miranda Avenue, Palo Alto, CA 94304

MVP Core Operations

- Lori Churby, B.S., Director, MVP Regulatory Affairs
VA Palo Alto Health Care System, 3801 Miranda Avenue, Palo Alto, CA 94304
- Stacey B. Whitbourne, Ph.D., Director, MVP Cohort Management
VA Boston Healthcare System, 150 S. Huntington Avenue, Boston, MA 02130
- Jessica V. Brewer, M.P.H., Director, MVP Recruitment & Enrollment
VA Boston Healthcare System, 150 S. Huntington Avenue, Boston, MA 02130
- Shahpoor (Alex) Shayan, M.S., Director, MVP Recruitment and Enrollment Informatics
VA Boston Healthcare System, 150 S. Huntington Avenue, Boston, MA 02130
- Luis E. Selva, Ph.D., Executive Director, MVP Biorepositories
VA Boston Healthcare System, 150 S. Huntington Avenue, Boston, MA 02130
- Saiju Pyarajan Ph.D., Director, Data and Computational Sciences
VA Boston Healthcare System, 150 S. Huntington Avenue, Boston, MA 02130
- Kelly Cho, M.P.H, Ph.D., Director, MVP Phenomics Data Core
VA Boston Healthcare System, 150 S. Huntington Avenue, Boston, MA 02130
- Scott L. DuVall, Ph.D., Director, VA Informatics and Computing Infrastructure (VINCI)
VA Salt Lake City Health Care System, 500 Foothill Drive, Salt Lake City, UT 84148
- Mary T. Brophy M.D., M.P.H., Director, VA Central Biorepository
VA Boston Healthcare System, 150 S. Huntington Avenue, Boston, MA 02130

MVP Coordinating Centers

- MVP Coordinating Center, Boston - J. Michael Gaziano, M.D., M.P.H.
VA Boston Healthcare System, 150 S. Huntington Avenue, Boston, MA 02130
- MVP Coordinating Center, Palo Alto – Philip S. Tsao, Ph.D.
VA Palo Alto Health Care System, 3801 Miranda Avenue, Palo Alto, CA 94304
- MVP Information Center, Canandaigua – Brady Stephens, M.S.
Canandaigua VA Medical Center, 400 Fort Hill Avenue, Canandaigua, NY 14424
- Cooperative Studies Program Clinical Research Pharmacy Coordinating Center, Albuquerque – Todd Connor, Pharm.D.; Dean P. Argyles, B.S., M.S.

New Mexico VA Health Care System, 1501 San Pedro Drive SE, Albuquerque, NM 87108

MVP Publications and Presentations Committee

- Co-Chair: Themistocles L. Assimes, M.D., Ph. D
VA Palo Alto Health Care System, 3801 Miranda Avenue, Palo Alto, CA 94304
- Co-Chair: Adriana Hung, M.D.; M.P.H
VA Tennessee Valley Healthcare System, 1310 24th Ave. South, Nashville, TN 37212
- Co-Chair: Henry Kranzler, M.D.
Philadelphia VA Medical Center, 3900 Woodland Avenue, Philadelphia, PA 19104

MVP Local Site Investigators

- Samuel Aguayo, M.D., Phoenix VA Health Care System
650 E. Indian School Road, Phoenix, AZ 85012
- Sunil Ahuja, M.D., South Texas Veterans Health Care System
7400 Merton Minter Boulevard, San Antonio, TX 78229
- Kathrina Alexander, M.D., Veterans Health Care System of the Ozarks
1100 North College Avenue, Fayetteville, AR 72703
- Xiao M. Androulakis, M.D., Columbia VA Health Care System
6439 Garners Ferry Road, Columbia, SC 29209
- Prakash Balasubramanian, M.D., William S. Middleton Memorial Veterans Hospital
2500 Overlook Terrace, Madison, WI 53705
- Zuhair Ballas, M.D., Iowa City VA Health Care System
601 Highway 6 West, Iowa City, IA 52246-2208
- Jean Beckham, Ph.D., Durham VA Medical Center
508 Fulton Street, Durham, NC 27705
- Sujata Bhushan, M.D., VA North Texas Health Care System
4500 S. Lancaster Road, Dallas, TX 75216

- Edward Boyko, M.D., VA Puget Sound Health Care System
1660 S. Columbian Way, Seattle, WA 98108-1597
- David Cohen, M.D., Portland VA Medical Center
3710 SW U.S. Veterans Hospital Road, Portland, OR 97239
- Louis Dellitalia, M.D., Birmingham VA Medical Center
700 S. 19th Street, Birmingham AL 35233
- L. Christine Faulk, M.D., Robert J. Dole VA Medical Center
5500 East Kellogg Drive, Wichita, KS 67218-1607
- Joseph Fayad, M.D., VA Southern Nevada Healthcare System
6900 North Pecos Road, North Las Vegas, NV 89086
- Daryl Fujii, Ph.D., VA Pacific Islands Health Care System
459 Patterson Rd, Honolulu, HI 96819
- Saib Gappy, M.D., John D. Dingell VA Medical Center
4646 John R Street, Detroit, MI 48201
- Frank Gesek, Ph.D., White River Junction VA Medical Center
163 Veterans Drive, White River Junction, VT 05009
- Jennifer Greco, M.D., Sioux Falls VA Health Care System
2501 W 22nd Street, Sioux Falls, SD 57105
- Michael Godschalk, M.D., Richmond VA Medical Center
1201 Broad Rock Blvd., Richmond, VA 23249
- Todd W. Gress, M.D., Ph.D., Hershel “Woody” Williams VA Medical Center
1540 Spring Valley Drive, Huntington, WV 25704
- Samir Gupta, M.D., M.S.C.S., VA San Diego Healthcare System
3350 La Jolla Village Drive, San Diego, CA 92161
- Salvador Gutierrez, M.D., Edward Hines, Jr. VA Medical Center
5000 South 5th Avenue, Hines, IL 60141
- John Harley, M.D., Ph.D., Cincinnati VA Medical Center
3200 Vine Street, Cincinnati, OH 45220
- Kimberly Hammer, Ph.D., Fargo VA Health Care System
2101 N. Elm, Fargo, ND 58102
- Mark Hamner, M.D., Ralph H. Johnson VA Medical Center
109 Bee Street, Mental Health Research, Charleston, SC 29401
- Adriana Hung, M.D., M.P.H., VA Tennessee Valley Healthcare System
1310 24th Avenue, South Nashville, TN 37212
- Robin Hurley, M.D., W.G. (Bill) Hefner VA Medical Center
1601 Brenner Ave, Salisbury, NC 28144

- Pran Iruvanti, D.O., Ph.D., Hampton VA Medical Center
100 Emancipation Drive, Hampton, VA 23667
- Frank Jacono, M.D., VA Northeast Ohio Healthcare System
10701 East Boulevard, Cleveland, OH 44106
- Darshana Jhala, M.D., Philadelphia VA Medical Center
3900 Woodland Avenue, Philadelphia, PA 19104
- Scott Kinlay, M.B.B.S., Ph.D., VA Boston Healthcare System
150 S. Huntington Avenue, Boston, MA 02130
- Jon Klein, M.D., Ph.D., Louisville VA Medical Center
800 Zorn Avenue, Louisville, KY 40206
- Michael Landry, Ph.D., Southeast Louisiana Veterans Health Care System
2400 Canal Street, New Orleans, LA 70119
- Peter Liang, M.D., M.P.H., VA New York Harbor Healthcare System
423 East 23rd Street, New York, NY 10010
- Suthat Liangpunsakul, M.D., M.P.H., Richard Roudebush VA Medical Center
1481 West 10th Street, Indianapolis, IN 46202
- Jack Lichy, M.D., Ph.D., Washington DC VA Medical Center
50 Irving St, Washington, D. C. 20422
- C. Scott Mahan, M.D., Charles George VA Medical Center
1100 Tunnel Road, Asheville, NC 28805
- Ronnie Marrache, M.D., VA Maine Healthcare System
1 VA Center, Augusta, ME 04330
- Stephen Mastorides, M.D., James A. Haley Veterans' Hospital
13000 Bruce B. Downs Blvd, Tampa, FL 33612
- Elisabeth Mates M.D., Ph.D., VA Sierra Nevada Health Care System
975 Kirman Avenue, Reno, NV 89502
- Kristin Mattocks, Ph.D., M.P.H., Central Western Massachusetts Healthcare System
421 North Main Street, Leeds, MA 01053
- Paul Meyer, M.D., Ph.D., Southern Arizona VA Health Care System
3601 S 6th Avenue, Tucson, AZ 85723
- Jonathan Moorman, M.D., Ph.D., James H. Quillen VA Medical Center
Corner of Lamont & Veterans Way, Mountain Home, TN 37684
- Timothy Morgan, M.D., VA Long Beach Healthcare System
5901 East 7th Street Long Beach, CA 90822
- Maureen Murdoch, M.D., M.P.H., Minneapolis VA Health Care System
One Veterans Drive, Minneapolis, MN 55417

- James Norton, Ph.D., VA Health Care Upstate New York
113 Holland Avenue, Albany, NY 12208
- Olaoluwa Okusaga, M.D., Michael E. DeBakey VA Medical Center
2002 Holcombe Blvd, Houston, TX 77030
- Kris Ann Oursler, M.D., Salem VA Medical Center
1970 Roanoke Blvd, Salem, VA 24153
- Ana Palacio, M.D., M.P.H., Miami VA Health Care System
1201 NW 16th Street, 11 GRC, Miami FL 33125
- Samuel Poon, M.D., Manchester VA Medical Center
718 Smyth Road, Manchester, NH 03104
- Emily Potter, Pharm.D., VA Eastern Kansas Health Care System
4101 S 4th Street Trafficway, Leavenworth, KS 66048
- Michael Rauchman, M.D., St. Louis VA Health Care System
915 North Grand Blvd, St. Louis, MO 63106
- Richard Servatius, Ph.D., Syracuse VA Medical Center
800 Irving Avenue, Syracuse, NY 13210
- Satish Sharma, M.D., Providence VA Medical Center
830 Chalkstone Avenue, Providence, RI 02908
- River Smith, Ph.D., Eastern Oklahoma VA Health Care System
1011 Honor Heights Drive, Muskogee, OK 74401
- Peruvemba Sriram, M.D., N. FL/S. GA Veterans Health System
1601 SW Archer Road, Gainesville, FL 32608
- Patrick Strollo, Jr., M.D., VA Pittsburgh Health Care System
University Drive, Pittsburgh, PA 15240
- Neeraj Tandon, M.D., Overton Brooks VA Medical Center
510 East Stoner Ave, Shreveport, LA 71101
- Philip Tsao, Ph.D., VA Palo Alto Health Care System
3801 Miranda Avenue, Palo Alto, CA 94304-1290
- Gerardo Villareal, M.D., New Mexico VA Health Care System
1501 San Pedro Drive, S.E. Albuquerque, NM 87108
- Agnes Wallbom, M.D., M.S., VA Greater Los Angeles Health Care System
11301 Wilshire Blvd, Los Angeles, CA 90073
- Jessica Walsh, M.D., VA Salt Lake City Health Care System
500 Foothill Drive, Salt Lake City, UT 84148
- John Wells, Ph.D., Edith Nourse Rogers Memorial Veterans Hospital
200 Springs Road, Bedford, MA 01730

- Jeffrey Whittle, M.D., M.P.H., Clement J. Zablocki VA Medical Center
5000 West National Avenue, Milwaukee, WI 53295
- Mary Whooley, M.D., San Francisco VA Health Care System
4150 Clement Street, San Francisco, CA 94121
- Allison E. Williams, N.D., Ph.D., R.N, Bay Pines VA Healthcare System
10,000 Bay Pines Blvd Bay Pines, FL 33744
- Peter Wilson, M.D., Atlanta VA Medical Center
1670 Clairmont Road, Decatur, GA 30033
- Junzhe Xu, M.D., VA Western New York Healthcare System
3495 Bailey Avenue, Buffalo, NY 14215-1199
- Shing Shing Yeh, Ph.D., M.D., Northport VA Medical Center
79 Middleville Road, Northport, NY 11768

Supplementary Note 2: Ethics and sample descriptions

GWAS results based on some of the included cases and controls have been published previously by the International OCD Foundation (IOCDF-GC; Stewart et al., 2013, IOCDF & OCGAS et al., 2018) and the OCD Collaborative Genetics Association Study (OC GAS; Mattheisen et al., 2015, IOCDF & OCGAS et al., 2018). These data were re-analyzed for the current publication using newly matched control participants that were genotyped with the same microarrays as the cases, making up 2,828 cases and 4,887 controls. GWAS results based on a subset of the cohorts are currently available as preprints ($N_{cases} = 14,140$, $N_{controls} = 562,117$, Strom et al., 2021 and $N_{cases} = 37,015$, $N_{controls} = 948,616$, Strom et al., 2024). Of those cohorts, three (EstBB, FinnGen, iPSYCH) were updated to include additional OCD cases and controls compared to the samples in the preprints. Seven cohorts are new to this study and were not included in any of the previously published GWAS ($N_{cases} = 6,120$, $N_{controls} = 430,999$). Here we describe each individual sample that was included in the OCD meta-analysis, in alphabetical order. The header of each sample lists study identifier, study principal investigator(s) (PI(s)), country or site name, and if there has been a previous publication in connection with the data, also PubMed ID(s) or

doi. See **Supplementary Table 1** for an overview of sample sizes, number of included SNPs, reference panel, GWAS analysis tool, and Lambda1000 estimate for each individual cohort.

Ethics approvals

All relevant ethics approvals have been obtained by the respective cohort's. 23andMe: Participants provided informed consent and volunteered to participate in the research online, under a protocol approved by the external AAHRPP-accredited IRB, Ethical & Independent (E&I) Review Services. As of 2022, E&I Review Services is part of Salus IRB (<https://www.versitclinicaltrials.org/salusirb>). AGDS: All study protocols were approved by the QIMR Berghofer Medical Research Institute Human Research Ethics Committee. The protocol for approaching participants through the DHS, enrolling them in the study, and consenting for all phases of the study (including invitation to future related studies) and accessing MBS and PBS records was approved by the Ethics Department of the Department of Human Services. BioVU: The Vanderbilt University Medical Center Institutional Review Board oversees BioVU and approved this project (IRB201609). COGA: Institutional review boards at all sites approved the study and all participants provided informed consent. EGOS: Ethical approvals were obtained from the Institutional Review Board (IRB) at the Icahn School of Medicine at Mount Sinai, New York, NY, and the Regional Ethical Review Board in Stockholm. EPOC: The study was in accordance with the revised Declaration of Helsinki and approved by the local ethics committees of the Charite University Medicine Berlin and the University Hospital Bonn. EstBB: At recruitment, participants signed a consent allowing follow-up linkage of their electronic health records (EHRs), thereby providing a longitudinal collection of their phenotypic information. FinnGen: The Ethical Review Board of the Hospital District of Helsinki and Uusimaa approved the FinnGen study protocol Nr. HUS/990/2017. The FinnGen project was approved by Finnish Institute for Health and Welfare (THL), approval numbers THL/2031/6.02.00/2017, amendments THL/341/6.02.00/2018, THL/2222/6.02.00/2018, and THL/283/6.02.00/2019. HUNT: The HUNT study was approved by the Regional Committee for Medical and Health Research Ethics, Norway (2015/575). IOCDF/OCGAS: This work was approved by the relevant IRBs at all participating

sites, and all participants provided written informed consent. iPSYCH: The study was approved by the Regional Scientific Ethics Committee in Denmark and the Danish Data Protection Agency. MVP: The U.S. Department of Veterans Affairs (VA) Million Veteran Program (MVP) is collecting genetic and electronic health record (EHR) data in the U.S with ethical approval given by the Central VA Institutional Review Board (IRB) and site-specific IRBs. All relevant ethical regulations for work with human subjects were followed in the conduct of the study, and informed consent was obtained from all participants. MoBa: The establishment of MoBa and initial data collection was based on a license from the Norwegian Data Protection Agency and approval from The Regional Committees for Medical and Health Research Ethics. The MoBa cohort is now based on regulations related to the Norwegian Health Registry Act. NORDiC-SWE: This study was approved by the Regional Ethics Committee, Stockholm (EPN Stockholm) and the Institutional Review Board (IRB) at the University of North Carolina at Chapel Hill and all subjects provided informed consent. NORDiC-NOR: The NORDiC-NOR study was approved by the Norwegian Regional Committee for Medical and Health Research Ethics (IRB00001872 REK West) under project number 2018/52 REKVest (PI: Bjarne Hansen) and project number: 2014/75 REKVest (PI: Jan Haavik) and all subjects provided informed consent. OCGAS-all: Ethics approvals for the OCGAS study were obtained from the Hopkins Medicine Institutional Review Boards, the Butler Institutional Review Board, the UCLA Institutional Review Boards, the Mass General Brigham Human Research Committee, the Columbia University Institutional Review Boards, and the National Institutes of Health Institutional Review Board (NIH IRB). OCGAS-nestadt: Ethics approvals for the OCGAS study were obtained from the Hopkins Medicine Institutional Review Boards. OCGAS-ab: The study was approved by REB at Hospital for Sick Children. Ethics approvals for the OCGAS study were obtained from the Hopkins Medicine Institutional Review Boards. OCGAS-gh: Informed written consent was obtained in all cases by the participants or their parents. The study was approved by the ethical commissions of all involved universities in accordance with the latest version of the Declaration of Helsinki, including an ethical permission granted by the Ethic Committees from Aachen, Wuerzburg, Marburg, Freiburg, and the Cantonal Ethic Commission of Zuerich (Ref. Nr. 39/97, 140/3 and EK: KEK-ZHNr. 2010-0340/3). OCD-

WWF: The study was approved by the ethics committee of the University of Wuerzburg, Germany and was conducted according to the ethical principles of the Helsinki Declaration. All patients gave written informed consent prior to participation. Psych_Broad: Both studies have been approved by the Clinical Research Ethics Committee (CREC) of Hospital Universitari Vall d'Hebron. All methods were performed in accordance with the relevant guidelines and regulations and written informed consent was obtained from participant parents before inclusion into the study. UKBB: Research on the UK Biobank is conducted under a generic Research Tissue Bank approval from the UK North West Multi-centre Research Ethics Committee (MREC). This research was approved to be conducted under that approval by the governing Research Ethics Committee of the UK Biobank. The analyses in this paper were performed under an approved extension to project 16577. Yale-Penn: Participants were recruited from eastern U.S sites and provided written informed consent as approved by the institutional review board at each site.

We confirm that all necessary patient/participant consent has been obtained and the appropriate institutional forms have been archived, and that any patient/participant/sample identifiers included were not known to anyone (e.g., hospital staff, patients or participants themselves) outside the research group so cannot be used to identify individuals.

23andMe | 23andMe | USA | <https://doi.org/10.1101/2024.03.06.24303776>

Samples of European ancestry were drawn from the customer base of 23andMe Inc., a private consumer genetics company. The 23andMe cohort has been described in detail elsewhere (Hyde et al., 2016). Participants provided informed consent and volunteered to participate in the research online, under a protocol approved by the external AAHRPP-accredited IRB, Ethical & Independent (E&I) Review Services. As of 2022, E&I Review Services is part of Salus IRB (<https://www.versiticlinicaltrials.org/salusirb>). The cohort was selected from participant data

available on 1st September 2017 and the analysis was reviewed and approved by a private institutional review board (www.eandireview.com). All individuals identified as cases have reported being diagnosed with OCD, while controls have reported not having been diagnosed with OCD. 30,167 OCD cases and 929,804 controls were included in the final GWAS. Of the cases, 10,808 (35.83%) were male and 19,359 (64.17%) were female. Of the controls, 443,336 (47.68%) were male and 486,468 (52.32%) were female. Of the cases, 7,040 (23.34%) individuals were below the age of 30, 11,087 (36.75%) between 30 and 45, 7,056 (23.39%) between 45 and 60, and 4,984 (16.52%) above the age of 60. Of the controls, 91,287 (9.82%) individuals were below the age of 30, 230,696 (24.81%) between 30 and 45, 257,793 (37.64%) between 45 and 60, and 350,028 (27.73%) above the age of 60. Extraction of DNA and genotyping was performed by the National Genetics Institute (NGI), a CLIA licensed clinical laboratory and a subsidiary of Laboratory Corporation of America. Individuals were genotyped on four different genotype platforms. Two (V1, V2) platforms were variants of the Illumina HumanHap550+ BeadChip, including 25,000 custom SNPs, one platform (V3) was the Illumina OmniExpress+ BeadChip, with custom SNPs to increase overlap with V2, and one platform (V4) is in current use and a fully customized array. Individuals that failed to meet a 98.5% call rate were re-analyzed. Only individuals with > 97% European ancestry were included in the analysis. European ancestry was determined through an analysis of local ancestry using a support vector machine to classify individual haplotypes into one of 31 reference populations. Those classifications were then fed into a hidden Markov model (HMM) which accounts for incorrect assignments and switch errors, thereby giving probabilities for each reference population per window (100 SNPs). Simulated admixed individuals were then used to re-calibrate the HMM probabilities so the assigned ancestries were consistent with the simulated individuals. Publicly available datasets (Human Genome Diversity Project, HapMap (Altshuler et al., 2010), and 1000 Genomes (The 1000 Genomes Project Consortium, 2015)) and 23andMe customers with four grandparents from the same country served as the reference population. Identity-by-descent (IBD) estimation was used to define a maximal set of unrelated individuals for each analysis. Individuals who shared less than 700 cM IBD were defined as unrelated (e.g., approximately less related than first cousins).

The merged UK10K and 1000 Genomes Phase 3 panel was used for imputation. Finch, an internally developed tool that implements the Beagle haplotype graph-based phasing algorithm, modified to separate the steps of graph construction and phasing, was used for phasing. Imputation was then performed as an estimated allele dosage averaged over a set of possible imputed haplotypes for each individual. Genetic association testing was performed using logistic regression, assuming an additive model for allelic effects, including age, sex, first five principal components (PCs), and the genotype platform as covariates. For imputed data, the imputed dosages rather than the best-guess genotypes were used. The association p-value was computed using a likelihood ratio test.

AGDS | Martin, N. | QIMR, Brisbane, Australia

The Australian Genetics of Depression study (AGDS) was established to recruit a large cohort of individuals who have been diagnosed with depression at some point in their lifetime. The purpose of establishing this cohort is to investigate genetic and environmental risk factors for depression and response to commonly prescribed antidepressants. All cases and controls were diagnosed with Major Depressive Disorder. OCD cases were identified based on self-reported clinical diagnosis for OCD while OCD controls had no self-reported clinical diagnosis for OCD and obtained a score below 10 on the Obsessive-Compulsive Inventory-Revised (OCI-R). A total of 20,689 participants were recruited through the Australian Department of Human Services and a media campaign, 75% of whom were female. Participants were recruited to the Australian Genetics of Depression Study (www.geneticsofdepression.org.au) between 2016 and 2019. All study protocols were approved by the QIMR Berghofer Medical Research Institute Human Research Ethics Committee. The protocol for approaching participants through the DHS, enrolling them in the study, and consenting for all phases of the study (including invitation to future related studies) and accessing MBS and PBS records was approved by the Ethics Department of the Department of Human Services. The average age of participants was 43 years \pm 15 years. Participants completed an online questionnaire that consisted of a compulsory module that assessed self-reported psychiatric history, clinical depression using the Composite Interview

Diagnostic Interview Short Form, and experiences of using commonly prescribed antidepressants. Further voluntary modules assessed a wide range of traits of relevance to psychopathology. Participants who reported they were willing to provide a DNA sample (75%) were sent a saliva kit in the mail. In the present study, we included 757 cases and 5,368 controls for whom genotype data were available. Samples from the AGDS were genotyped in three different genotyping centers using the same array (GSAMD-24v1-0_20011747). Genotype calling was performed with GenomeStudio. A common set of high QC markers between the different genotyping batches was obtained prior to joint imputation. Marker exclusion criteria (prior to imputation) included: unknown or ambiguous map position and strand alignment in a BLAST search, missingness > 5%, HWE test $P < 1 \times 10^{-6}$, MAF < 1%, GenTrain score < 0.6. The Michigan imputation server (Das et al., 2016) was used to impute the genotypes using the HRCr1.1 as a reference panel. Individuals were excluded based on a high missingness (missing rate > 3%), inconsistent (and unresolvable) sex, or if deemed ancestry outliers from the European population (6 standard deviations from the first two genetic PCs). The GWAS was done employing a logistic regression using PLINK 1.9 (Purcell et al., 2007) and imputed dosage genotypes while correcting for the genotyping center and the first twenty ancestry PCs as covariates.

BioVU | Davis, L. K. | Nashville, Tennessee, USA

Vanderbilt University Medical Center (VUMC) is a tertiary care center that provides inpatient and outpatient care in Nashville, TN. The VUMC electronic health record (EHR) system was established in 1990 and includes data on billing codes from the International Classification of Diseases, 9th and 10th editions (ICD-9 and ICD-10), Current Procedural Terminology (CPT) codes, laboratory values, reports, and clinical documentation. In 2007, VUMC launched a biobank, BioVU, which links a patient's DNA sample to their EHR. The BioVU Consent form is provided to patients in the outpatient clinic environments at VUMC. The form states policies on data sharing and privacy, and should a signature be obtained, makes any blood leftover from clinical care eligible for BioVU banking. The Vanderbilt University Medical Center Institutional Review Board oversees BioVU and approved this project (IRB201609). OCD case status was

determined through a combination of ICD codes, medications, and natural language processing for EHR notes. Using ICD codes, cases were defined as individuals with any codes for OCD (ICD9: 300.3, ICD10: F42, F63.3, F45.22). Additional cases were gathered by first finding individuals with “obsessive-compulsive”, “obsessive compulsive”, or “OCD” in clinic notes, problem lists, discharge summaries, or clinical communications, where the first instance occurred before 55 years of age. Mentions of “OCD” were excluded if the keyword “osteochondritis” or “osteochondritis” occurred in the same document, or if certain ICD codes related to pervasive developmental disorders, osteochondritis dissecans, and bariatric surgery status (ICD9: 299*, 732.7, 649.2, 649.20, 649.21, 649.23, 649.24, V45.86, ICD10: F84*, M93.20, O99.840, O99.841, O99.842, O99.843, O99.844, O99.845, Z98.84) occurred on the same date as the mention. Keyword mentions were excluded if negating terms or terms related to familial relationships occurred within 30 characters on either side of the mention. Next, these individuals were required to have OCD medication or cognitive behavioral therapy in their EHR. Finally, individuals with metabolic disorder codes (ICD9: 277.89, 277.9, ICD10: E88.9, E88.89) were excluded. Controls were defined as any individual without OCD codes, metabolic disorder codes, “obsessive-compulsive”, “obsessive compulsive”, or “OCD” in their clinical notes, and without evidence of OCD medication. The initial sample included 1062 cases (median age in years across EHR: 32.75, 58.6% female) and 40,316 controls (median age in years across EHR: 53.57% female). We obtained genotype information on 94,474 BioVU individuals genotyped on the Illumina MEGA EX array. Using PLINK v1.95 (Purcell et al., 2007), genotypes were filtered for SNP and individual call rates, sex discrepancies, and excessive heterozygosity. We selected individuals of European ancestry using principal component analysis (PCA) implemented in FlashPCA (Abraham & Inouye, 2014) and confirmed the absence of genotyping batch effects through logistic regression with batch as the phenotype. Autosomes were imputed to the HRC panel using Michigan Imputation Server (Das et al., 2016) in five batches. After imputation, genotypes were converted to hard calls with PLINK using the default threshold settings. SNPs with multiple alleles or imputation quality less than R^2 of 0.3 were excluded. Next, SNPs with minor allele frequency (MAF) less than 0.005 or genotyping rates less than 0.98 were excluded. Individuals with call

rates less than 0.98 were excluded. We ran a series of PCA to determine BioVU individuals of European genetic ancestry. First, we performed PCA using FlashPCA on BioVU combined with CEU, YRI, and CHB reference sets from 1000 Genomes Project Phase 3 (The 1000 Genomes Project Consortium, 2015). PCs were scaled so that the axes could be interpreted as proportions of genetic ancestry. We selected BioVU individuals who were within 40% of the CEU cluster along the CEU-CHB axis and within 30% of the CEU cluster on the CEU-YRI axis, generating a once-PCA filtered European set. To ensure subsequent steps would remove SNPs associated with reduced quality rather than cryptic population substructure, we filtered the previously identified BioVU European cluster to identify individuals falling within the CEU, TSI, and GIH 1000 genomes populations, producing a twice filtered European set. Using the twice-filtered European set we conducted a series of SNP checks. First, we filtered individuals with IBS greater than 0.2 and calculated PCs to use as covariates. Next, we checked for imputation batch effects by conducting pairwise logistic regression of the five imputation batches using sex and top 10 PCs as covariates. SNPs with p-values less than 0.001 in the additive model were flagged. We then compared MAF between BioVU and the CEU reference population. Any SNPs with a MAF difference greater than 0.1 were flagged. SNPs with a Hardy-Weinberg Equilibrium (HWE) p-value less than 1×10^{-10} were flagged. Finally, the flagged SNPs from the batch effect, MAF difference, and HWE were excluded from the once-PCA filtered BioVU European set, resulting in 9,386,383 SNPs for analysis. The final dataset consisted of 1041 cases and 38,613 controls. To account for the large case-control imbalance, we used SAIGE (Zhou et al., 2018) for the GWAS. Covariates included were sex, median age across medical records, and top 10 PCs.

Children's Hospital of Philadelphia (CHOP) | Hakonarson, H; Gur, R. E. | USA | 25840117, 2588255

Data were collected in the Philadelphia Neurodevelopmental Cohort (PNC) from the Children's Hospital of Philadelphia (CHOP) (Satterthwaite et al., 2016). 9428 participants aged 8-21 years completed a computerized structured screener based on the Kiddie-Schedule for Affective

Disorders and Schizophrenia (K-SADS) called GO-ASSESS (Calkins et al., 2014, 2015), which asked about the lifetime presence of any obsessive or compulsive symptoms as well as the severity, level of impairment and age of onset of symptoms. Samples were genotyped using Illumina Human610-Quadv1_B BeadChip array. As previously described (Burton et al., 2021), QC filtering was done using standard methods including removing SNPs with MAF < 0.01, and imputation quality < 0.6. GWAS was conducted using R (v3.5.1) with a logistic regression and using age, sex, and 3 PC scores as covariates.

COGA | Porjesz, B.; Foroud, T.; Agrawal, A. | USA | 31270906, 31090166

Participants of the Collaborative Study on the Genetics of Alcoholism (COGA) were recruited from 7 sites across the U.S. Institutional review boards at all sites approved the study and all participants provided informed consent. OCD diagnosis was determined using DSM-III-R or DSM-IV. COGA European ancestry (EA) data were genotyped using three arrays: Illumina Human1M array, Illumina Human OmniExpress V1 array, Smokescreen genotyping array (two different batches) (Lai et al., 2019, 2020). To confirm pedigree structure and estimate PCs of population stratification, we used a set of independent (defined as linkage disequilibrium (LD) $r^2 < 0.5$) and high-quality variants (MAF > 10%, HWE p-value > 0.001, missing rate < 2%, 47,000 variants) genotyped on all arrays. Pairwise identity by descent was computed using PLINK (Chang et al., 2015; Purcell et al., 2007) to confirm family relationships and family structures were updated if necessary. PCs were calculated using Eigenstrat (Price et al., 2006). Based on the first two PCs,

each individual was assigned an ancestry (i.e., African ancestry, European ancestry, or other) if they clustered with the corresponding 1000 Genomes project populations. Before imputation, Mendelian errors were detected using Pedcheck (O'Connell & Weeks, 1998) and inconsistencies were set to missing. Palindromic variants were excluded to avoid strand ambiguities. Then variants with missing rates > 5%, MAF < 3%, and HWE P values < 0.0001 were excluded. SHAPEIT2 (Delaneau et al., 2013) was used to phase haplotypes of each sample then Minimac 3.32 (Das et al., 2016) was used for imputation. 1000 genomes project samples were used as the reference panel (Phase 3, version 5; (The 1000 Genomes Project Consortium, 2015)). Samples on each array were imputed separately due to the different array content. After imputation, variants with $R^2 < 0.60$ were excluded. Genotype probabilities were converted to genotypes if they were > 0.90, then Mendelian error checking was performed again. All genotype and imputed variants with missing rates less than 20%, $MAF \geq 1\%$, and HWE P values > 0.000001 were kept in analyses. Unrelated samples selected from COGA EA datasets were included in analysis (131 cases and 1358 controls) using PLINK. Sex, age, the first 10 PCs, and array indicator were included as covariates.

EGOS | Grice, D. E. | Sweden | 31907560

The EGOS source population consists of individuals born in Sweden between January 1954 and December 1998 that presented at least two diagnoses of OCD or chronic tic disorders (CTD) at

different time points in the Swedish National Patient Register (NPR) and were followed between January 1997 and December 2012 (N = 20,374). The International Classification of Disease (ICD) was used for the identification of OCD cases (F49 ICD version 10). Detailed information for each individual was obtained through linkage to the Swedish national registers, e.g., family relatedness, identification of additional psychiatric diagnoses, medical diagnoses, birth-related variables, and relevant demographic and social data. To create an epidemiologically valid subset of the source cohort that also includes biospecimens and additional phenotyping, individuals were contacted from within the source population. Study participants could elect whether to donate blood or saliva. Individuals are aged 16-64 years. All samples were genotyped using the Global Screening Array (GSA). Further description of the EGOS cohort has been represented by Mahjani et al. (Mahjani et al., 2020) The dataset that was used in this meta-analysis consists of 1026 OCD cases from EGOS and 1208 controls from LifeGene. EGOS was supported by a grant from the Beatrice and Samuel A. Seaver Foundation to DEG. Ethical approvals were obtained from the Institutional Review Board (IRB) at the Icahn School of Medicine at Mount Sinai, New York, NY, and the Regional Ethical Review Board in Stockholm.

EPOC | Wagner, M.; Kathmann, N. | Germany | 30744714, 30008679, 29890378, 29721727, 29159055, 28541065, 28481032, 28160276

The EPOC (Endophenotypes of OCD) sample comprises (epi)genetic and deep phenotype data of OCD patients, unaffected first-degree relatives of OCD patients and healthy controls that were collected at two sites in Germany (Berlin and Bonn) between 2014 and 2017. In the present analysis, data from 195 patients with OCD (55.9 % female, 44.1 % male) and 204 controls (63.2 % female, 36.8 % male) were included. Mean age was 33.37 (SD = 10.76; range: 18–64) for patients and 34.72 (SD = 12.64; range: 18–64) for controls. Lifetime comorbidity rates of patients were 60.0% for depression; 11.3% for panic disorder or agoraphobia; 7.7% for tic disorder; 7.7% for specific phobia; 6.7% for social phobia; 4.6% for generalized anxiety disorder; 4.1% for PTSD; 4.1% for anorexia; 3.1% for hypochondria; and 2.6% for hoarding disorder. OCD patients were

recruited via the outpatient clinics at the Department of Psychology of Humboldt University, Berlin and at the Department of Psychiatry and Psychotherapy of the University Hospital Bonn. Healthy volunteers were recruited from the general population via public advertisements in the same cities. All participants were examined by trained clinical psychologists using the Structured Clinical Interview for DSM-IV (SCIDI) to assess OCD diagnosis and potential comorbidities. To establish cross-site reliability of clinical ratings, all instructions were standardized, and raters completed assessments of four training videos. Patients were only included if they: (a) met diagnostic criteria of OCD based on the SCID-I interview; (b) were free of past or present psychotic, bipolar or substance-related disorders; (c) did not take neuroleptic medication for the previous four weeks; and (d) did not use benzodiazepines in the prior two weeks. Healthy controls were excluded if they: (a) took any psychoactive medication in the previous three months; (b) had a current Axis I disorder; (c) had a lifetime diagnosis of OCD or tic disorder; or (d) had a family history of OCD. Written informed consent was obtained and participants were compensated for their time. The study was in accordance with the revised Declaration of Helsinki and approved by the local ethics committees of the Charité University Medicine Berlin and the University Hospital Bonn. Genotyping of cases and controls was performed on the Illumina Global Screen Array (GSA) at the Life Brain Center, Bonn. Genotype quality control was done using Plink-1.9 (Purcell et al., 2007), and R (version 3.5.1). We checked the data for sex inconsistencies and grossly failing markers (call rate < 0.5). Individuals with a call rate of < 0.95 were removed. The heterozygosity rate for each subject was calculated; outliers (± 3 SD from the mean heterozygosity rate) were identified and removed. On marker level, SNPs were removed if at least one of the following conditions was true: significant difference of missing rate between cases and controls, call rate < 0.95; deviation of HWE ($P < 1 \times 10^{-6}$); and MAF < 0.05 (computed separately in cases and controls). Furthermore, all A/T or C/G SNPs were removed. To check for population stratification, PCA was performed, and the first two PCs were checked for outliers. Seven individuals, whose non-European ethnicity was additionally validated based on their demographic data, were excluded. The genotyped data were imputed on the Michigan Imputation Server (Das et al., 2016) using the 1000 Genomes Phase 3 (version 5) reference panel (The 1000 Genomes Project

Consortium, 2015). GWAS was conducted using SNPTEST (version 2.5.2) (Marchini et al., 2007) with the first four PCs as covariates.

EstBB | Metspalu, A. | Estonia

The Estonian Biobank (EstBB) is a population-based cohort with a rich variety of phenotypic and health-related information collected for each participant (Leitsalu et al., 2015). At recruitment, participants signed a consent allowing follow-up linkage of their electronic health records (EHRs), thereby providing a longitudinal collection of their phenotypic information. The EstBB database includes health records from the national Health Insurance Fund Treatment Bills (from 2004 onwards), Tartu University Hospital (from 2008), and North Estonia Medical Center (from 2005), and data from different registries (causes of death, cancer, etc.). For all the participants EstBB provides information on the diagnoses in ICD-10 coding and information on drug dispensing data, including drug ATC codes, prescription status and purchase date (if available).

Genotyping of DNA samples from the Estonian Biobank was done at the Core Genotyping Lab of the Institute of Genomics, University of Tartu using the Illumina Global Screening Arrays (GSAv1.0, GSAv2.0, and GSAv2.0_EST). Samples were genotyped and then PLINK format files were created using Illumina GenomeStudio v2.0.4. During the quality control all individuals with call-rate < 95% or mismatching sex that was defined based on the heterozygosity of X chromosome and sex in the phenotype data, were excluded from the analysis. Variants were filtered by call-rate < 95% and HWE p-value < 1×10^{-04} (autosomal variants only). Variant positions were updated to Genome Reference Consortium Human Build 37 and all variants were changed to be from TOP strand using reference information provided by Dr. Will Rayner from the University of Oxford (<https://www.well.ox.ac.uk/~wrayner/strand/>). After QC the dataset contained 202,910 samples for imputation. Before imputation variants with MAF < 1% and Indels were removed. Prephasing was done using the Eagle v2.3 software (Loh, Danecek, et al., 2016; Loh, Palamara, et al., 2016). The number of conditioning haplotypes Eagle2 uses when phasing each

sample was set to: $-Kpbwt=20000$. Imputation was done using Beagle v.18May20.d20 (B. L. Browning et al., 2018; S. R. Browning & Browning, 2007) with effective population size $N_{Effective} = 20,000$. As a reference, Estonian population specific imputation reference of 2297 WGS samples was used (Mitt et al., 2017). Further, EstBB samples were combined with the 1000 genomes phase 3 dataset for ancestry analysis. Genetic principal components were calculated using a subset of quality controlled and pruned genotyped SNPs. This was further used to identify and remove samples that deviated from the main cluster.

For the genome-wide study of OCD, EstBB cases were defined as participants with F42* ICD10 diagnosis codes in their EHRs. We conducted a GWAS on individuals of European ancestry, including 772 cases and 196,079 controls. The analysis was performed with the REGENIE software (2.2.4) (Mbatchou et al., 2021) including related individuals and adjusting for the first 10 PCs of the genotype matrix, as well as for birth year, birth year squared and sex.

FinnGen | Kaprio, J. A. | Finland

Finnish samples are population-based biobank samples collected between August 2017 and August 2019 (collection further ongoing), including legacy samples collected since the 1980s. The Ethical Review Board of the Hospital District of Helsinki and Uusimaa approved the FinnGen study protocol Nr. HUS/990/2017. The FinnGen project was approved by Finnish Institute for Health and Welfare (THL), approval numbers THL/2031/6.02.00/2017, amendments THL/341/6.02.00/2018, THL/2222/6.02.00/2018, and THL/283/6.02.00/2019. The here included data has not been published before but a general description of the FinnGen study can be found elsewhere (Mars et al., 2020). Cases were recruited through hospital records (inpatient and outpatients) from 1970 onwards, based on clinical diagnoses used in patient care, supplemented by cause of death diagnoses (if clinical diagnosis was underlying and/or contributing cause). Case definitions were based on a diagnosis of ICD-10 F42 and ICD-8 3003, there were no case exclusion criteria. The mean age of cases was 45.36 ($SD = 15.44$) years with a minimum age of

16.5 and a maximum age of 91.4, including 330 males and 460 females. Controls were defined as all other participants in FinnGen with GWAS data. The mean age of controls was 60.02 ($SD = 17.28$) years with a minimum age of 0 and a maximum age of 105.73, including 71,252 males and 90,138 females. Genotyping of cases and controls was performed in multiple batches on multiple arrays in multiple centers. For genotype quality control the following filters were applied to the data: exclusion of chromosome X, exclusion of variants with INFO score < 0.95 , with missingness > 0.01 , or with MAF < 0.05 . To check for population stratification, PCA was performed, and a Bayesian algorithm was used to spot outliers. Individuals were excluded if there was a mismatch between imputed sex and sex in registry data. For calculating the genetic relationship matrix, we used the genotype dataset where genotypes with GP < 0.95 have been set missing. Only variants imputed with an INFO score > 0.95 in all batches were used. Variants with $> 3\%$ missing genotypes were excluded as well as variants with MAF $< 1\%$. The remaining variants were LD pruned with a 1Mb window and r^2 threshold of 0.1. Imputation was performed using Eagle 2.4/Beagle 4.1 (B. L. Browning, 2017; Loh, Danecek, et al., 2016) using Finnish WGS (depth up to 30x) samples as reference with a total amount of 16962023 variants. GWAS was performed using SAIGE (v0.35.8.8) (Zhou et al., 2018) including age, sex, the first 10 PCs, genotyping batch and kinship matrix as covariates.

HUNT | Zwart, J.-A. | Norway | 22879362

The Trøndelag Health Study (HUNT) consists of three different population-based health surveys conducted in the county of Nord-Trøndelag, Norway over approximately 20 years (HUNT1: 1984-1986, HUNT2: 1995-1997 and HUNT3: 2006-2008). The HUNT study was approved by the Regional Committee for Medical and Health Research Ethics, Norway (2015/575). For each survey, the entire adult population (≥ 20 years) was invited to participate by completing questionnaires, attending clinical examinations and interviews. Participation rates in HUNT1,

HUNT2 and HUNT3 were 89.4% (N = 77,212), 69.5% (N = 65 237) and 54.1% (N = 50 807), respectively. Taken together, the study included more than 120,000 different individuals from NordTrøndelag County. Biological samples including DNA have been collected for approximately 70000 participants. The entire HUNT Study has been described in more detail elsewhere (Krokstad et al., 2013). For the present study, we included participants from HUNT2 and HUNT3. Cases and controls were defined by linkage to hospital diagnostic registries from the time period 1987-2017. Cases were defined as those with a hospital diagnosis of obsessive-compulsive disorder (ICD10 code F42). Controls were defined as those without ICD10 code F42 and without ICD-9 code 300 (“anxiety, dissociative and somatoform disorders”). The study was approved by the Regional Committee for Medical and Health Research Ethics (ref. 2015/575). In total, DNA from 71,860 HUNT samples was genotyped at the Genomics Core Facility at the Norwegian University of Science and Technology using one of three different Illumina HumanCoreExome arrays (HumanCoreExome12 v1.0, HumanCoreExome12 v1.1 and UM HUNT Biobank v1.0). Samples that failed to reach a 99% call rate, had contamination >2.5% as estimated with BAF Regress (Jun et al., 2012), large chromosomal copy number variants, lower call rate of a technical duplicate pair and twins, gonosomal constellations other than XX and XY, or whose inferred sex contradicted the reported gender, were excluded. Samples that passed quality control were analyzed in a second round of genotype calling following the Genome Studio quality control

protocol described elsewhere (Guo et al., 2014). Genomic position, strand orientation and the reference allele of genotyped variants were determined by aligning their probe sequences against the human genome (Genome Reference Consortium Human genome build 37 and revised Cambridge Reference Sequence of the human mitochondrial DNA; <http://genome.ucsc.edu>) using BLAT (Dunham et al., 2012). PLINK v1.90 (Purcell et al., 2007) was then used to exclude variants if their probe sequences could not be perfectly mapped, cluster separation was < 0.3 , GenTrain score < 0.15 , showed deviations from HWE in unrelated samples of European ancestry with $p\text{-value} < 0.0001$, had a call rate $< 99\%$, or another assay with higher call rate genotyped the same variant. Ancestry of all samples was inferred by projecting all genotyped samples into the space of the PCs of the Human Genome Diversity Project (HGDP) reference panel (938 unrelated individuals; downloaded from <http://csg.sph.umich.edu/chaolong/LASER/>) (Li et al., 2008; Wang et al., 2014) using PLINK. Recent European ancestry was defined as samples that fell into an ellipsoid spanning exclusively European populations of the HGDP panel. The different arrays were harmonized by reducing to a set of overlapping variants and excluding variants that showed frequency differences $> 15\%$ between data sets, or that were monomorphic in one and had $\text{MAF} > 1\%$ in another data set. The resulting genotype data were phased using Eagle2 v2.3 (Loh, Danecek, et al., 2016). Imputation was performed on the 69,716 samples of recent European ancestry using Minimac3 (v2.0.1, <http://genome.sph.umich.edu/wiki/Minimac3>, Das et al., 2016) with default settings (2.5 Mb reference based chunking with 500kb windows) and a customized Haplotype Reference consortium release 1.1 (HRC v1.1) for autosomal variants and HRC v1.1 for chromosome X variants (McCarthy et al., 2016). The customized reference panel represented the merged panel of two reciprocally imputed reference panels: (a) 2201 low-

coverage whole genome sequences samples from the HUNT study and (b) HRC v1.1 with 1023 HUNT WGS samples removed before merging. We excluded imputed variants with $R_{sq} < 0.3$ or minor allele count < 1 , resulting in 24.2 million well-imputed variants. After restricting to those with available phenotype information 66,476 individuals (284 cases and 66,192 controls) were included in the analysis. Association analyses were conducted using SAIGE (Zhou et al., 2018), a generalized mixed effects model approach, to account for cryptic population structure and relatedness when modeling the association between genotype probabilities (dosages) and OCD. Models were adjusted for sex, birth year, genotyping batch and four PCs. PCs were computed using PLINK.

IOCDF-GC and 610k_trio | Multiple | Multiple | 22889921, 28761083

The results of the International OCD Foundation-Genetics Consortium (IOCDF-GC) study was previously published (Arnold et al., 2018; Stewart et al., 2013). The IOCDF-GC case-control cohort (IOCDF-GC) consists of 1519 European ancestry cases and 3541 matching controls from IOCDF-GC and three cohorts previously genotyped, including the Alzheimer's Disease Genetics Initiative (Lee et al., 2008), the Center for Applied Genomics (CAG) at Children's Hospital of Philadelphia (CHOP) (Gur et al., 2012), and the Breast and Prostate Cancer Cohort Consortium (BPC3; (Schumacher et al., 2011)) The IOCDF-GC trio sample (610k_trio) consists of 323 European ancestry complete trios. All cases and trios were recruited predominantly from OCD specialty clinics, and controls were recruited from Bonn, Germany and from Cape Town, South Africa. This work was approved by the relevant IRBs at all participating sites, and all participants provided written informed consent. For study inclusion, all cases and trio probands were required to have a DSM-IV diagnosis of OCD. The controls from Bonn had an absent lifetime history of all axis I disorders and the South African controls were diagnostically unscreened. The sample description of the three cohorts refer to the primary article. All samples were genotyped on Illumina Human610-Quadv1_B SNP array (Illumina, San Diego, CA, USA). Standard quality control (QC) protocol was conducted with PLINK (Purcell et al., 2007). Samples were removed for call rates $< 98\%$, sex discrepancy and ambiguous genomic sex, related samples with $pi_{hat} > 0.2$. SNP QC

included removing monomorphic SNPs, CNV-targeted SNP probes, SNPs with genotyping rate < 98%, SNPs with MAF < 0.01, strand-ambiguous SNPs with significant allele frequency differences or aberrant LD correlations with adjacent SNPs based on the entire HapMap2 reference panel, SNPs with $P < 1 \times 10^{-6}$ in HWE test among controls or $P < 1 \times 10^{-10}$ among cases, SNPs with differential missing rate between cases and controls (> 0.02), and SNPs with batch effect ($P < 1 \times 10^{-5}$) between different control cohorts. Multidimensional scaling (MDS) analyses were performed in PLINK2 (Chang et al., 2015) and samples were removed when they were significant outliers in the first five MDS dimensions or when there were no matching cases or controls on these MDS dimensions. The genotyped data were phased by SHAPEIT2 (Delaneau et al., 2013) and imputed by Minimac3 (Das et al., 2016) using the HRC (McCarthy et al., 2016) release 1.1 as the reference panel. GWAS was performed on SNPs with INFO score > 0.8 and MAF > 0.01 , using a logistic regression model in PLINK2 (Chang et al., 2015) with the first five and the 7th MDS components as covariates. Same QC was conducted on the 323 trios, with an additional filter of removing SNPs with Mendelian errors. In each trio, the transmitted alleles and untransmitted alleles were converted into one case and one pseudo-control. Phasing and imputation were conducted on the cases and pseudo-controls in the same way as the case-control cohort. GWAS was performed without covariates.

iPSYCH | Borglum, A.D.; Mors, O.; Mattheisen, M. | Denmark

In the scope of the Danish OCD and Tourette Study (DOTS) within The Lundbeck Foundation Initiative for Integrative Psychiatric Research (iPSYCH), Danish nation-wide population-based case-cohort samples were collected and genotyped. The study was approved by the Regional Scientific Ethics Committee in Denmark and the Danish Data Protection Agency. All analyses of the samples were performed on the secured national GenomeDK high performance computing cluster in Denmark (<https://genome.au.dk>). Samples stem from the newly updated baseline cohort iPSYCH2015 including singletons born between 1981 and 2008 who were born to a known mother and resided in Denmark on their first birthday. Genetic information was obtained by the

Statens Serum Institut (SSI) at the Danish Neonatal Screening Biobank (DNSB) from heel prick blood samples that had been collected from all newborn babies in Denmark. The genetic information was linked with the Danish Civil Registration System and thereby coupled with the Danish Psychiatric Central Research Register which collects patient data of individuals treated in psychiatric hospitals (from 1969 onwards) or in outpatient psychiatric clinics (from 1995 onwards). See (Bybjerg-Grauholm et al., 2020; C. B. Pedersen et al., 2018) for a detailed description of the overall cohort, array genotyping and quality control. Cases included in the present study were not primarily ascertained for OCD; OCD cases were drawn from cases that also presented a diagnosis of one of the core disorders iPSYCH primarily collected for and from randomly ascertained population “controls” with a diagnosis of OCD. OCD cases were diagnosed by a healthcare professional and met ICD10 (F42) criteria, controls were randomly from the controls and excluded individuals with an F42 diagnosis. Samples for iPSYCH2012 were genotyped on the PsychChip v 1.0 array (Illumina, San Diego, CA, USA), while samples for iPSYCH2015i were genotyped on the Illumina Global Screening (GSA) v2 Array, both at the Broad Institute of MIT and Harvard (Cambridge, MA, USA). Genotype calling of markers with MAF > 0.01 was performed by merging call sets from GenCall (Illumina: Illumina GenCall Data Analysis Software) and Birdseed (Korn et al., 2008), and less frequent variants were called with zCall (Goldstein et al., 2012). Genotyping and data analysis was performed in 23 waves. Genotype data were processed using the Rapid Imputation and COmputational PIpeLine for Genome-Wide Association Studies (RICOPILI; (Lam et al., 2020) to perform stringent QC, imputation, PC analysis, and primary association analysis. SHAPEIT (Delaneau et al., 2011) was used for phasing, imputation was conducted with IMPUTE2 (Howie et al., 2009), using the HRC as a reference panel (McCarthy et al., 2016). We removed samples with a call rate below 95%, with a sex mismatch, between the sex obtained from genotype data and from the register data, as well as related individuals. PC analysis was used to exclude ancestral outliers of non-European descent, excluding all individuals exceeding eight standard deviations from the mean on the first three PCs. The GWAS was performed using RICOPILI and included 10 PCs as covariates. The final dataset included 4509

OCD cases and 38,392 controls, 7,544,427 SNPs, had a Lambda of 1.074 and a Lambda1000 of 1.009.

Michigan/Toronto OCD Imaging Genomics Study | Noam Soreni; Gregory L.

Hanna; Kate D. Fitzgerald; David Rosenberg; Paul D. Arnold | USA and Canada

Individuals from the Michigan/Toronto OCD Imaging Genomics Study (Michigan/Toronto IGS) were recruited from four different academic child psychiatry sites: The Hospital for Sick Children, McMaster University, University of Michigan, and Wayne State University (Gazzellone et al., 2016). Cases were defined by a clinical investigator according to DSM-IV criteria. Samples were genotyped using three different arrays: HumanCoreExome, PsychArray and Omni2.5. Each genotyping array was processed separately; then imputed data from all arrays were combined. As previously described (Burton et al., 2021), QC filtering was done using standard methods including removing SNPs with MAF < 0.01, and imputation quality < 0.6. The GWAS was conducted using R (v3.5.1) with a logistic regression and using age, sex, and 3 PCs as covariates.

Million Veteran Program | Stein, M.B.; Gelernter, J. | USA | 26441289

The U.S. Department of Veterans Affairs (VA) Million Veteran Program (MVP) is collecting genetic and electronic health record (EHR) data in the U.S with ethical approval given by the Central VA Institutional Review Board (IRB) and site-specific IRBs (Gaziano et al., 2016; Harrington et al., 2019). All relevant ethical regulations for work with human subjects were followed in the conduct of the study, and informed consent was obtained from all participants. We used data release version 4; briefly, genotyping was performed with Affymetrix Axiom Biobank Array and quality control (Hunter-Zinck et al., 2020), phasing chromosomes with EAGLE2 (Loh, Danecek, et al., 2016) and imputation with Minimac3 (Das et al., 2016), using the 1000 Genomes Project reference panel, phase 3 (The 1000 Genomes Project Consortium, 2015). We selected European (EUR) ancestry data which were defined using PCA (Gaziano et al., 2016). For QC filtering, we set up imputation quality scores > 0.6, HWE filtering > 5×10^{-05} , MAF > 0.001, missing call rates

for variants < 0.1 , and missing call rates for samples < 0.1 . Data were aligned to the GRCh37 reference genome. Cases were defined having at least one International Classification of Disease (ICD) outpatient code for OCD (ICD-9: 300.3; ICD-10: F42.2, F42.3, F42.8, F42.9) by any provider, excluding individuals suffering from schizophrenia. To optimize the number of cases to keep when removing individuals with kinship, which was calculated by KING (Manichaikul et al., 2010) for a minimum threshold of 0.0884 corresponding to a second-degree relationship, we implemented the following algorithm where cases have priority compared to controls: between a case and a control, we remove the individual as control; if there is relatedness between two cases, we remove the one that has the highest number of relationships with other individuals; we do the same between two controls. In conclusion, we obtained 5129 cases and 422,860 controls. GWAS was performed using PLINK 2.0 (Chang et al., 2015) setting age, sex, and the first ten PCs scores as covariates.

MoBa | Ask, H. | Norway | 27063603

The Norwegian mother, father and child cohort study (MoBa) is a population-based pregnancy cohort study conducted by the Norwegian Institute of Public Health. The establishment of MoBa and initial data collection was based on a license from the Norwegian Data Protection Agency and approval from The Regional Committees for Medical and Health Research Ethics. The MoBa cohort is now based on regulations related to the Norwegian Health Registry Act. Participants were recruited from all over Norway between 1999 and 2008. The women consented to participation in 41% of the pregnancies. The cohort now includes 114,500 children, 95,200 mothers and 75,200 fathers. Blood samples were obtained from the mothers and fathers at 17–18 weeks of gestation and from mothers and children (umbilical cord) at birth. For the current study we used genotype data from 17,000 randomly selected trios, genotyped in three batches on three different arrays (harvest12: Illumina HumanCoreExome12v1.1, harvest24: Illumina HumanCoreExome24v1.0, rotterdam1: Illumina Global Screening Array MD v.1.0.). harvest12 and harvest24 were genotyped at the Genomics core facility in Trondheim, Norway while the rotterdam1 samples were genotyped at ERASMUS MC, Rotterdam, Netherlands. PLINK version

1.90 beta 3.365 (Purcell et al., 2007) was used to conduct quality control, details of QC have been previously described by Helgeland et al. (Helgeland et al., 2019). Individuals were excluded if they had a genotyping call rate below 95% or autosomal heterozygosity greater than four standard deviations from the sample mean. SNPs were excluded if they were ambiguous (A/T and C/G), had a genotyping call rate below 98%, MAF of less than 1%, or HWE P-value less than 1×10^{-06} . Population stratification was assessed using the HapMap phase 3 release 3 as a reference, by PCA using EIGENSTRAT (Price et al., 2006) version 6.1.4. Visual inspection identified a homogenous population of European ethnicity and individuals of non-European ethnicity were removed. Phasing was conducted using Shapeit (Delaneau et al., 2013) release 837 and the duoHMM approach was used to account for the pedigree structure. Imputation was conducted using the Haplotype reference consortium (HRC) release 1-1 as the genetic reference panel. The Sanger Imputation Server was used to perform the imputation with the Positional Burrows Wheeler Transform (PBWT). The phasing and imputation were conducted separately for each genotyping batch. A core homogeneous sample of European ethnicity across all batches and arrays were available for use in analysis (totals prior to analysis-specific exclusions for relatedness: $N_{Mothers} = 14,804$; $N_{Fathers} = 15,198$). OCD diagnosis was ascertained through linkage to the Norwegian Patient Registry (ICD-10 codes from specialist health care registered from 2008-2018). Case inclusion criteria were ICD-10 code F42 diagnosed at least once. Control inclusion criteria were no F-diagnosis. Control individuals were excluded if they were related to any of the cases ($\text{pihat} > 0.2$). Sex ratio in the cases was 40:60 (male:female) and in the controls 50:50. GWA analysis was based on 104 cases and 2193 controls. GWAS was performed using SAIGE (Zhou et al., 2018), including the first 10 PCs and genotyping batch as covariates.

NORDiC-SWE | Crowley, J; Mataix-Cols, D.; Rück, C. | Sweden | 31424634

A paper describing the rationale, design, and methods of the NORDiC study has been published previously (Mataix-Cols et al., 2020). NORDiC-SWE is the Swedish case-control arm of the study, and all samples were collected in Sweden between 2015 and 2019. This study was approved by the Regional Ethics Committee, Stockholm (EPN Stockholm) and the Institutional Review Board

(IRB) at the University of North Carolina at Chapel Hill and all subjects provided informed consent. OCD cases have a primary ICD-10 and/or DSM-5 diagnosis of OCD from a multidisciplinary specialist OCD team (established with a semi-structured instrument such as the MINI or the SCID). All patients were included in the study regardless of psychiatric comorbidity, as long as they fulfilled strict diagnostic criteria for OCD. Patients were excluded in cases of diagnostic uncertainty, such as OCD secondary to a neurological disorder or CNS insult, or where the differential diagnosis between OCD and an alternative condition was unclear. Our cases had a mean age at symptom debut of 12.1 years and 59% were female. Approximately 58% of patients had a documented psychiatric comorbidity. Controls were unrelated to any OCD case to the third degree and unaffected with OCD. Controls were excluded if they had a lifetime history of anorexia nervosa (controls were inherited from an anorexia GWAS). The controls included 95% females. The NORDiC-SWE cohort consists of 971 OCD cases and 2735 controls (LIFEGENE control batch 1: N = 1026 and LIFEGENE control batch 2: N = 1389). Subjects provided either blood or saliva for DNA extraction. All samples were genotyped on the Illumina Global Screening Array (GSA) at LIFE&BRAIN in Bonn, Germany. Samples were filtered for duplicates ($pi_{hat} \geq 0.95$) and cryptic relatedness ($pi_{hat} \geq 0.2$). We selected individuals of European ancestry using PCA implemented in PEDDY v0.4.3 (B. S. Pedersen & Quinlan, 2017). We defined variants as QC-failing if they met one of the following criteria: 1) maximum genotype missingness in a cohort > 0.02; 2) allele frequency < 0.001 in at least one cohort; 3) max – min allele frequency > 0.1 across all five cohorts; 4) max – min allele frequency > 0.03 across all three control cohorts; 5) genome-wide significant in a control vs. control synthetic GWAS. A total of 154,791 variants that met at least one of these criteria were excluded. We used the RICOPILI v2018_Dec_7.001 (Lam et al., 2020) pipeline to run an automated round of pre-imputation QC. The pre-imputation QC step

involved a series of hard filters on variant and sample level data, including removing variants with pre-sample pruning call rate < 0.95, samples with call rate < 0.98, FHET outside of ± 0.20 , samples with discrepancies between reported and derived sex, and post sample-pruning variants that meet any of the following : 1) call rate < 0.98, 2) missing difference > 0.02, 3) invariant positions, 4) MAF > 0.01, 5) HWE $p < 1 \times 10^{-06}$ in controls, and 6) HWE $p < 1 \times 10^{-10}$ in cases.

The final dataset consisted of 647,335 variant calls across a total of 1997 cases and 3943 controls. RICOPILI's impute_dirsub module was used to conduct imputation using the Haplotype Reference Consortium (HRC) reference panel. In our imputation run we used EAGLE v2.3.5 (Loh, Danecek, et al., 2016) for pre-phasing, and minimac3 v2.0.1 (Das et al., 2016) for imputation. We derived 3 different imputed callsets from this process: 1) a set of high confidence imputed genotypes (2,771,425 SNPs), 2) 7,112,906 imputed best-guess genotypes with medium level accuracy, and 3) genotypes for variants where imputation accuracy is lowered in order to increase the total number of variants included in the imputation (resulting in 8,995,398 SNPs). We elected to run our GWAS across the largest dataset of imputed variants that were generated during the imputation process on a subset of samples that were of European ancestry. We conducted PCA on these samples across high-confidence imputed genotypes using the pacer_sub RICOPILI module and tested the first 20 PCs for significant association with sample case/control status (significant = p-value < 0.05/20). We identified PCs 1, 3 and 14 as significant predictors and used them as covariates in the GWAS analysis. We used RICOPILI's postimp_navi module to conduct the final GWAS. Summary statistics were well controlled across the separate GWAS. We noted a lambda of 1.01 and a lambda1000 of 1.01 across a total of 7,679,714 tested SNPs.

NORDiC-NOR | Crowley, J.; Kvale, G.; Hansen, S. | Norway | 31424634

NORDiC-NOR is the Norwegian case-control arm of the NORDiC study and all samples were collected in Norway between 2016 and 2019. The NORDiC-NOR study was approved by the

Norwegian Regional Committee for Medical and Health Research Ethics (IRB00001872 REK West) under project number 2018/52 REKVest (PI: Bjarne Hansen) and project number: 2014/75 REKVest (PI: Jan Haavik) and all subjects provided informed consent. OCD Cases have a primary ICD-10 and/or DSM-5 diagnosis of OCD from a multidisciplinary specialist OCD team (established with a semi-structured instrument such as the MINI or the SCID). All patients were included in the study regardless of psychiatric comorbidity, as long as they fulfilled strict diagnostic criteria for OCD. Patients were excluded in cases of diagnostic uncertainty, such as OCD secondary to a neurological disorder or CNS insult, or where the differential diagnosis between OCD and an alternative condition was unclear. The mean age at symptom debut was 17 years and 65% were female. Approximately 51% of patients had a documented psychiatric comorbidity. Controls reported no OCD or first-degree family members with OCD. Among the controls, 50% were female. Subjects provided either blood or saliva for DNA extraction. All samples were genotyped on the Illumina Global Screening Array at LIFE&BRAIN in Bonn, Germany. The pre-GWAS QC applied to the NORDiC-NOR dataset (482 cases, 343 controls in the raw data) was nearly identical to that applied to NORDIC-SWE data after merging of separate genotype data, consisting of pruning of cryptic relatedness, marking of samples that are of likely European ancestry and pruning of the dataset for single variants where there was suggestive evidence of technical biases or batch effects. In our cryptic relatedness QC step, we identified 4 samples that had a mean pi_{hat} with other samples of ≥ 0.1 , 74 samples that had evidence of being a sample duplicate ($pi_{hat} \geq 0.95$) and 8 samples from the remaining cohort with evidence of cryptic relatedness ($pi_{hat} \geq 0.2$). We found a total of 6263 variants overlapped between the merged PLINK fileset and the 1000 genomes data included in PEDDY (B. S. Pedersen & Quinlan, 2017) and identified a total of 368 cases and 315 controls with likely European ancestry. We performed variant-level QC on a PC-pruned subset of 340 cases and 307 controls, defining variants as failing

if they met one of the following criteria: 1) maximum genotype missingness in a cohort > 0.02 ; 2) allele frequency of 0 in at least one cohort; 3) max – min allele frequency > 0.1 . A total of 136,525 variants that met at least one of these criteria were excluded, leaving us with a final case/control dataset consisting of genotype calls across 479,358 variants. We used calls across these variants in samples that had been pruned for relatedness issues (not including standard pairwise relatedness issues as RICOPILI can detect these) as input for the GWAS (407 cases, 340 controls). Imputation and GWAS was performed analogously to the NORDiC-SWE and EGOS analysis. The final data set consisted of 365 cases and 315 controls. Imputed callsets resulted in 1) 3,043,464 high confidence imputed genotypes 2) 7,277,174 genotypes with medium level, and 3) 8,964,589 genotypes with a low imputation accuracy. We identified PCs 1, 2, 3 and 4 as significant predictors and used them as covariates in the NORDiC-NOR GWAS. The analysis resulted in a lambda of 1.00 and a lambda1000 of 1.01 across a total of 7,518,582 tested SNPs.

OCGAS_all | Multiple | USA | 22889921

Results of The OCD Collaborative Genetics Association Study (OCGAS) have been published previously (Mattheisen et al., 2015). For detailed sample- and analysis description refer to the primary article. In brief: The OCGAS sample consists of 986 cases from OCGAS, a family based sample, and 1023 controls from the Genomic Psychiatry Cohort (GPC; (Pato et al., 2013)), a population based sample. The inclusion of control samples was slightly altered in this publication compared to the original study. For study inclusion, probands were required to meet DSM-IV criteria for OCD with onset of obsessions and/or compulsions before the age of 18 years (mean

= 9.4 years; SD = 6.35), as evaluated by a PhD-level clinical psychologist using the Structured Clinical Interview for DSM-IV modified and extended to include additional symptom and diagnostic information. Several mental- and brain disorders were reasons for exclusion. Genotyping was performed at the Johns Hopkins SNP Center using Illumina's HumanOmniExpress bead chips (Illumina, San Diego, CA, USA). A stringent quality control protocol was followed, including checking the relatedness of samples, sex comparison, Mendelian inconsistencies etc. One sample was removed per related pair with $p_{\text{ihat}} > 0.2$. Multidimensional scaling analyses were performed on singleton OCD cases and unselected controls, as implemented in PLINK (Purcell et al., 2007). Samples were removed when they significantly deviated in the first two multidimensional scaling dimensions (> 4 SD from the mean). The genotyped data were phased by SHAPEIT2 (Delaneau et al., 2013) and imputed by Minimac3 (Das et al., 2016) using the HRC release 1.1 (McCarthy et al., 2016) as the reference panel. GWAS was performed on SNPs with info score > 0.8 and MAF > 0.01 , using logistic regression model in PLINK2 (Chang et al., 2015) with the first four and the 9th MDS components as covariates. Ethics approvals for the OCGAS study were obtained from the Hopkins Medicine Institutional Review Boards, the Butler Institutional Review Board, the UCLA Institutional Review Boards, the Mass General Brigham Human Research Committee, the Columbia University Institutional Review Boards, and the National Institutes of Health Institutional Review Board (NIH IRB).

OCGAS-nestadt | Nestadt G., OCGAS Consortium | USA

The OCGAS-nestadt study was conducted at one of the five participating recruitment sites of the National Institute of Mental Health. The study was approved by the IRB boards at: Johns Hopkins University School of Medicine, Brown Medical School, New York State Psychiatric Institute and College of Physicians and Surgeons at Columbia University, University of California Los Angeles (UCLA) School of Medicine, Massachusetts General Hospital and Harvard Medical School, National Institute of Mental Health, and Keck School of Medicine at the University of Southern California. Samples were collected between 2007 and 2014. The sample comprised of trios (including an affected proband and both parents) or in some cases a proband and an unaffected

sibling. Each case was evaluated by a MD- or PhD-level clinical psychologist using the Structured Clinical Interview for DSM-IV (SCID). The checklist of obsessions and compulsions from the Y-BOCS, refined to include the age of onset, offset, and severity of each symptom, as well as the Y-BOCS scores for the worst episode (lifetime) was recorded. Course and treatment response variables were also included. A similar model was used for evaluating tics and Tourette disorder. Axis I disorder diagnoses were assigned using the JHU Diagnostic Assignment Checklist, an instrument that documents the criteria for over 20 DSM-IV disorders; this instrument also was the primary tool for the diagnostic consensus procedure. The SCID-II was used to evaluate four personality disorders (schizotypic, obsessive-compulsive, avoidant, and dependent), and the FISC was used to obtain additional information about each participant from a knowledgeable informant. Children over the age of eight were assessed in the same way, except that the Kiddie-SADS was used in place of the SCID. Final diagnostic status was assigned based on the consensus of two psychiatrists or psychologists reviewing the case independently. The agreement between diagnosticians using the Diagnostic Assignment Checklist has been studied and found to be excellent for variables such as age at onset of OCD. The chance-corrected percent agreement between the diagnosticians for the diagnosis of OCD was $K = 0.92$; for age at onset of OCD, $K = 0.88$ (for age ± 5 years), and Pearson's $r = 0.71$. The diagnostic information from each site was reviewed by one of the five members of the JHU diagnostic consensus committee to ensure comparability across sites. For study inclusion, probands were required to meet DSM-IV criteria for OCD - with onset of obsessions and/or compulsions before the age of 18 years (mean = 9.4 years; SD = 6.35). Subjects with an age-of-onset > 17 years, schizophrenia, severe mental retardation that does not permit an evaluation to characterize the psychiatric disorder, Tourette disorder or OCD occurring exclusively in the context of depression (secondary OCD) were excluded. In addition, individuals were removed from the sample if they were previously diagnosed with brain pathology including brain tumors, Huntington's disease, Parkinson's disease, or Alzheimer's disease. This resulted in a final sample of 212 cases and 212 controls. Samples were genotyped on the Illumina PsychChip array at USC. Ethics approvals for the OCGAS study were obtained from the Hopkins Medicine Institutional Review Boards. The

OCD Collaborative Genetics Association Study (OCGAS) is a collaborative research study and was funded by the following NIMH Grant Numbers: MH071507, MH079489, MH079487, MH079488 and MH079494.

OCGAS-ab | Arnold P., Burton C. | Canada | 27777633, 31772171

Samples in the OCGAS-ab cohort were collected between 2008-2015 and consist of trios. The study was approved by REB at Hospital for Sick Children. Cases were required to meet DSM-IV criteria for OCD and were selected based on the K-SADS and CY-BOCS questionnaires. Exclusion criteria for cases were age of onset < 18 years of age, psychosis, and history of severe neurological disorders other than Tourette's disorder. Of the 55 cases and 55 controls, 50% are female, the mean age was 16.38 (SD = 3.78). One individual presented a co-morbid diagnosis of ASD, 11 of ADHD, six of Tics/Tourette's syndrome, five of anxiety disorders, one of a eating disorder, two of depression, and two presented learning difficulties. Samples were genotyped on the Illumina PsychChip array at USC. Ethics approvals for the OCGAS study were obtained from the Hopkins Medicine Institutional Review Boards.

OCGAS-gh | Grünblatt E., Walitza S. | Switzerland and Germany | 28065182, 29102815

Participants in the early-onset OCD cohort were recruited at the Departments of Child and Adolescent Psychiatry of the Universities of Würzburg, Marburg, Aachen, and Freiburg in Germany and Zurich in Switzerland. Informed written consent was obtained in all cases by the participants or their parents. The study was approved by the ethical commissions of all involved universities in accordance with the latest version of the Declaration of Helsinki, including an ethical permission granted by the Ethic Committees from Aachen, Würzburg, Marburg, Freiburg, and the Cantonal Ethic Commission of Zürich (Ref. Nr. 39/97, 140/3 and EK: KEK-ZHNr. 2010-0340/3). Samples were collected between 2000 and 2016 and resulted in 56 cases and 56 controls in trios

or case-control samples. Patients were included if they fulfilled the diagnostic criteria for current OCD according to DSM-4 and ICD-10. To assess OCD diagnostic criteria, early-onset OCD patients and parents were interviewed separately by senior clinicians with a semi-structured diagnostic interview of psychiatric disorders in children and adolescents (Kinder-DIPS; children and parents version) (Margraf et al., 2017); the patients and parents located in Zürich underwent the German version of a semi-structured clinical interview (K-SADS-PL) (Kaufman et al., 2000). In addition, severity and additional characteristics of OCD symptoms were evaluated with the Children's Yale-Brown Obsessive Compulsive Scale (CY-BOCS) (Scahill et al., 1997). Kinder-DIPS or K-SADS-PL was used to screen for the existence of comorbid disorders (affective-, anxiety-, eating- and tic disorders, attention-deficit/hyperactivity disorder, conduct-, oppositional disorder, as well as substance use, abuse, psychosis and somatic diseases) in children and adolescents. CASCAP-D was used to screen for autistic spectrum disorders (Schmidt et al., 2000). Present and lifetime Tourette's syndrome and tic disorders were assessed with the adapted German version of the Child and Adult Schedule for Tourette and Other Behavioral Syndromes (STOBS) and the Yale Global Tic Severity Scale (YGTSS) (Leckman et al., 1989) in the Zürich patients. Case exclusion criteria were a lifetime history of Tourette's syndrome, psychotic disorder, autism spectrum disorder, mental retardation ($IQ < 70$) or alcohol dependence. Patients with comorbid disorders were only included if OCD was the primary diagnosis. Controls were excluded if they presented a major psychiatric disorder or an $IQ < 70$. All pediatric OCD patients received cognitive behavior therapy, while when insufficient, drug treatment was added, most commonly with an SSRI. Samples were genotyped on the ILMN PsychChip array and genotyped at USC in one batch.

OCD-WWF | Domschke, K; Berberich G. | Germany

For the obsessive-compulsive disorder - Windach Würzburg Freiburg (OCD-WWF) study, 129 inpatients with OCD (mean age \pm SD: 34.47 ± 11.86 years; 66 female) were recruited at the

Psychosomatic Hospital Windach, Windach, Germany, between 2014 and 2017. OCD diagnosis was ascertained on the basis of a structured clinical interview according to DSMIV criteria (SCID-I) by experienced psychiatrists and/or clinical psychologists. Inclusion criteria were age at inclusion between 18 and 80 years and European descent (self-report up to third generation). Exclusion criteria comprised severe somatic and neurological disorders, the consumption of illegal drugs, and pregnancy. Comorbid tic disorder, trichotillomania, skin-picking disorder or other current axis I diagnoses except for depression ($n = 71$), specific phobias ($n = 10$), generalized anxiety disorder ($n = 1$), social phobia ($n = 5$), panic disorder ($n = 2$), agoraphobia ($n = 5$) or posttraumatic stress disorder ($n = 3$) were excluded. The study was approved by the ethics committee of the University of Würzburg, Germany and was conducted according to the ethical principles of the Helsinki Declaration. All patients gave written informed consent prior to participation. Cases and controls were genotyped on Illumina's Global screening array. GWAS analysis was performed using RICOPII (Lam et al., 2020), employing standard parameters. First, we ran an automated round of pre-imputation QC. The pre-imputation QC step involved a series of hard filters on variant and sample level data, including removing variants with pre-sample pruning call rate < 0.95 , samples with call rate < 0.98 , FHET outside of ± 0.20 , samples with discrepancies between reported and derived sex, and post sample-pruning variants that meet any of the following : 1) call rate < 0.98 , 2) missing difference > 0.02 , 3) invariant positions, 4) MAF $>$

0.01, 5) HWE $p < 1E-6$ in controls, and 6) HWE $p < 1E-10$ in cases. RICOPILI's impute_dirsub module was used to conduct imputation using the Haplotype Reference Consortium (HRC) reference panel. RICOPILI's impute_dirsub module was used to conduct imputation using the 1000's genomes (1000G) reference panel (The 1000 Genomes Project Consortium, 2015). We conducted PCA on these samples across high-confidence imputed genotypes using the pacer_sub RICOPILI module and tested the first 20 PCs for significant association with sample case/control status. In a final step, we used RICOPILI's postimp_navi module to conduct the GWAS analysis.

Psych_Broad | Mathews C | Netherlands, Italy, USA, Spain

Psych_Broad sample consists of 1396 European ancestry cases and 4009 population matched controls. The cases were recruited predominantly from OCD specialty clinics in the US, Spain, The Netherlands, and Italy. The Spanish controls were part of the Mental-Cat clinical sample or the INSchool population-based cohort. A total of 1757 controls from the Mental-Cat cohort (60.3% males) were evaluated and recruited prospectively from a restricted geographic area at the Hospital Universitari Vall d'Hebron of Barcelona (Spain) and consisted of unrelated healthy blood donors. The INSchool sample consists of 765 children (76.2% males) from schools in Catalonia. Both studies have been approved by the Clinical Research Ethics Committee (CREC) of Hospital Universitari Vall d'Hebron. All methods were performed in accordance with the relevant guidelines and regulations and written informed consent was obtained from participant parents before inclusion into the study. Detailed information has been published previously (Bosch et al., 2021; Rovira et al., 2020). Genomic DNA samples were obtained either from peripheral blood lymphocytes by the salting out procedure or from saliva using the Oragene DNA Self-Collection

Kit (DNA Genotek, Kanata, Ontario Canada). DNA concentrations were determined using the Pico- Green dsDNA Quantitation Kit (Molecular Probes, Eugene, OR). All cases were required to have a DSM-IV diagnosis of OCD. The population based unscreened controls were recruited from the same countries. All samples were genotyped on Infinium PsychArray-24 at Broad Institute (Cambridge, MA, USA). Standard quality control (QC) protocol was conducted with PLINK2 (Chang et al., 2015). Samples were removed for call rates $< 98\%$, sex discrepancy and ambiguous genomic sex, related samples with $\text{pihat} > 0.2$. SNP QC included removing monomorphic SNPs, SNPs with genotyping rate $< 98\%$, SNPs with $\text{MAF} < 0.01$, SNPs with $P < 1 \times 10^{-6}$ in HWE test among controls or $P < 1 \times 10^{-10}$ among cases, SNPs with differential missing rate between cases and controls (> 0.02), and SNPs with batch effect ($P < 1 \times 10^{-5}$). Multidimensional scaling (MDS) analyses were performed in PLINK2 and samples were removed when they were significant outliers in the first six MDS dimensions or when there were no matching cases or controls on the MDS dimensions. The genotyped data were phased by SHAPEIT2 (Delaneau et al., 2013) and imputed by Minimac3 (Das et al., 2016) using the HRC release 1.1 (McCarthy et al., 2016) as the reference panel. GWAS was performed on SNPs with INFO score > 0.8 and $\text{MAF} > 0.01$, using a logistic regression model in PLINK2 with the first six and the 8th MDS PCA components as covariates. The GWAS result of this data set has not been previously published.

UKBB | Breen, G. | United Kingdom | 30305743

The UK Biobank sample consists of 776 OCD cases and 125,729 controls, see (Bycroft et al., 2018) for a general description of the UK Biobank resource. The data was derived from an online mental health questionnaire, completed by participants between July 2016 and July 2017. Research on the UK Biobank is conducted under a generic Research Tissue Bank approval from the UK North West Multi-centre Research Ethics Committee (MREC). This research was

approved to be conducted under that approval by the governing Research Ethics Committee of the UK Biobank. The analyses in this paper were performed under an approved extension to project 16577. Cases were defined by self-report of a professional diagnosis of OCD ("Have you been diagnosed with [Obsessive compulsive disorder (OCD)] by a professional, even if you don't have it currently?"). The median age of cases at the time of report (not age at diagnosis) was 62 (interquartile range (IQR) = 55-68), 59% of cases are female. All participants who did not report an OCD diagnosis were included as controls. The median age of controls at the time of report was 65 (IQR = 58-70), 56% of controls are female. Genotyping of cases and controls was performed on the Affymetrix Axiom UK Biobank Array / Affymetrix Axiom UK BiLEVE array and genotyped in the Affymetrix Research Services Laboratory, Santa Clara, CA in several batches. For genotype quality control the following filters were applied to the data: MAF > 0.01 and call-rate > 98%. 4-means clustering was applied on the first two PCs to determine participants of European ancestries (as described by (Warren et al., 2017)). Further, quality assurance outliers marked by UKBB were removed. Relatives (KING relatedness > 0.044) greedily (e.g. keeping the parents in a parent-child trio) and participants with mismatched sex were removed (reported females with FX \geq 0.6, reported males with FX \leq 0.85; FX determined using chrX SNPs in approximate linkage equilibrium, $r^2 < 0.2$). Imputation was performed with IMPUTE4 (Bycroft et al., 2018) using a combined HRC/UK-10k reference panel. Imputed data was filtered on MAF \geq

0.01 and INFO \geq 0.4. GWAS was performed using SAIGE (Zhou et al., 2018), including the first six PCs from participants of European ancestries, factors for genotyping batch and assessment center as covariates.

Yale-Penn | Gelernter, J.; CO-PI: Kranzler, H. | USA | 24166409, 32492095

Yale-Penn data consists of three cohorts, called Yale-Penn 1 (also abbreviated GWCIDR), 2 (also abbreviated EXHCE), and 3 (also abbreviated MEGA123). Participants were recruited from eastern U.S sites and provided written informed consent as approved by the institutional review board at each site. Certificates of confidentiality were issued by the National Institute on Drug Abuse and the National Institute on Alcohol Abuse and Alcoholism. As previously described (Cheng et al., 2018), Yale-Penn 1 samples were genotyped using Illumina HumanOmni1-Quad v1.0 microarray, targeting approximately one million SNPs. Yale-Penn 2 samples were genotyped using Illumina HumanCore Exome array, including approximately 0.5 million SNPs. Yale-Penn 3 samples were genotyped using Illumina Multi-Ethnic Genotyping Array, including approximately 1.7 million SNPs. Genotyping was performed separately in each cohort using the 1000 Genomes Project reference panel, phase 1 (Durbin et al., 2010). First 10 PCs, calculated with EIGENSOFT (Price et al., 2006), were used to differentiate European and African ancestries through k-means clustering (Sherva et al., 2016; Solovieff et al., 2010). Imputation was performed with Minimac3 (Das et al., 2016), using the 1000 Genomes Project reference panel, phase 3 (The 1000 Genomes Project Consortium, 2015). For QC filtering, we kept SNPs with imputation quality scores > 0.6 , HWE filtering $> 5 \times 10^{-05}$, MAF > 0.05 , missing call rates for variants < 0.5 , and missing call rates for samples < 0.5 . Imputation score was > 0.8 . Data were aligned to the GRCh37 reference genome. Cases were defined according to the Diagnostic and Statistical Manual of mental disorders 4th edition (DSM-IV) by interviewers. As described for the MVP dataset, to optimize the number of cases to keep when removing individuals with kinship, which was calculated by KING (Manichaikul et al., 2010) for a minimum threshold of 0.0884 corresponding to a second-degree

relationship, we implemented the following algorithm where cases have priority compared to controls: between a case and a control, we remove the control; if there is relatedness between two cases, we remove the one that has the highest number of relationships with other individuals; we do the same between two controls. In conclusion, we obtained 74 cases and 1635 controls for Yale-Penn 1, 52 cases and 1646 controls for Yale-Penn 2, and 53 cases and 1926 controls for Yale-Penn 3, making up a total of 179 cases and 5207 controls. GWAS was performed separately on each of the three cohorts using PLINK 2.0 (Chang et al., 2015) setting age, sex, and the first ten PCs as covariates.

Supplementary Note 3: Definition of independent significant loci

In total, the main meta-analysis contained 1672 SNPs with a p-value smaller than 5×10^{-8} . To identify independent genome-wide significant loci, we first performed automated LD-based ‘clumping’ of genome-wide SNPs using PLINK within the RICOPILI pipeline. A locus was defined by considering genomic regions harboring one or more genome-wide significant SNPs and testing for LD with neighboring SNPs in a 500-kb range. If a locus contained several genome-wide significant SNPs in LD ($r^2 > 0.1$), the SNP with the lowest p-value was selected as lead SNP. We then used GCTA-COJO to determine if a locus contained multiple independent SNPs (see also **Online Methods**). For this approach, we used the stepwise conditional regression approach to correct the betas and p-values of neighboring SNPs (in a sliding window of 10 Mb) based on the LD between the SNPs. We used the default p-value threshold of 5×10^{-8} to define a genome-wide significant hit. The LD reference panel was created from 73,005 individuals from the QIMR Berghofer Medical Research Institute genetic epidemiology cohort. GCTA-COJO has the advantage that it can identify multiple independent signals within a single locus, which the standard PLINK clumping (as implemented into Ricopili) alone might miss.

Supplementary Note 4: X-chromosome analysis in 23andMe

Results for the association between X-chromosome SNPs and OCD are only reported for a subset of our study samples (23andMe), as this information was only available for a limited number of samples. We are currently working in parallel on a manuscript that focuses on sex-stratified analyses and will include the full meta-analysis of all samples that have information for the X-chromosome available in that manuscript. For the 23andMe sample analysis of the X-chromosome was equivalent to the analysis of autosomes (see cohort description above), with the difference that for phasing 23andMe built separate haplotype graphs for the non-pseudoautosomal region and each pseudoautosomal region, and then phased these regions separately. Association tests were computed with male genotypes coded as if they were homozygous diploid for the observed allele. No SNP on the X-chromosome reached genome-wide significance.

Supplementary Note 5: Subgroup, MTAG & common-factor analyses (GenomicSEM)

Heterogeneity was assessed with Cochran's Q statistic and I^2 statistic (Higgins et al., 2002; 2003). Cochran's Q is calculated as the weighted sum of squared differences between individual study effects and the pooled effect across studies, with the weights being those used in the pooling method. Q is distributed as a chi-square statistic with k (number of studies) minus 1 degrees of freedom. The I^2 statistic describes the percentage of variation across studies that is due to heterogeneity rather than chance. Unlike Q, it does not inherently depend upon the number of studies considered. As mentioned in the main article, we did not find evidence of significant heterogeneity in the results of our 30 genome-wide SNPs. Nevertheless, this does not rule out heterogeneity in study results at other loci or across the entire genome. Therefore, we conducted several analyses to explore whether the heterogeneity in sample ascertainment methods influenced genetic findings. In total, we found 39 independent loci that were associated with either

our meta-analysis, a common factor GenomicSEM analysis, subgroup analyses, MTAG analyses, or more than one of the analyses mentioned (see further below for details). An overview of these results and loci can be found in **Supplementary Table 3**.

Subgroup GWASs and meta analysis without 23andMe

To investigate loci that are predominantly associated with the individual subgroups (clinical, comorbid, biobanks, 23andMe), we conducted separate GWASs for each subgroup, identifying seven SNPs associated with OCD in 23andMe (rs2836950 uniquely in this analysis; see **Supplementary Figure 37A and 37B** for Manhattan-plots and QQ-plots), and two in the clinical data (rs11856716 uniquely in this analysis; see **Supplementary Figure 34A and 34B** for Manhattan-plots and QQ-plots, and **Supplementary Table 3** for a list of all significant SNPs, as well as **Supplementary Figure 35A, 35B and 36A and 36B** for Manhattan and QQ-plots of the comorbid and biobank specific GWASs). When conducting a combined GWAS of clinical, comorbid, and biobank cohorts (excluding 23andMe), two SNPs reached significance (**Supplementary Figure 33**). Genetic correlations (r_G) between the four sub-groups ranged from 0.63 ($SE = 0.11, P = 2.1 \times 10^{-09}$) between the comorbid and biobank sub-groups to 0.93 ($SE = 0.066, P = 9.89 \times 10^{-45}$) between the comorbid and 23andMe sub-groups. The genetic correlation between all clinically diagnosed cases (clinical, comorbid and biobank) and the 23andme cases was 0.92 ($SE = 0.040, P = 2.58 \times 10^{-113}$; See **Supplementary Table 7** for all correlation estimates).

Multi-Trait Analysis of GWAS (MTAG)

As the sample sizes for the different subgroups in our study varied, some of the subgroup analyses had limited power to detect subgroup-specific genetic variations. To address this, we conducted four multi-trait analyses to increase our power to detect these variations while still being able to identify subgroup-specific signals. MTAG is more commonly used to combine different but related traits into one meta-analysis by leveraging the shared heritability among the

different traits and thereby gaining power. While we do not have different traits, our aim was to generate ascertainment-specific estimates, while boosting power by leveraging the high shared heritability among the subgroups. In addition to stronger evidence (meaning lower p-value) for several of the OCD hits defined in the main GWAS, 11 SNPs were associated with the biobank samples, of which one SNP (rs138354) was not found in any of the other analyses (including the meta-analysis, one-factor GWAS (see further below for details), subgroup-specific GWASs, or other MTAG analyses), 19 SNPs were associated with the clinical subgroup, of which one SNP (rs4838077) was not found in any of the other analyses, 24 SNPs were associated with the self-report subgroup, with one SNP (rs56346215) uniquely found in this analysis, and 20 SNPs associated with the comorbid subgroup (see **Supplementary Table 3** for a full list of results, and **Supplementary Figures S34-37** for Manhattan-plots).

Genomic structural equation modeling (GenomicSEM)

First, we ran a common-factor model without individual SNP effects, following the tutorial ‘Models without individual SNP effects’ on the GenomicSEM github website (Grotzinger et al., 2019, see web resources). In brief, the summary statistics were first harmonized and filtered (with the munge-function), using HapMap3 as the reference file, using the effective sample size (clinical: $N_{\text{eff}}=21,562$, comorbid: $N_{\text{eff}}=18,794$, biobanks: $N_{\text{eff}}=36,124$, 23andMe: $N_{\text{eff}}=116,876$) as the input sample size, and filtering SNPs to $\text{INFO} > 0.9$ and $\text{MAF} > 0.01$. In a second step, multivariable LDSC was run to obtain the genetic covariance matrix and corresponding sampling covariance matrix, using precomputed European-ancestry LD scores, a sample prevalence of 0.5 and a population prevalence of 0.02. In a third step we ran a confirmatory factor analysis (CFA) using the pre-packaged common factor model in GenomicSEM using diagonally weighted least squares (DWLS) estimation. Second, we ran a multivariate GWAS of the common factor. In multivariate GWAS, the common factor defined by genetic indicators is regressed on each SNP, thereby generating summary statistics for the common factor (details can be found in the tutorial “GenomicSEM for Common Factor GWAS”, see web resources). First, summary statistics for all four subgroups were prepared for multivariate GWAS with the ‘sumstats’ function in

GenomicSEM, which aligns and merges all files. Next, with the 'userGWAS' function, the S (genetic covariance) and V (corresponding sampling covariance) matrices from the LDSC output (from the model without SNP effects) and the summary statistics were combined to create a separate S and V matrix for each SNP containing the effect estimate. The function also transforms the effect estimates from the summary statistics and their SEs into covariances and SEs of covariances by taking the product of the regression coefficient and SNP variance from the reference file (1000 genomes phase 3). The CFA showed some evidence of sample ascertainment impacting our results at a genome-wide scale, however, not beyond what has been observed with closely related psychiatric illnesses (Wray et al., 2018; Strom et al., 2024) and yielded a satisfactory fit for a one-factor model ($\chi^2 = 3.46, df = 2, P_{\chi^2} = 0.18, CFI = 0.998, SRMR = 0.056$), with high and similar loadings across all four subgroups (**Supplementary Figure 39** and **Supplementary Table 8**). Conducting a GWAS of this common factor model, we identified 20 genome-wide significant loci, all of which also reached significance in the main meta-analysis GWAS. Notably, the Manhattan-plot for the GenomicSEM GWAS closely resembles that of the main meta-analysis, though the common-factor GWAS shows reduced power to detect associated loci. This overall reduced power may partly stem from GenomicSEM's exclusion of subgroup-specific unique variances. Additionally, GenomicSEM's modeling approach may overlook localized SNP-heritability deviations, which are preserved in a straightforward meta-analysis, allowing for greater power. While these unique variances of the subgroups may introduce heterogeneity, this heterogeneity may still be OCD-related, potentially capturing variations in OCD symptom profiles across samples (e.g., one sample may overrepresent checking symptoms, while another has a higher prevalence of intrusive thoughts or religious OCD cases). If such sample-specific variations align with distinct genetic factors, the unique variances in GenomicSEM could reflect genuine OCD-related variance, which a simple meta-analysis might retain, leading to stronger associations.

Genetic correlations of subgroups with published GWASs

We also conducted bivariate LDSC analyses for the 112 published GWASs (see Supplementary Note 5 for details) and our four subgroup GWASs to study potential heterogeneity between our subgroups. As expected, the comorbid subgroup showed stronger genetic correlations for the traits for which these studies were primarily ascertained (for AGDS and iPSYCH MDD, for iPSYCH alone also ADHD and ASD). All three traits (MDD, ASD, ADHD) in contrast were less strongly associated with the clinical and biobank cohorts, with the 23andMe sub-group showing estimates in between (**Supplementary Figure 40** and **Supplementary Table 19**).

Supplementary Note 6: Cross-trait genetic correlations

As mentioned in the main paper, we used bivariate LDSC to investigate the extent of genetic correlations between OCD and a curated list of 112 previously published GWASs of psychiatric, substance use, cognition/socioeconomic status (SES), personality, psychological, neurological, autoimmune, cardiovascular, anthropomorphic/diet, fertility, and other traits. We conducted the same analysis for the four OCD sub-groups (**Supplementary Figure 41** and **Supplementary Table 19**) as well as comparing the OCD GWAS of only 23andMe to the OCD GWAS excluding 23andMe (**Supplementary Figure 42** and **Supplementary Table 19**). The co-morbid sub-group, as expected because the iPSYCH samples are enriched for MDD, ADHD, and ASD cases, tends to correlate higher with most psychiatric disorders they were primarily ascertained for, while clinical and biobank cohorts tend to correlate less strongly, with the 23andMe sub-group in between.

Beyond the strong genetic correlations of the main GWAS with psychiatric disorders described in the main text, high genetic correlations were also shown for several smoking phenotypes (positively with nicotine dependence, cigarettes per day, smoking initiation, and negatively with age smoking initiation), as well as a negative genetic correlation with alcohol dependence ($r_G = -0.1458$, 95% CI $[-0.22, -0.29]$, $P_{adj} = 0.0001$) was observed. We also found a negative

correlation with educational attainment (EA, $r_G = -0.1015$, 95% $CI [-0.14, -0.203]$, $P_{adj} = 7.32 \times 10^{-06}$) and other cognition traits, though there was some heterogeneity across the four subgroups with the biobank sub-cohort correlating positively with EA ($r_G = 0.1675$, 95% $CI [0.084, 0.251]$, $P_{adj} = 0.0004$), while the 23andMe and co-morbid sub-groups correlated negatively (23andMe: $r_G = -0.1608$, 95% $CI [-0.211, -0.111]$, $P_{adj} = 7.61 \times 10^{-10}$; Comorbid: $r_G = -0.1298$, 95% $CI [-0.205, -0.054]$, $P_{adj} = 0.0018$), and the clinical cohorts not showing a significant correlation ($r_G = -0.0412$, 95% $CI [-0.041, 0.022]$, $P_{adj} = 0.3191$). Moreover, some of the anthropomorphic traits (BMI, body fat, waist-hip-circumference, hip-circumference, and diet including fat) and auto-immune disorders (Crohn's disease, ulcerative colitis, and inflammatory bowel disease) showed a significant negative correlation with OCD, while asthma (adult onset) showed a positive correlation. In addition, significant correlations of substantial effect ($r_G > |0.25|$) were found with suicide attempt, tiredness, migraine (all positive), and with subjective well-being, childhood maltreatment, self-rated health, and risk taking-auto speeding (negative).

Supplementary Note 7: Gene-based results.

mBAT-combo

A gene-based analysis was conducted using multivariate Set-Based Association Test (mBAT-combo) (Li et al., 2023) within GCTA version 1.94.1 (Yang et al., 2011). mBAT-combo has the advantage of being better powered than other gene-based association test methods to detect multi-SNP associations in the context of masking effects (i.e., when the product of the true SNP effect sizes and the LD correlation is negative). To ensure that the overall power is maximized independent of masking effects at specific loci, mBAT-combo combines mBAT and fastBAT test statistics through a Cauchy combination method, which allows the combination of different test statistics without a priori knowledge of the correlation structure. mBAT-combo identified 207 protein-coding genes significantly associated with OCD at a Bonferroni-corrected threshold of $P =$

2.67×10^{-06} (**Supplementary Table 10**). Of these 104 were located on chromosome 6 in the extended major histocompatibility (MHC) locus (25MB - 35MB).

Transcriptome-wide association study (TWAS)

The transcriptome-wide association study (TWAS) using human prefrontal cortex gene expression weights generated by the psychENCODE consortium identified 24 significant protein-coding genes (Bonferroni-corrected $P = 4.76 \times 10^{-06}$), of which 15 showed strong evidence for a co-localised signal (COLOC PP4 > 0.8; i.e., a single SNP associated with both OCD and gene expression implicating a causal role for this association) (**Supplementary Table 11**). Conditional analyses of each locus identified 14 genes with statistically independent TWAS signals, 11 of which also had evidence of a co-localized signal (COLOC PP4 > 0.8) (**Supplementary Table 11**).

Summary-based Mendelian Randomization (SMR)

We complemented the TWAS analysis with summary-based Mendelian Randomisation (SMR) using eQTL meta-analysis results from the eQTLGen (Vösa et al., 2021) (whole blood) and MetaBrain consortia (de Klein et al., 2023). A total of 46 unique protein-coding genes (21 of which located in the MHC region) reached significance (**Supplementary Table 12**), of which seven were significant in both the MetaBrain (14 genes in total) and eQTLGen (39 genes in total) datasets. Of the 46 genes, 24 (six in MHC region) had a HEIDI (heterogeneity in dependent instruments) $P > 0.05$ in blood or brain tissue, indicating that a single causal variant underlies the GWAS and eQTL association (**Supplementary Table 12**). We found three significant genes with HEIDI $P > 0.05$ in both the MetaBrain and eQTLGen datasets (*WDR6*, *ARIH2*, and *FLOT1*), one of which (*WDR6*) also had a significant and colocated signal in the TWAS analysis.

Psychiatric Omnilocus Prioritisation Score (PsyOPS)

We used the Psychiatric Omnilocus Prioritisation Score (PsyOPS) (Wainberg et al., 2022) in a supervised approach that uses both positional mapping and biological annotations to prioritize putative causal genes for psychiatric traits. This approach expanded the list of candidate genes within OCD loci. For example, *CSRNP3* is nearest the index SNP rs9287859 on chromosome 2

and had a significant mBAT-combo association ($P = 3.99 \times 10^{-08}$). However, PsyOPS prioritized the sodium channel protein subunit genes *SCN1A*, *SCN2A*, and *SCN3A*, which are located further away from the index SNP but are given greater weight by PsyOPS than other genes in the locus due to their loss-of-function intolerance, brain-specific expression, and involvement in neurodevelopmental disorders.

TWAS-COLOC and SMR-HEIDI

Colocalization of the significant TWAS results (TWAS-COLOC) identified 15 genes (PPH4 > 0.8), 11 of which were conditionally independent, while HEIDI prioritized 22 significant SMR associations (HEIDI $P > 0.05$). In total, 35 genes significantly associated with OCD passed colocalization and/or HEIDI criteria.

Convergence across methods

From the individual results for the four approaches, the following overlap in significant genes was observed: For the positional mapping approaches 123 genes showed significance in both the mBAT-combo and PsyOPS analyses. For the functional eQTL approaches, ten genes showed significance in both approaches (TWAS and SMR). For the group of 37 genes that was identified by at least one positional approach (i.e., either mBAT-combo or PsyOPS) and one functional eQTL approach (i.e., either TWAS or SMR), 27 genes were identified by mBAT-combo and either TWAS (four genes) or SMR (23 genes), and 27 genes were identified by PsyOPS and either TWAS (nine genes) or SMR (18 genes). See the main manuscript for the genes that were either conditional independent (TWAS-COLOC) or were significant in both tested SMR tissues (whole blood and brain). Out of the 123 genes identified in both positional mapping approaches (see above), 24 genes overlapped with significant results for the SMR analysis using eQTLs from whole blood, and eight genes overlapped with significant results for the SMR analysis using eQTLs from brain tissue. Out of these, five genes showed significance across the positional mapping approaches and SMR analyses for both tissues. These included the aforementioned *WDR6*, *ARIH2*, *NCKIPSD* genes from the previously identified chromosome 3 region, *BTN3A2* in

the extended MHC region, and the Poly(A) Binding Protein Cytoplasmic 1 Like (*PABPC1L*) gene on chromosome 20. Out of the 123 genes identified in both positional mapping approaches (see above), 12 genes were also significant in the TWAS analysis. Of these, two genes were conditionally independent (*CTNND1*, and *WDR6*).

Overlap of gene-based results with rare-coding variant results

We performed a bi-directional look-up to assess the overlap of gene-findings from common-variance GWAS and rare-variance studies. We assessed 1) whether gene-findings from our GWAS showed evidence for rare variant involvement and 2) vice versa, whether findings from rare variant testing showed evidence of common variant association in our GWAS, using data from Halvorsen et al. (2021). We did not observe any evidence for differential case/control burden in the 251 GWAS-derived genes (**Supplementary Table 14**) for de novo synonymous variants ($p = 0.42$), missense non-damaging variants ($p = 0.58$) or missense damaging variants ($p = 0.58$). We observed the lowest p-value for loss of function variants ($p = 0.08$), which was driven by two such variants in cases and none in controls. These de novo variants impact the genes *QRICH1* and *ZKSCAN3*. We note that *QRICH1* in particular is loss of function intolerant ($pLI > 0.995$) and was observed to carry a separate loss of function de novo variant in the largest de novo variant study published for Tourette Syndrome (Wang et al., 2018).

Next, we looked up 200 genes from Halvorsen et al. (2021) with a probability of loss of function intolerance > 0.995 (defined by Lek et al., 2016) and effect size estimate > 1 . We found varying degrees of evidence for four different genes: *QRICH1*, *CTNND1*, *PTPRD*, *CHD3* (all look-ups can be found in **Supplementary Table 6**). Nevertheless, we found no evidence for unequal proportions of the 251 GWAS highlighted genes in the exome summary statistics ($OR = 1.25$, $p = 0.76$), most likely due to the exome study being severely underpowered (around 1300 total cases).

Supplementary Note 8: Sample size and number of GWS loci

For OCD, like with other complex traits, there seems to be a relationship between the number of genome-wide significant loci that were identified in the GWAS and the sample size in the analysis. To highlight this relationship, we conducted additional meta-analyses combining different subsamples of previously published GWASs (e.g., combining the PGC OCD2 GWAS with the 23andMe sample from this GWAS). A graphical representation can be found in **Supplementary Figure 43**. Between 25,000 (PGC OCD2 GWAS) and 30,000 cases (23andMe sample), the relationship between genome-wide significant loci and sample size seems to become linear, and (at least within European ancestry samples) each additional ~ 1000 cases will add one additional genome-wide significant SNP on average (see **Supplementary Table 20**).

Supplementary Note 9: Assumptions and Limitations of SMR (and TWAS)

Transcriptome-wide association studies and SMR are commonly used for the identification and prioritization of target genes from GWAS of complex traits. A key advantage of these approaches is their use of summary-based GWAS and QTL datasets, meaning they can be applied across a range of experimental settings. In addition, TWAS and SMR only consider genetically regulated gene expression and are therefore not confounded by non-genetic factors that may influence expression (e.g., medication use). Nonetheless, the approaches are associated with several important assumptions and limitations (PMID: 30926968). For TWAS, the prediction accuracy of gene expression relies on the heritability of gene expression and therefore requires well-powered test SNP genotype and expression datasets for the development of expression weights. Meanwhile, SMR using a single causal variant is unable to distinguish between pleiotropy, where the same causal variant affects the trait and gene expression, and causality, where the effect of a causal variant on a trait is mediated by gene expression. Both approaches also assume the genotype is associated with gene expression and genes with a higher cis-genetic correlation to

trait are more likely to be causal. In addition, they ignore long-range regulatory effects (trans-eQTLs) that may explain a large proportion of heritability, lack eQTL data from different ancestry groups, and cannot identify causal tissues or cell types. Despite these limitations, TWAS and SMR can identify important relationships between QTL datasets and GWAS to find target genes (PMID:36778001). We combined information from these approaches and other gene-based methods with different underlying assumptions (e.g., mBAT-combo or PsyOPS) to increase confidence in our gene-based associations and prioritize target genes within OCD loci.

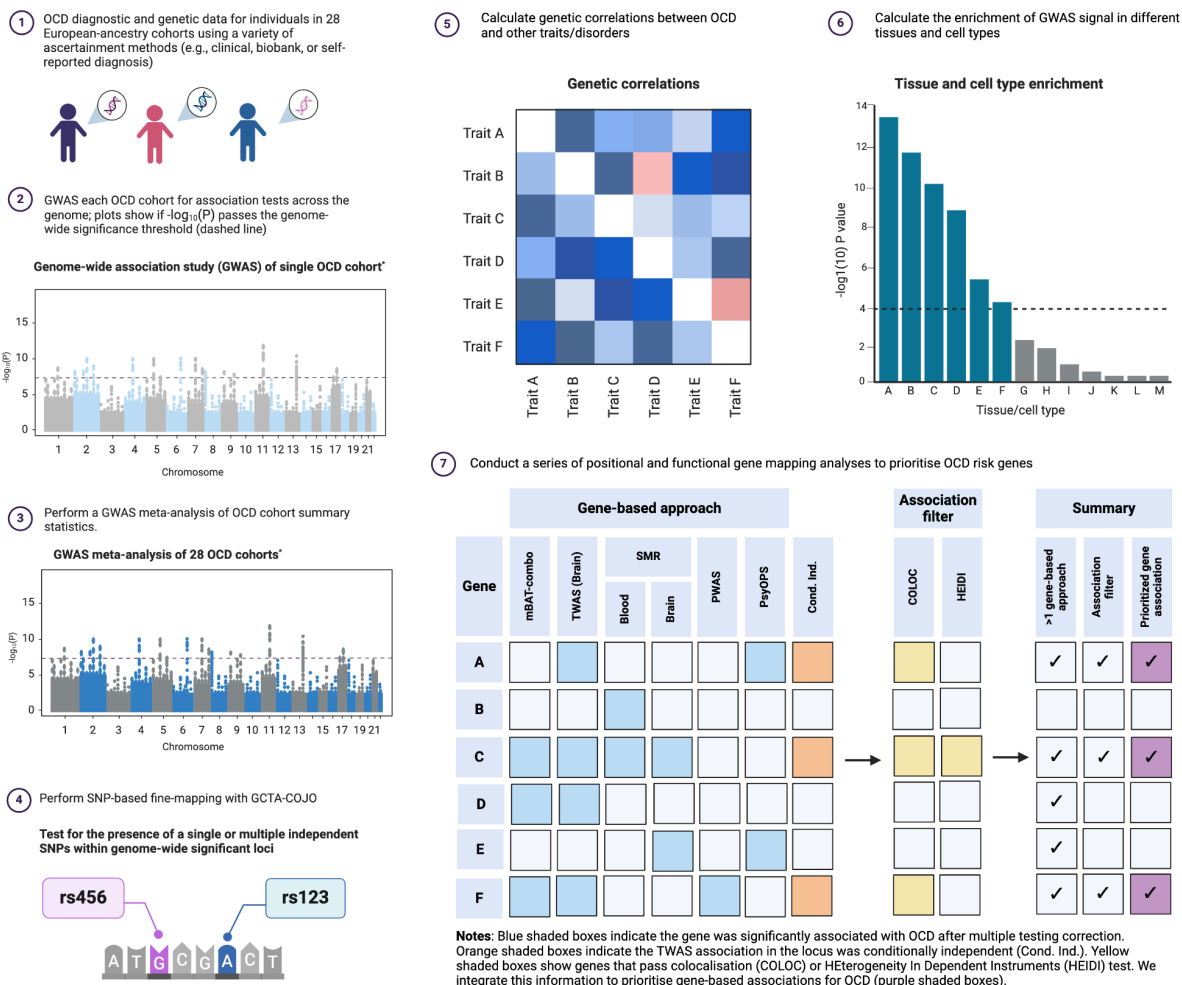
Supplementary Note 10: A note on liability scale heritability estimates and sample ascertainment

When calculating liability scale heritability, both sample prevalence and population prevalence are important considerations. The sample prevalence is typically estimated from the number of cases and controls, but in biobank-scale studies, specialized analytical approaches (e.g., REGENIE) may obscure the true sample prevalence underlying the analysis. For population prevalence, robust estimates from entire populations are often used (e.g., from meta-analysis or national disease registers); however, ascertainment differences can complicate the use of precise population prevalences, as the specific fraction of the overall patient spectrum in individual cohorts may be unknown. Our study is limited by the use of a uniform prevalence estimate of 1% across all cohorts, which is on the lower end compared to previous studies and likely does not reflect the true prevalence across all enrolled cohorts. We adopted this approach due to the lack of reliable population prevalence data from prior meta-analyses or national registries, especially in Nordic countries, where available data is biased toward more severe and chronic cases based on hospital diagnoses. Consequently, liability scale heritability estimates from this manuscript should be interpreted with caution. Furthermore, when comparing estimates of (liability scale) heritability across studies, it is essential to consider the analytical approach and software that was used; comparisons should only be made when the same approaches are employed. For example,

previous (higher) heritability estimates for OCD from GWASs studies (e.g., Mattheisen et al., 2015, or Stewart et al., 2013) used GCTA (Yang et al., 2011) for estimations while the current study uses LDSC. It is known that heritability estimates from these two (and other) approaches do not compare sufficiently (Evans et al., 2018).

Supplementary Figures

Genome-wide Association Study of Obsessive Compulsive Disorder (OCD)

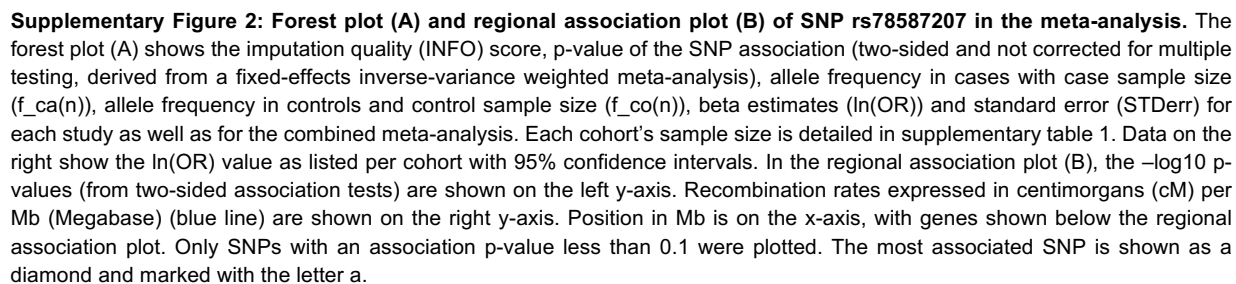


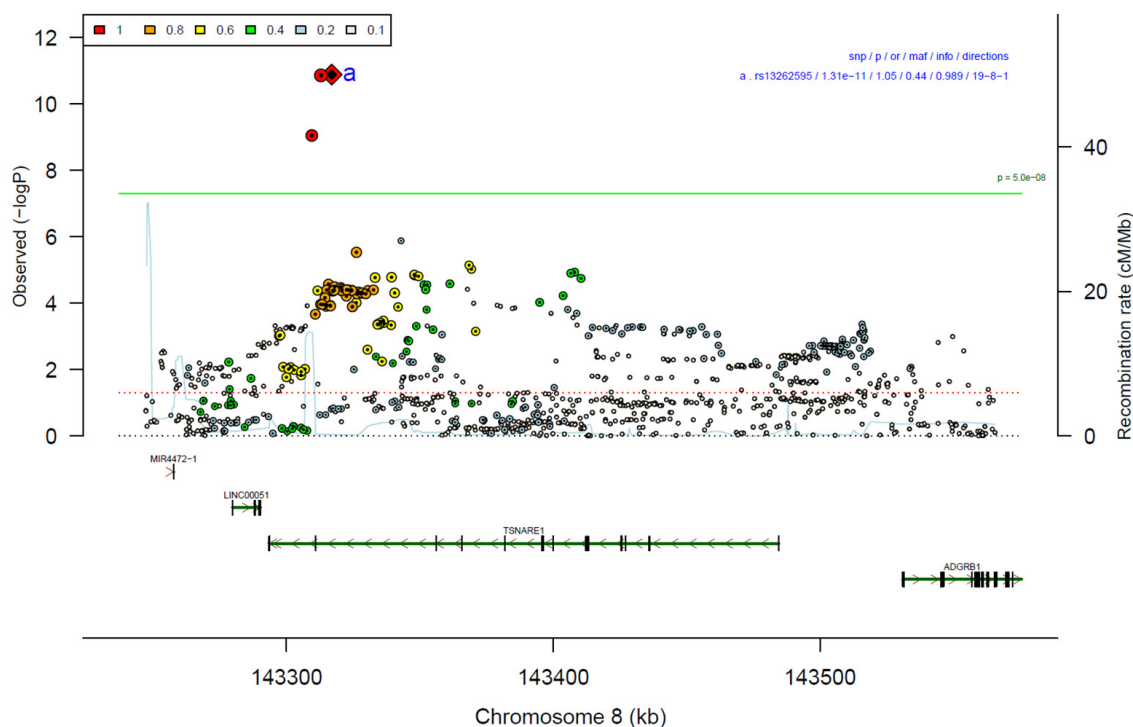
Supplementary Figure 1: Schematic overview of the main bioinformatic analyses presented in the manuscript.

Steps 1-4 outline the process for identifying SNP-based association signals: 1) Collecting clinical and genetic data from OCD patients and controls across 28 different cohorts at various institutes; 2) Conducting individual GWAS for each OCD cohort; 3) Meta-analyzing the individual GWAS results to obtain combined summary statistics; 4) Characterizing the SNP findings through fine-mapping to determine whether each locus contains an independent signal. *Steps 5-7 detail additional analyses performed to further characterize the SNP signals identified in steps 1-4, including:* 5) Genetic correlation analysis between OCD and other traits/disorders; 6) Analyzing tissue and cell-type enrichment of the OCD-associated variants; and 7) A series of positional and functional gene mapping analyses to prioritize OCD risk genes.

* GWAS Manhattan plots are not based on real data. Figure was created in BioRender

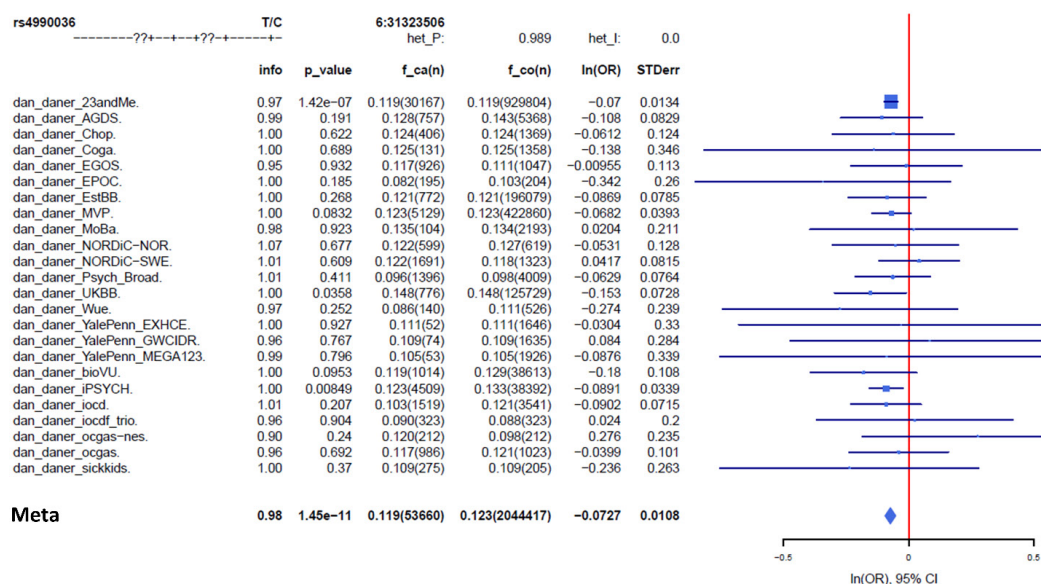
Regional association plots and forest plots of the 30 significant SNPs



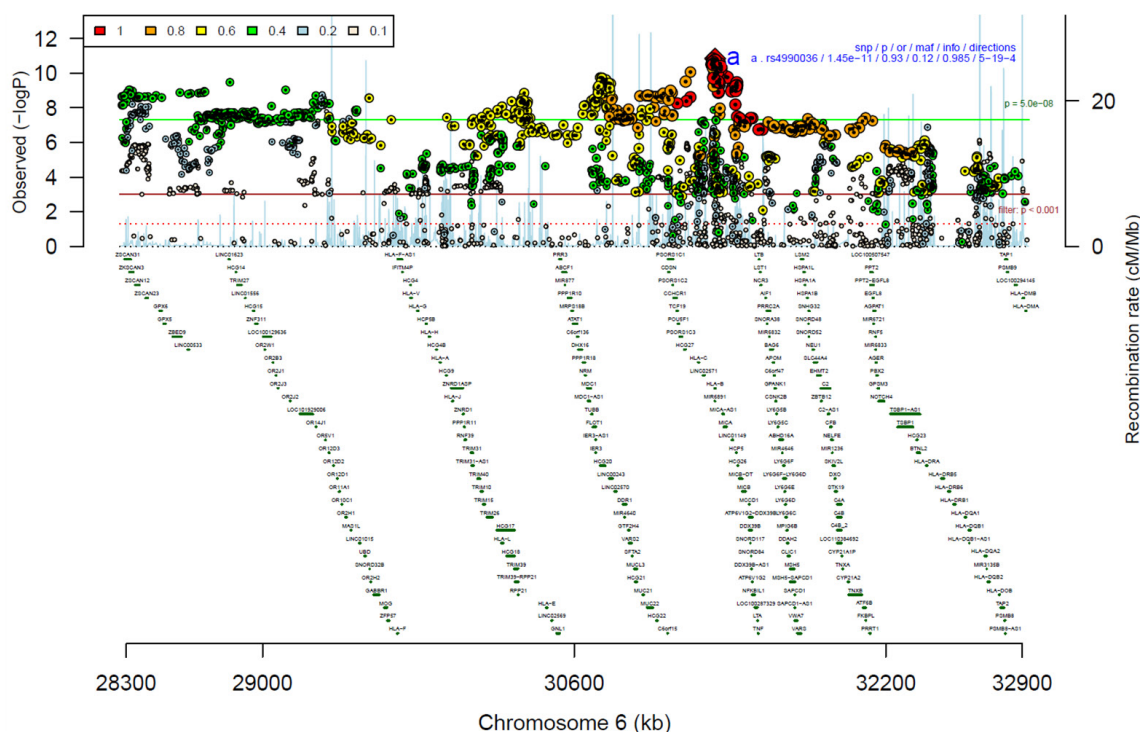


Supplementary Figure 3: Forest plot (A) and regional association plot (B) of SNP rs13262595 in the meta-analysis. The forest plot (A) shows the imputation quality (INFO) score, p-value of the SNP association (two-sided and not corrected for multiple testing, derived from a fixed-effects inverse-variance weighted meta-analysis), allele frequency in cases with case sample size ($f_{ca}(n)$), allele frequency in controls and control sample size ($f_{co}(n)$), beta estimates ($\ln(OR)$) and standard error (STDerr) for each study as well as for the combined meta-analysis. Each cohort's sample size is detailed in supplementary table 1. Data on the right show the $\ln(OR)$ value as listed per cohort with 95% confidence intervals. In the regional association plot (B), the $-\log_{10}$ p-values (from two-sided association tests) are shown on the left y-axis. Recombination rates expressed in centimorgans (cM) per Mb (Megabase) (blue line) are shown on the right y-axis. Position in Mb is on the x-axis, with genes shown below the regional association plot. Only SNPs with an association p-value less than 0.1 were plotted. The most associated SNP is shown as a diamond and marked with the letter a.

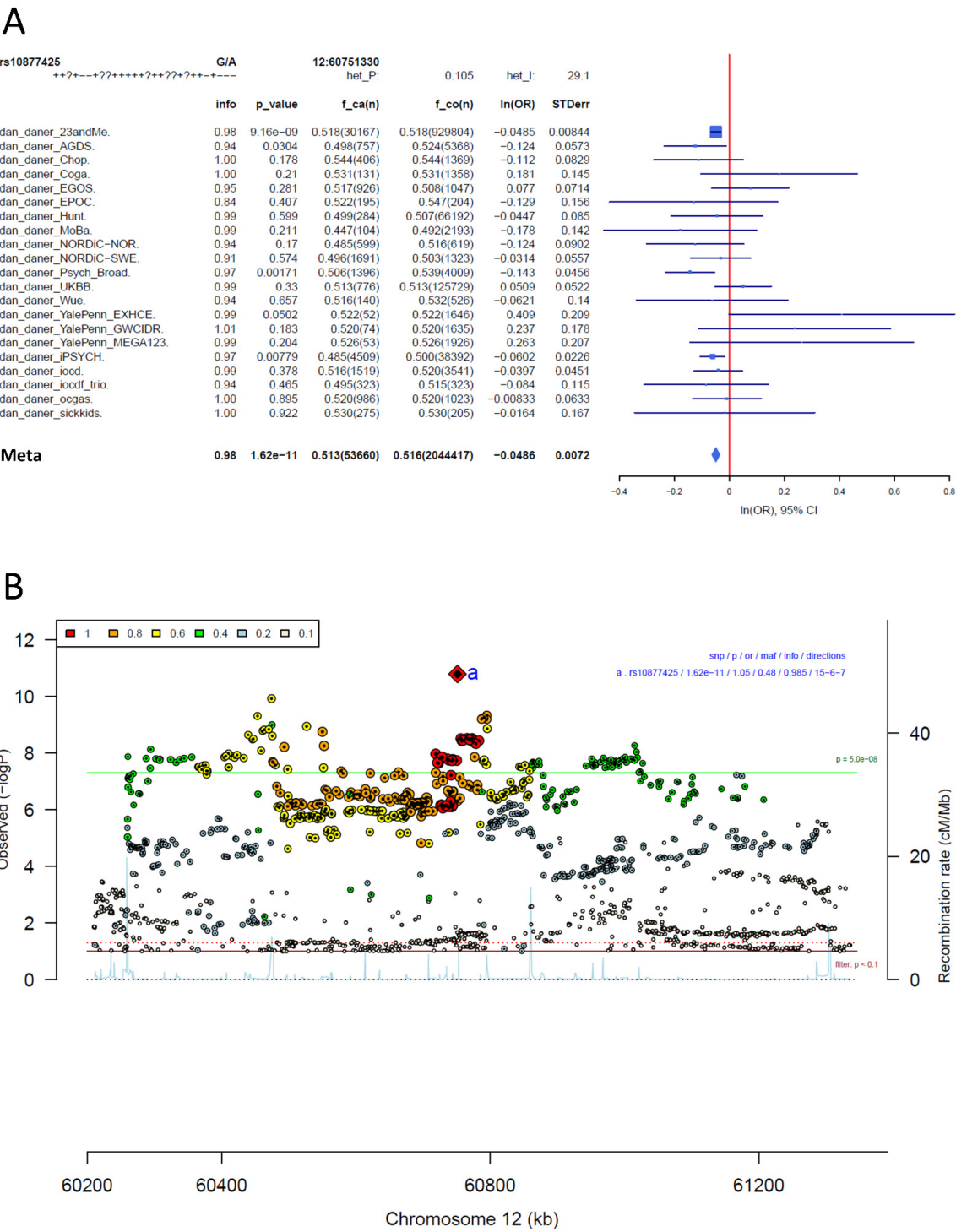
A



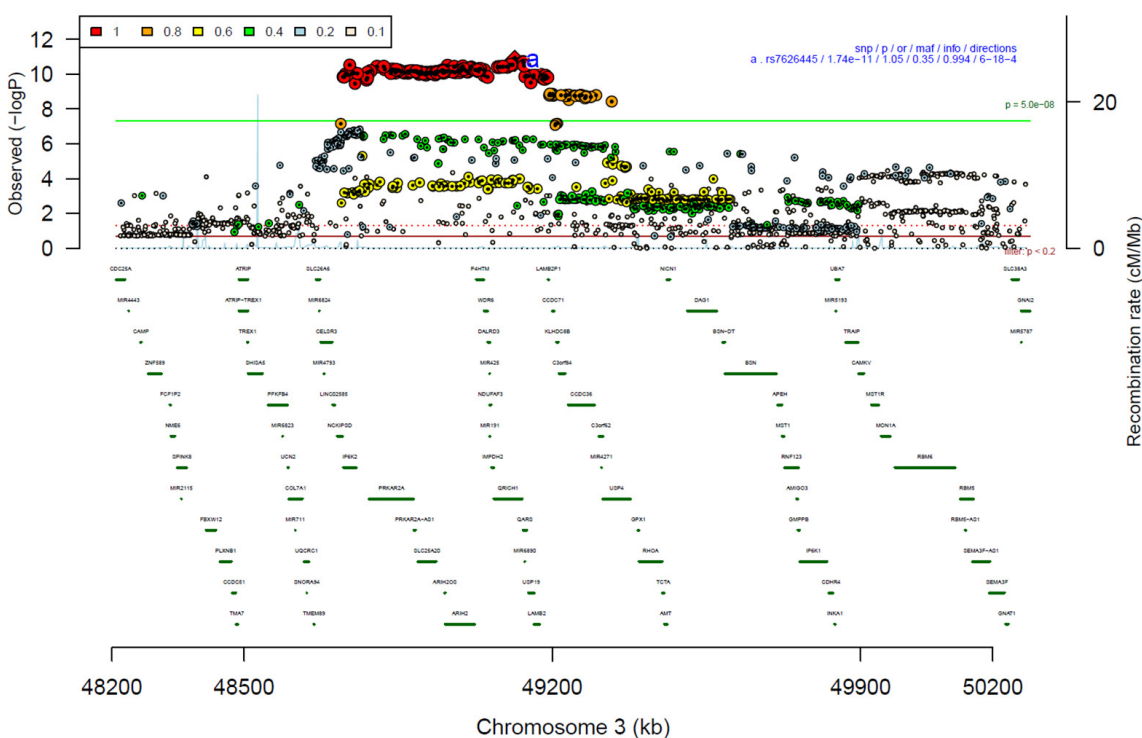
B



Supplementary Figure 4: Forest plot (A) and regional association plot (B) of SNP rs4990036 in the meta-analysis: The forest plot (A) shows the imputation quality (INFO) score, p-value of the SNP association (two-sided and not corrected for multiple testing, derived from a fixed-effects inverse-variance weighted meta-analysis), allele frequency in cases with case sample size ($f_{ca}(n)$), allele frequency in controls and control sample size ($f_{co}(n)$), beta estimates ($\ln(OR)$) and standard error (STDerr) for each study as well as for the combined meta-analysis. Each cohort's sample size is detailed in supplementary table 1. Data on the right show the $\ln(OR)$ value as listed per cohort with 95% confidence intervals. In the regional association plot (B), the $-\log_{10}$ p-values (from two-sided association tests) are shown on the left y-axis. Recombination rates expressed in centimorgans (cM) per Mb (Megabase) (blue line) are shown on the right y-axis. Position in Mb is on the x-axis, with genes shown below the regional association plot. Only SNPs with an association p-value less than 0.1 were plotted. The most associated SNP is shown as a diamond and marked with the letter a.

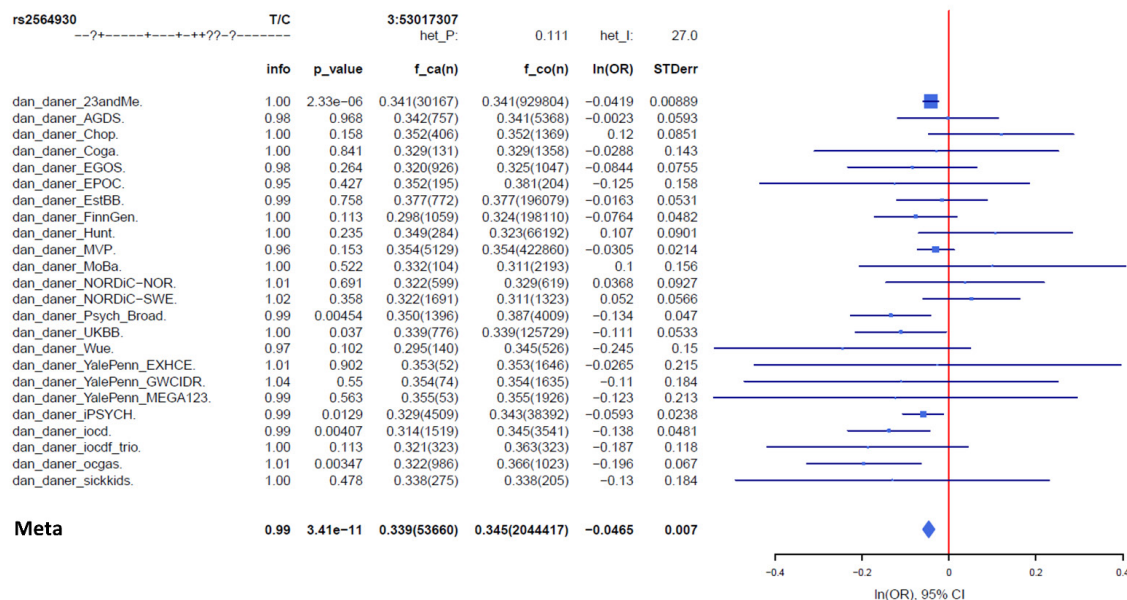


Supplementary Figure 5: Forest plot (A) and regional association plot (B) of SNP rs10877425 in the meta-analysis. The forest plot (A) shows the imputation quality (INFO) score, p-value of the SNP association (two-sided and not corrected for multiple testing, derived from a fixed-effects inverse-variance weighted meta-analysis), allele frequency in cases with case sample size (f_{ca}(n)), allele frequency in controls and control sample size (f_{co}(n)), beta estimates (ln(OR)) and standard error (STDerr) for each study as well as for the combined meta-analysis. Each cohort's sample size is detailed in supplementary table 1. Data on the right show the ln(OR) value as listed per cohort with 95% confidence intervals. In the regional association plot (B), the -log₁₀ p-values (from two-sided association tests) are shown on the left y-axis. Recombination rates expressed in centimorgans (cM) per Mb (Megabase) (blue line) are shown on the right y-axis. Position in Mb is on the x-axis, with genes shown below the regional association plot. Only SNPs with an association p-value less than 0.1 were plotted. The most associated SNP is shown as a diamond and marked with the letter a.

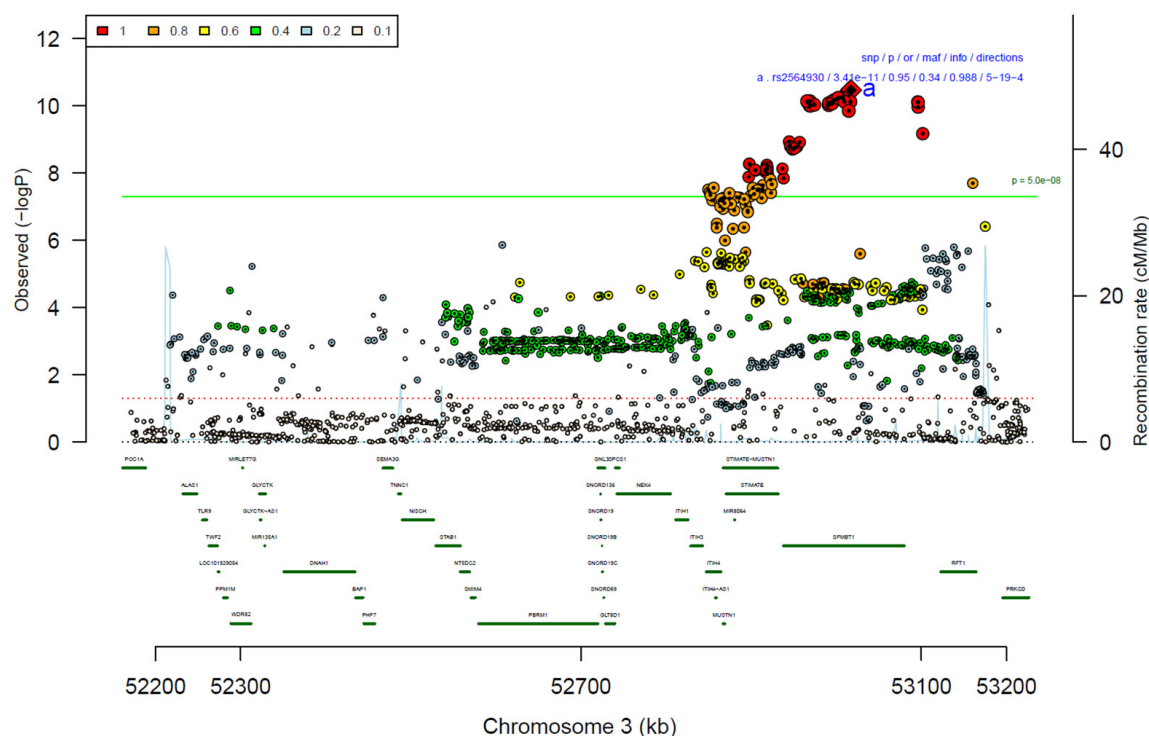


Supplementary Figure 6: Forest plot (A) and regional association plot (B) of SNP rs7626445 in the meta-analysis. The forest plot (A) shows the imputation quality (INFO) score, p-value of the SNP association (two-sided and not corrected for multiple testing, derived from a fixed-effects inverse-variance weighted meta-analysis), allele frequency in cases with case sample size ($f_{ca}(n)$), allele frequency in controls and control sample size ($f_{co}(n)$), beta estimates ($\ln(OR)$) and standard error (STDer) for each study as well as for the combined meta-analysis. Each cohort's sample size is detailed in supplementary table 1. Data on the right show the $\ln(OR)$ value as listed per cohort with 95% confidence intervals. In the regional association plot (B), the $-\log_{10}$ p-values (from two-sided association tests) are shown on the left y-axis. Recombination rates expressed in centimorgans (cM) per Mb (Megabase) (blue line) are shown on the right y-axis. Position in Mb is on the x-axis, with genes shown below the regional association plot. Only SNPs with an association p-value less than 0.1 were plotted. The most associated SNP is shown as a diamond and marked with the letter a.

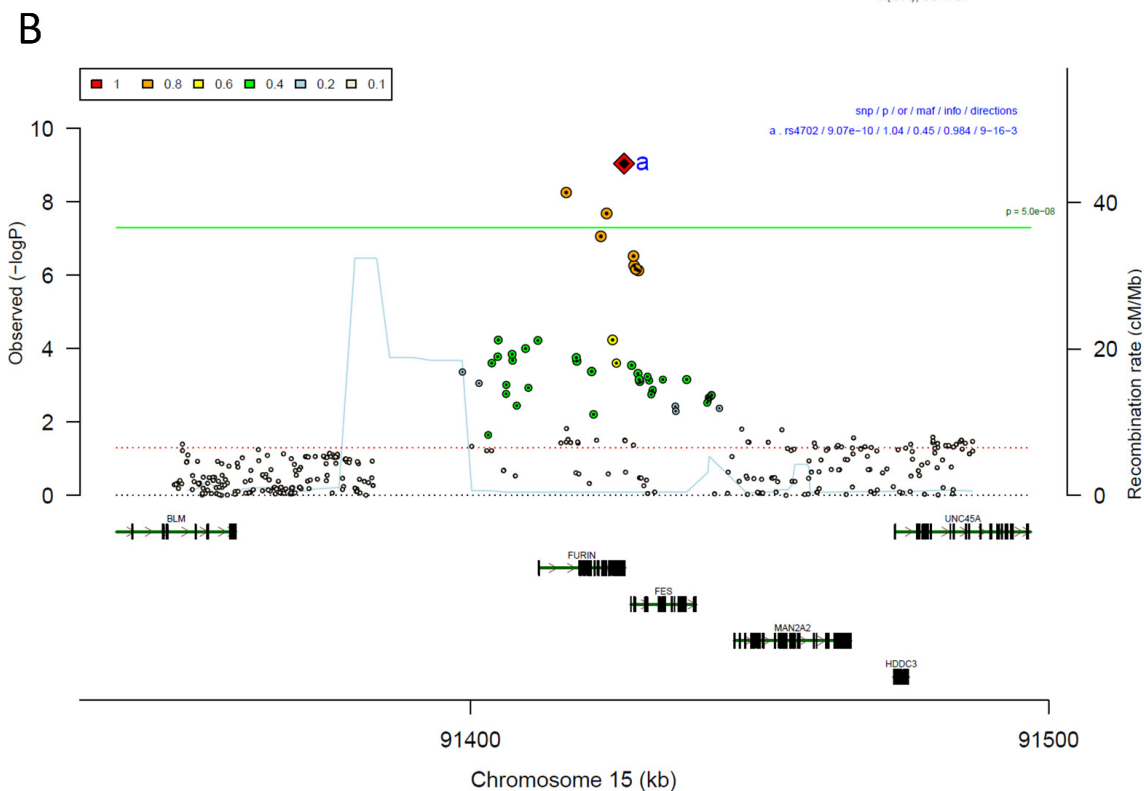
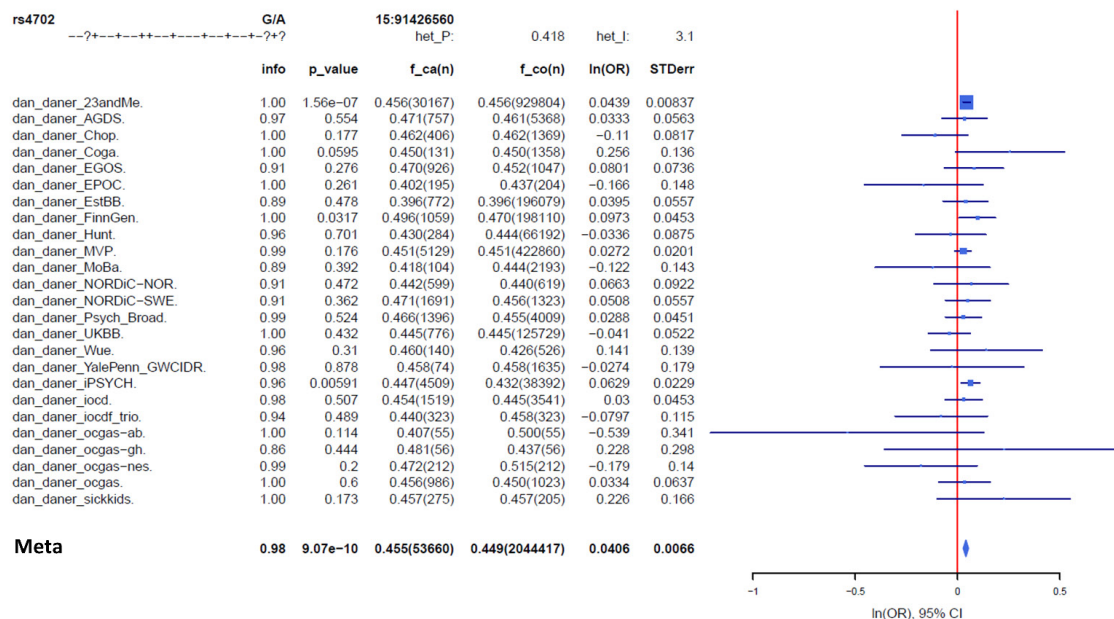
A



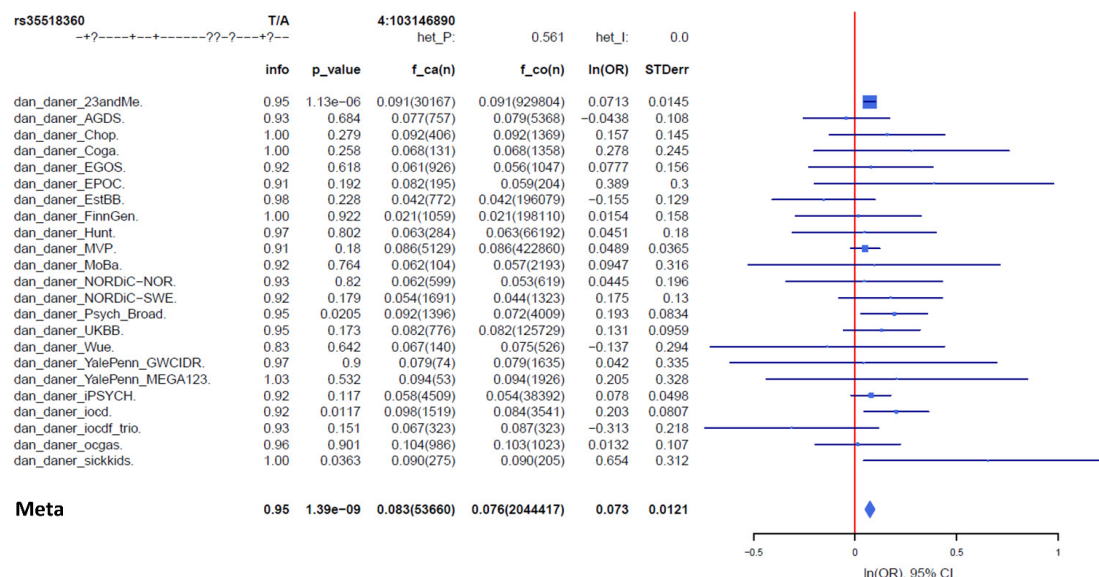
B



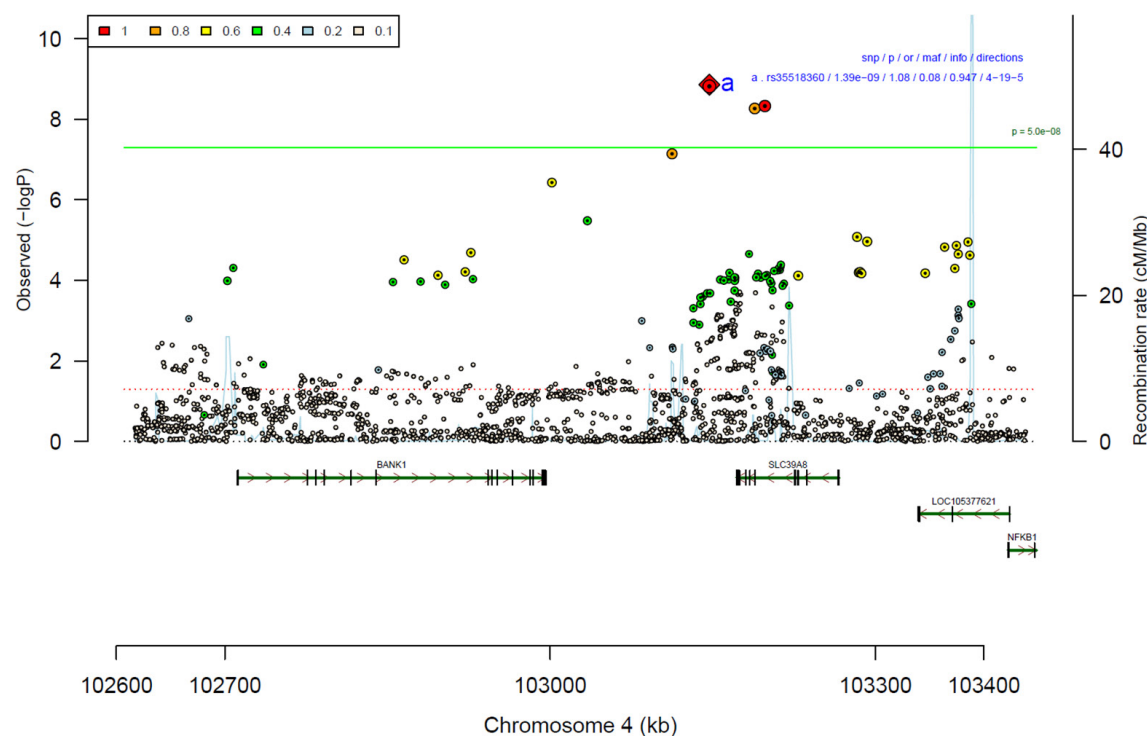
Supplementary Figure 7: Forest plot (A) and regional association plot (B) of SNP rs2564930 in the meta-analysis. The forest plot (A) shows the imputation quality (INFO) score, p-value of the SNP association (two-sided and not corrected for multiple testing, derived from a fixed-effects inverse-variance weighted meta-analysis), allele frequency in cases with case sample size ($f_{ca}(n)$), allele frequency in controls and control sample size ($f_{co}(n)$), beta estimates ($\ln(OR)$) and standard error (STDerr) for each study as well as for the combined meta-analysis. Each cohort's sample size is detailed in supplementary table 1. Data on the right show the $\ln(OR)$ value as listed per cohort with 95% confidence intervals. In the regional association plot (B), the $-\log_{10}$ p-values (from two-sided association tests) are shown on the left y-axis. Recombination rates expressed in centimorgans (cM) per Mb (Megabase) (blue line) are shown on the right y-axis. Position in Mb is on the x-axis, with genes shown below the regional association plot. Only SNPs with an association p-value less than 0.1 were plotted. The most associated SNP is shown as a diamond and marked with the letter a.



A

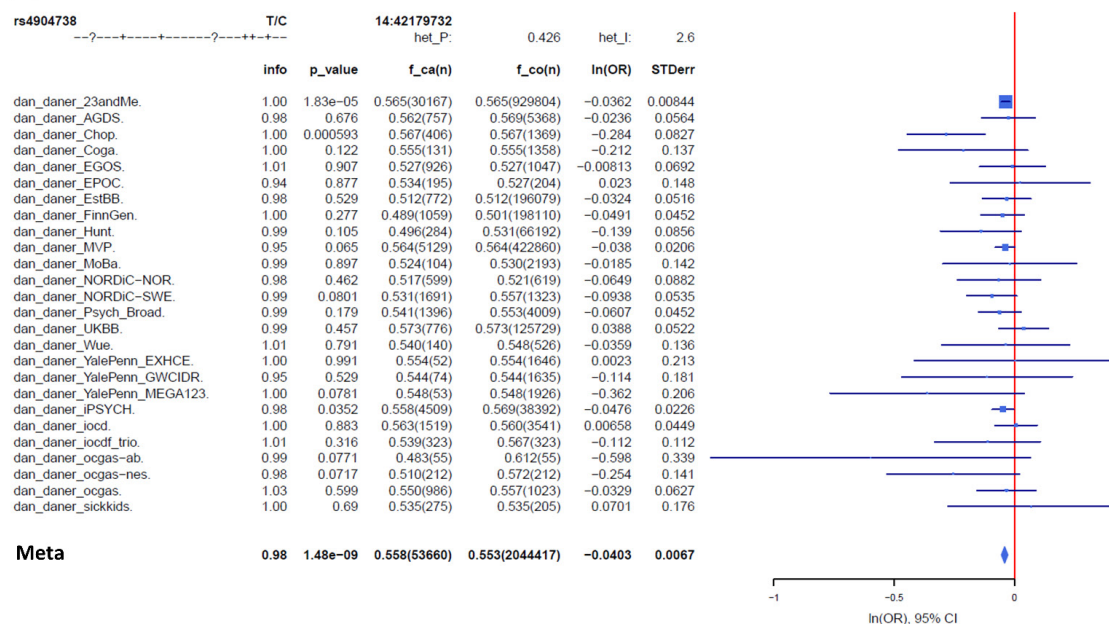


B

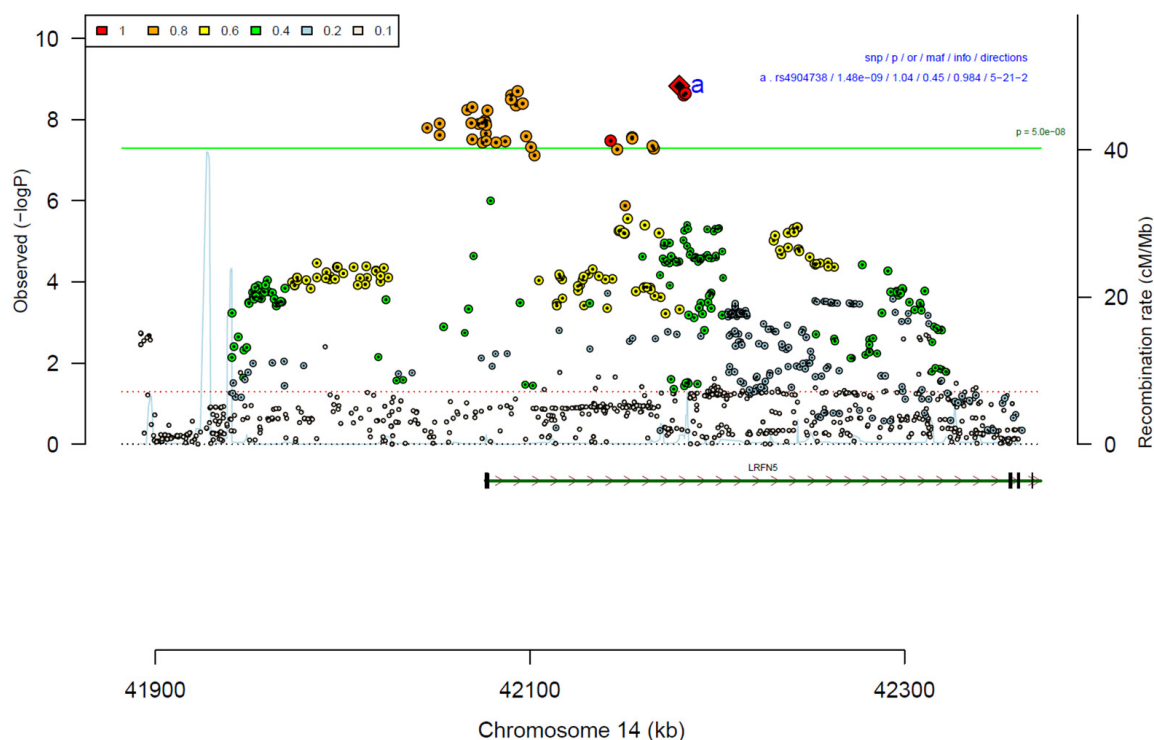


Supplementary Figure 9: Forest plot (A) and regional association plot (B) of SNP rs35518360 in the meta-analysis. The forest plot (A) shows the imputation quality (INFO) score, p-value of the SNP association (two-sided and not corrected for multiple testing, derived from a fixed-effects inverse-variance weighted meta-analysis), allele frequency in cases with case sample size ($f_{ca}(n)$), allele frequency in controls and control sample size ($f_{co}(n)$), beta estimates ($\ln(OR)$) and standard error (STDerr) for each study as well as for the combined meta-analysis. Each cohort's sample size is detailed in supplementary table 1. Data on the right show the $\ln(OR)$ value as listed per cohort with 95% confidence intervals. In the regional association plot (B), the $-\log_{10}$ p-values (from two-sided association tests) are shown on the left y-axis. Recombination rates expressed in centimorgans (cM) per Mb (Megabase) (blue line) are shown on the right y-axis. Position in Mb is on the x-axis, with genes shown below the regional association plot. Only SNPs with an association p-value less than 0.1 were plotted. The most associated SNP is shown as a diamond and marked with the letter a.

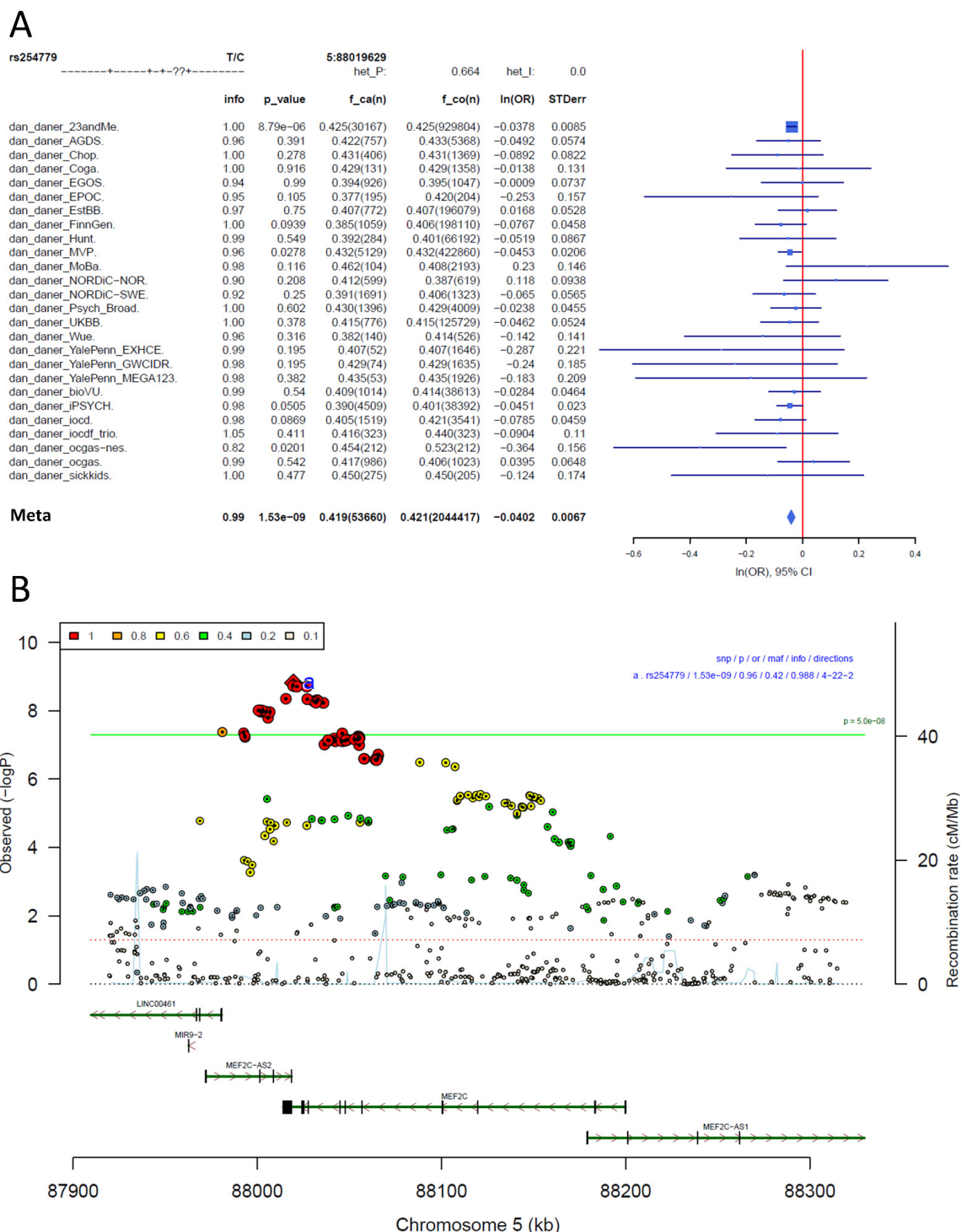
A



B

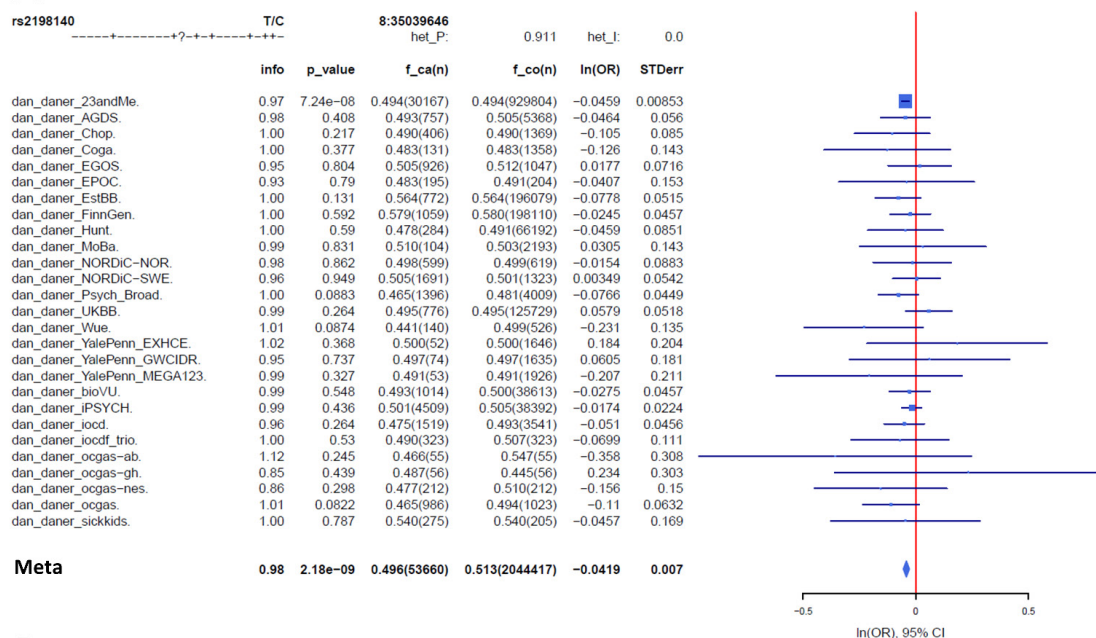


Supplementary Figure 10: Forest plot (A) and regional association plot (B) of SNP rs4904738 in the meta-analysis. The forest plot (A) shows the imputation quality (INFO) score, p-value of the SNP association (two-sided and not corrected for multiple testing, derived from a fixed-effects inverse-variance weighted meta-analysis), allele frequency in cases with case sample size ($f_{ca}(n)$), allele frequency in controls and control sample size ($f_{co}(n)$), beta estimates ($\ln(OR)$) and standard error (STDerr) for each study as well as for the combined meta-analysis. Each cohort's sample size is detailed in supplementary table 1. Data on the right show the $\ln(OR)$ value as listed per cohort with 95% confidence intervals. In the regional association plot (B), the $-\log_{10}$ p-values (from two-sided association tests) are shown on the left y-axis. Recombination rates expressed in centimorgans (cM) per Mb (Megabase) (blue line) are shown on the right y-axis. Position in Mb is on the x-axis, with genes shown below the regional association plot. Only SNPs with an association p-value less than 0.1 were plotted. The most associated SNP is shown as a diamond and marked with the letter a.

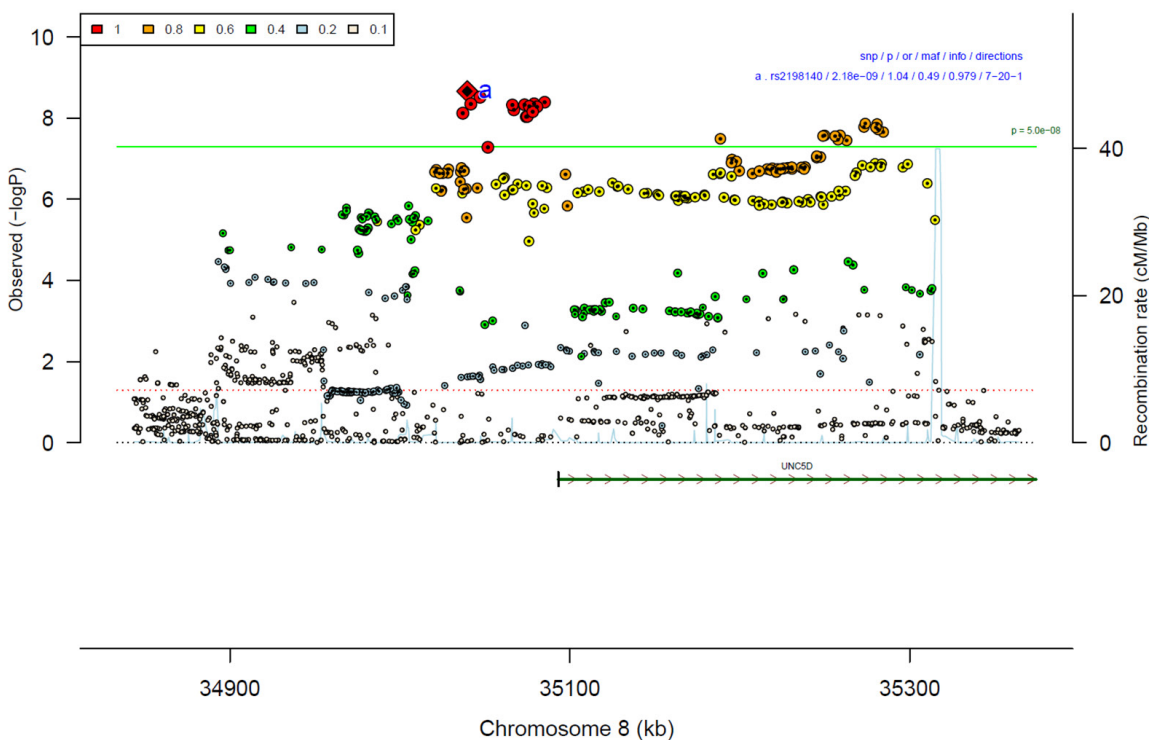


Supplementary Figure 11: Forest plot (A) and regional association plot (B) of SNP rs254779 in the meta-analysis. The forest plot (A) shows the imputation quality (INFO) score, p-value of the SNP association (two-sided and not corrected for multiple testing, derived from a fixed-effects inverse-variance weighted meta-analysis), allele frequency in cases with case sample size ($f_{ca}(n)$), allele frequency in controls and control sample size ($f_{co}(n)$), beta estimates ($\ln(OR)$) and standard error (STDerr) for each study as well as for the combined meta-analysis. Each cohort's sample size is detailed in supplementary table 1. Data on the right show the $\ln(OR)$ value as listed per cohort with 95% confidence intervals. In the regional association plot (B), the $-\log_{10}$ p-values (from two-sided association tests) are shown on the left y-axis. Recombination rates expressed in centimorgans (cM) per Mb (Megabase) (blue line) are shown on the right y-axis. Position in Mb is on the x-axis, with genes shown below the regional association plot. Only SNPs with an association p-value less than 0.1 were plotted. The most associated SNP is shown as a diamond and marked with the letter a.

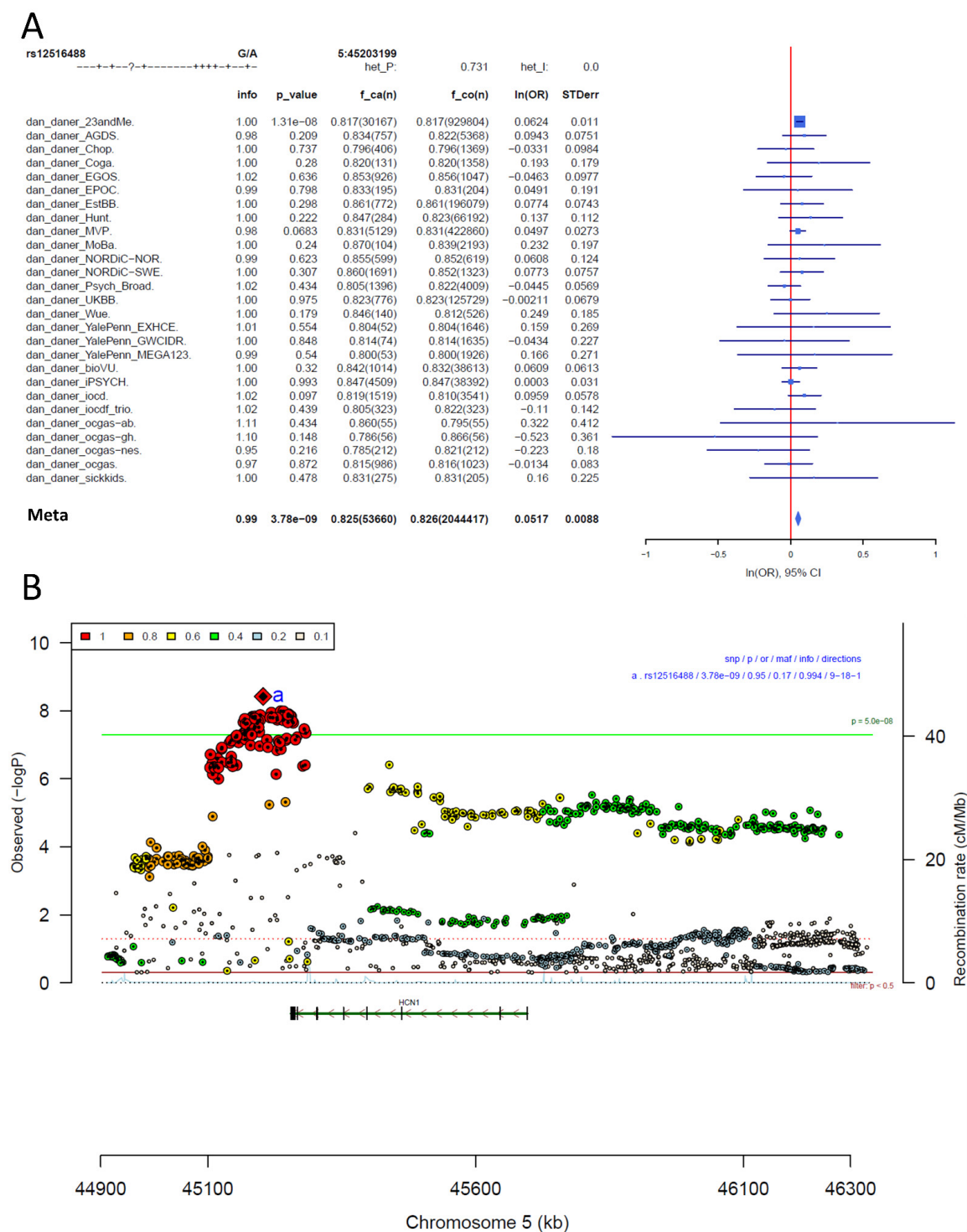
A



B

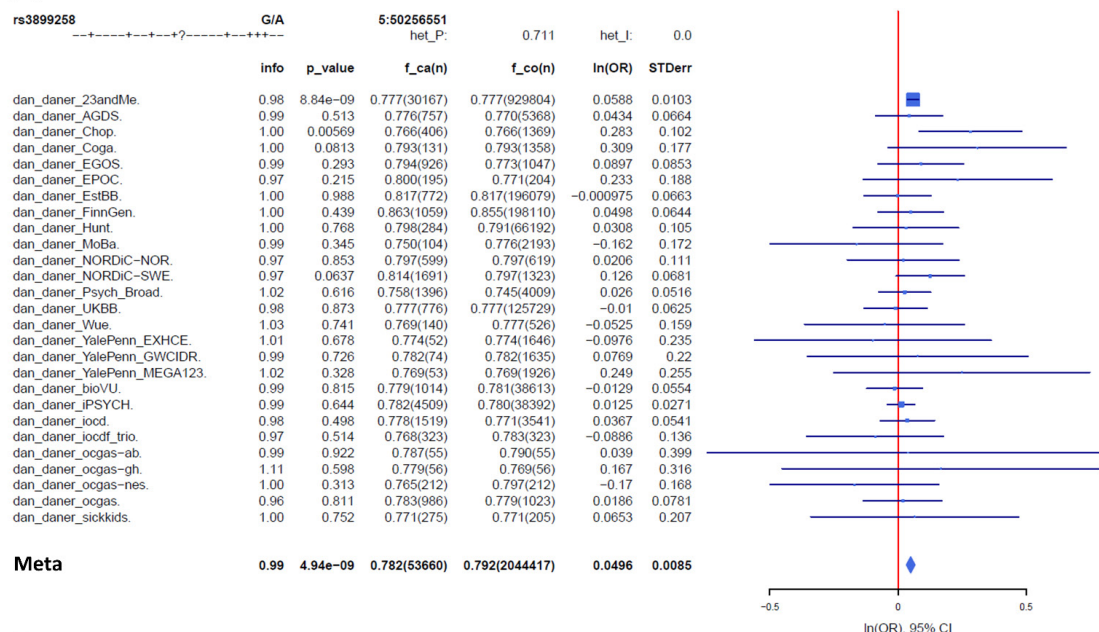


Supplementary Figure 12: Forest plot (A) and regional association plot (B) of SNP rs2198140 in the meta-analysis. The forest plot (A) shows the imputation quality (INFO) score, p-value of the SNP association (two-sided and not corrected for multiple testing, derived from a fixed-effects inverse-variance weighted meta-analysis), allele frequency in cases with case sample size ($f_{ca}(n)$), allele frequency in controls and control sample size ($f_{co}(n)$), beta estimates ($\ln(OR)$) and standard error (STDerr) for each study as well as for the combined meta-analysis. Each cohort's sample size is detailed in supplementary table 1. Data on the right show the $\ln(OR)$ value as listed per cohort with 95% confidence intervals. In the regional association plot (B), the $-\log_{10}$ p-values (from two-sided association tests) are shown on the left y-axis. Recombination rates expressed in centimorgans (cM) per Mb (Megabase) (blue line) are shown on the right y-axis. Position in Mb is on the x-axis, with genes shown below the regional association plot. Only SNPs with an association p-value less than 0.1 were plotted. The most associated SNP is shown as a diamond and marked with the letter a.

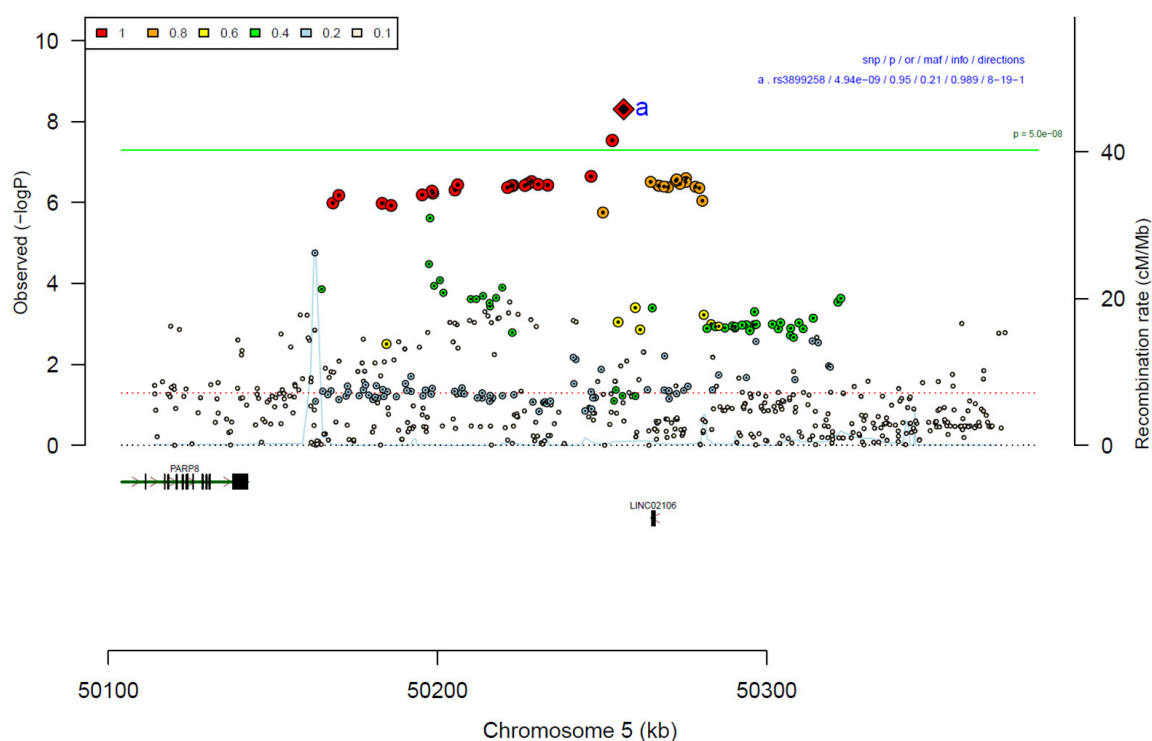


Supplementary Figure 13: Forest plot (A) and regional association plot (B) of SNP rs12516488 in the meta-analysis. The forest plot (A) shows the imputation quality (INFO) score, p-value of the SNP association (two-sided and not corrected for multiple testing, derived from a fixed-effects inverse-variance weighted meta-analysis), allele frequency in cases with case sample size ($f_{ca}(n)$), allele frequency in controls and control sample size ($f_{co}(n)$), beta estimates ($\ln(OR)$) and standard error (STDerr) for each study as well as for the combined meta-analysis. Each cohort's sample size is detailed in supplementary table 1. Data on the right show the $\ln(OR)$ value as listed per cohort with 95% confidence intervals. In the regional association plot (B), the $-\log_{10}$ p-values (from two-sided association tests) are shown on the left y-axis. Recombination rates expressed in centimorgans (cM) per Mb (Megabase) (blue line) are shown on the right y-axis. Position in Mb is on the x-axis, with genes shown below the regional association plot. Only SNPs with an association p-value less than 0.1 were plotted. The most associated SNP is shown as a diamond and marked with the letter a.

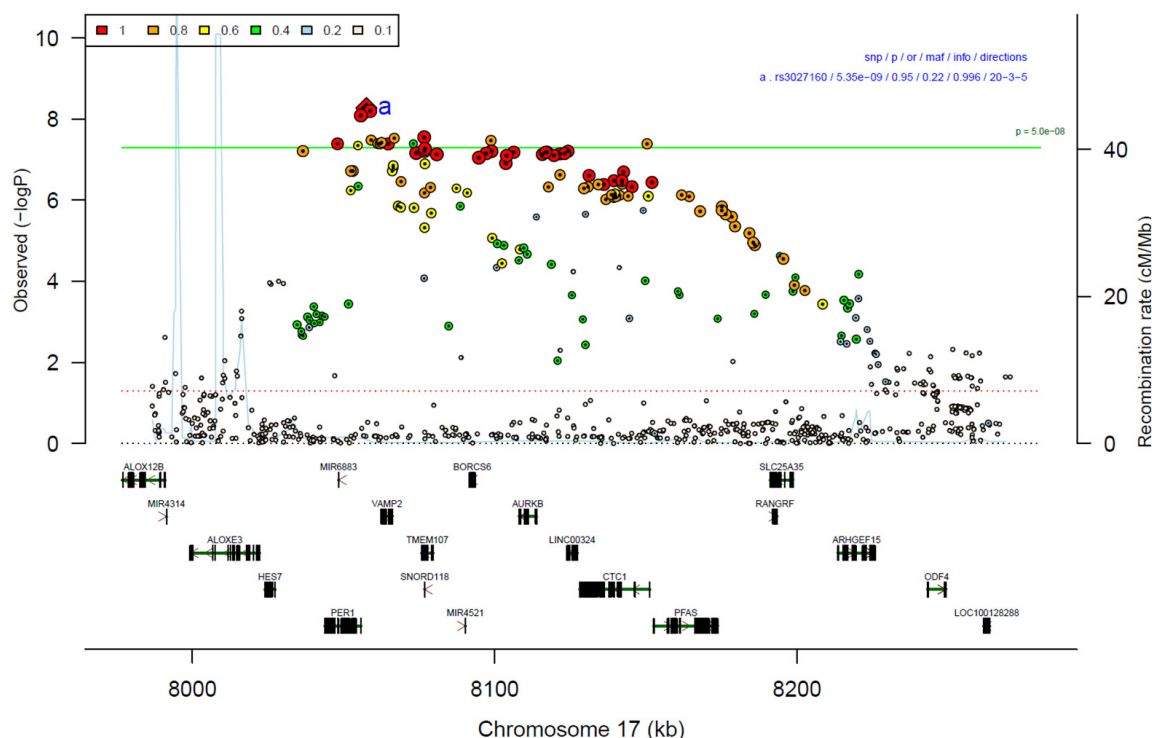
A



B

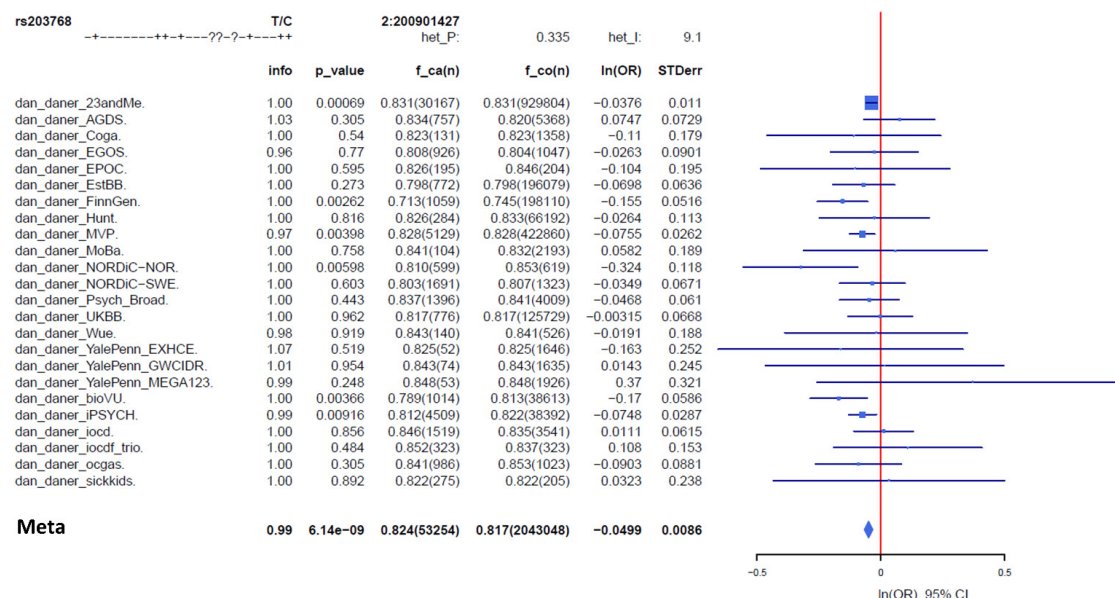


Supplementary Figure 14: Forest plot (A) and regional association plot (B) of SNP rs3899258 in the meta-analysis. The forest plot (A) shows the imputation quality (INFO) score, p-value of the SNP association (two-sided and not corrected for multiple testing, derived from a fixed-effects inverse-variance weighted meta-analysis), allele frequency in cases with case sample size ($f_{ca}(n)$), allele frequency in controls and control sample size ($f_{co}(n)$), beta estimates ($\ln(OR)$) and standard error (STDerr) for each study as well as for the combined meta-analysis. Each cohort's sample size is detailed in supplementary table 1. Data on the right show the $\ln(OR)$ value as listed per cohort with 95% confidence intervals. In the regional association plot (B), the $-\log_{10}$ p-values (from two-sided association tests) are shown on the left y-axis. Recombination rates expressed in centimorgans (cM) per Mb (Megabase) (blue line) are shown on the right y-axis. Position in Mb is on the x-axis, with genes shown below the regional association plot. Only SNPs with an association p-value less than 0.1 were plotted. The most associated SNP is shown as a diamond and marked with the letter a.

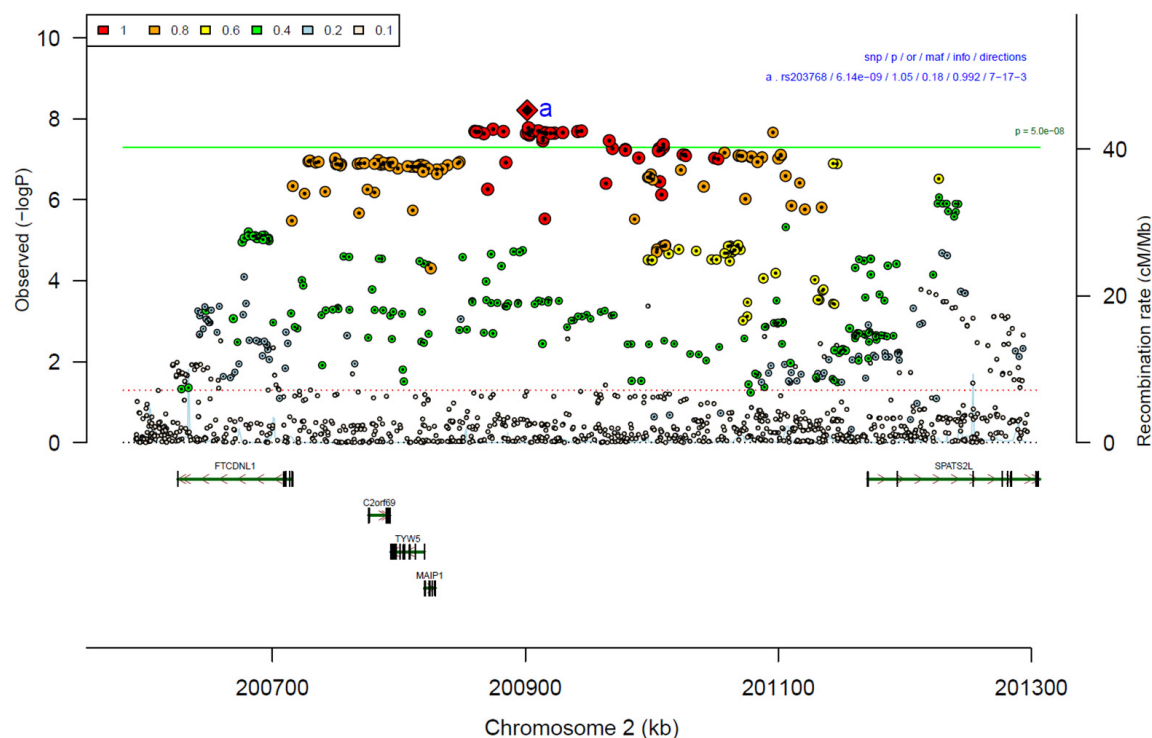


Supplementary Figure 15: Forest plot (A) and regional association plot (B) of SNP rs3027160 in the meta-analysis. The forest plot (A) shows the imputation quality (INFO) score, p-value of the SNP association (two-sided and not corrected for multiple testing, derived from a fixed-effects inverse-variance weighted meta-analysis), allele frequency in cases with case sample size ($f_{ca}(n)$), allele frequency in controls and control sample size ($f_{co}(n)$), beta estimates ($\ln(OR)$) and standard error (STDerr) for each study as well as for the combined meta-analysis. Each cohort's sample size is detailed in supplementary table 1. Data on the right show the $\ln(OR)$ value as listed per cohort with 95% confidence intervals. In the regional association plot (B), the $-\log_{10}$ p-values (from two-sided association tests) are shown on the left y-axis. Recombination rates expressed in centimorgans (cM) per Mb (Megabase) (blue line) are shown on the right y-axis. Position in Mb is on the x-axis, with genes shown below the regional association plot. Only SNPs with an association p-value less than 0.1 were plotted. The most associated SNP is shown as a diamond and marked with the letter a.

A

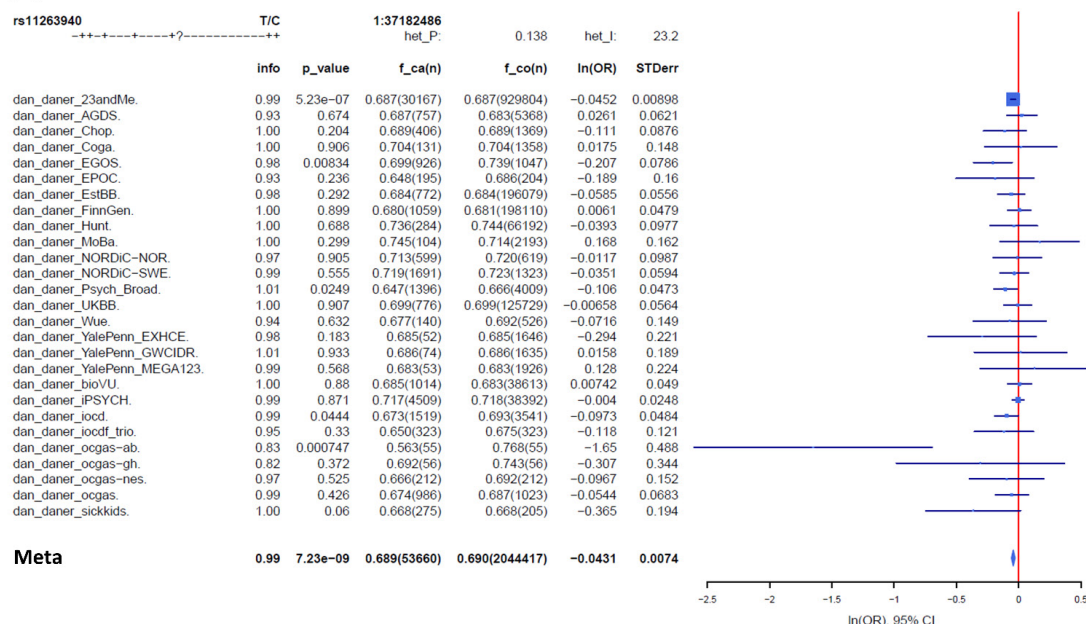


B

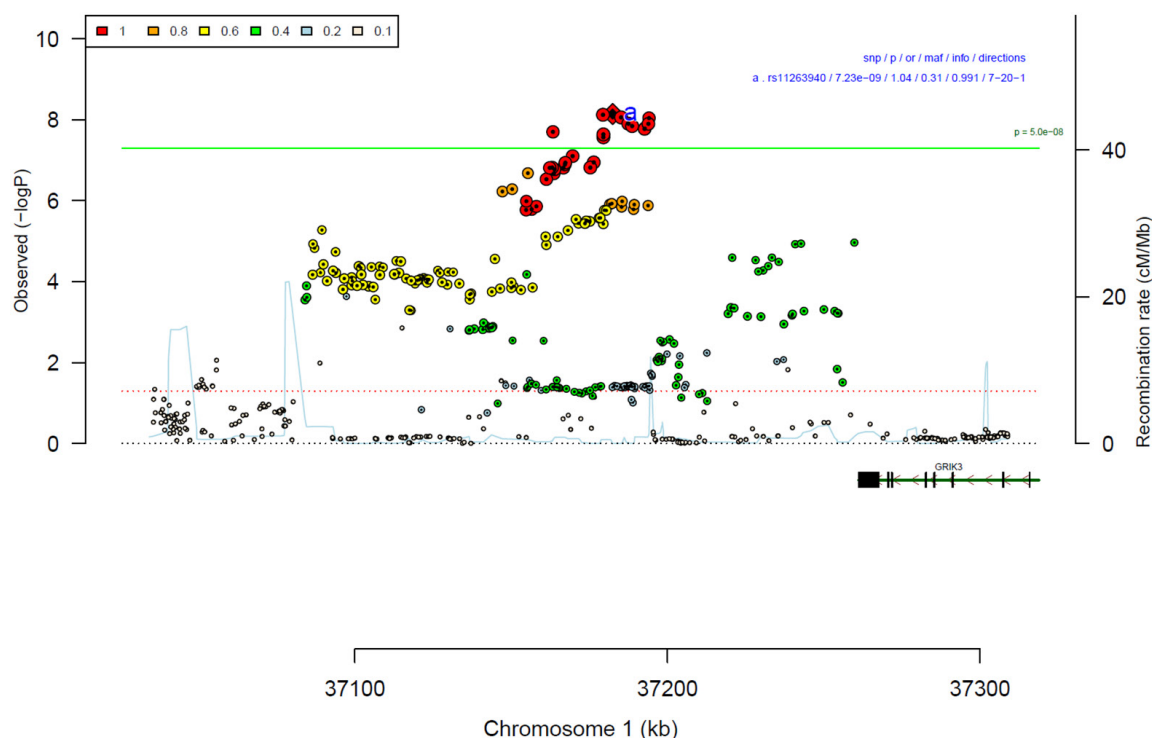


Supplementary Figure 16: Forest plot (A) and regional association plot (B) of SNP rs203768 in the meta-analysis. The forest plot (A) shows the imputation quality (INFO) score, p-value of the SNP association (two-sided and not corrected for multiple testing, derived from a fixed-effects inverse-variance weighted meta-analysis), allele frequency in cases with case sample size ($f_{ca}(n)$), allele frequency in controls and control sample size ($f_{co}(n)$), beta estimates ($\ln(OR)$) and standard error (STDerr) for each study as well as for the combined meta-analysis. Each cohort's sample size is detailed in supplementary table 1. Data on the right show the $\ln(OR)$ value as listed per cohort with 95% confidence intervals. In the regional association plot (B), the $-\log_{10}$ p-values (from two-sided association tests) are shown on the left y-axis. Recombination rates expressed in centimorgans (cM) per Mb (Megabase) (blue line) are shown on the right y-axis. Position in Mb is on the x-axis, with genes shown below the regional association plot. Only SNPs with an association p-value less than 0.1 were plotted. The most associated SNP is shown as a diamond and marked with the letter a.

A

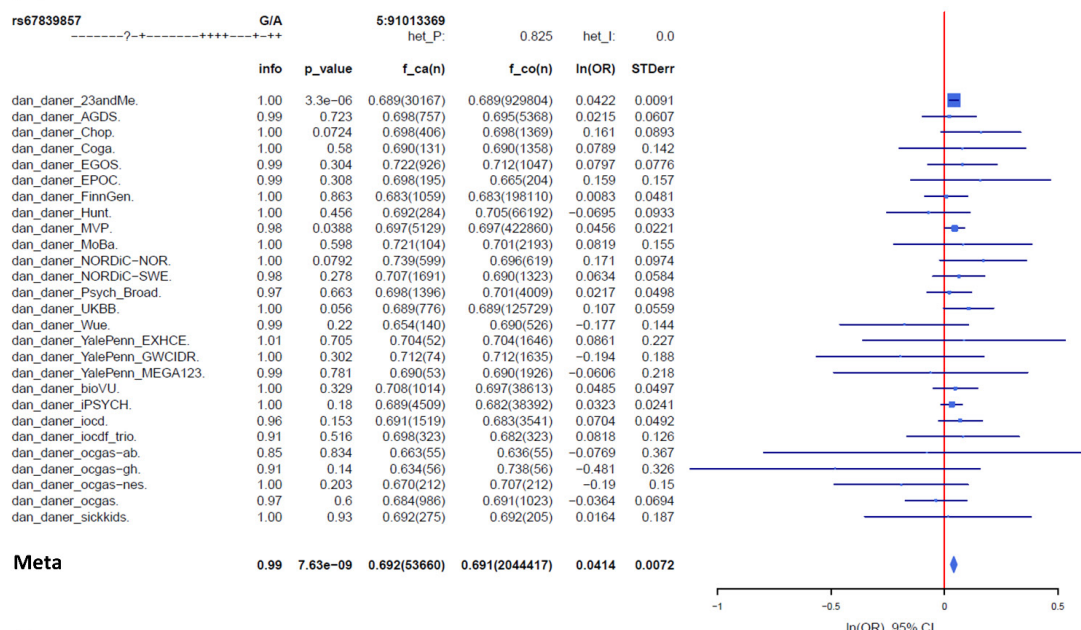


B

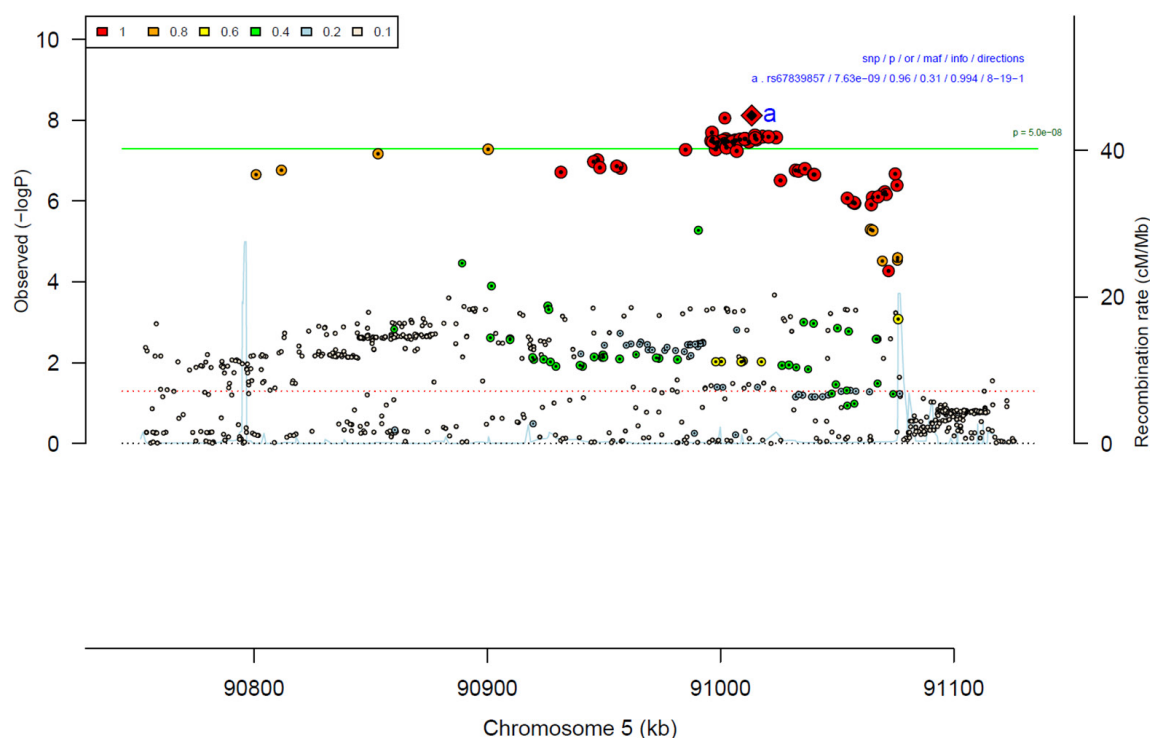


Supplementary Figure 17: Forest plot (A) and regional association plot (B) of SNP rs11263940 in the meta-analysis. The forest plot (A) shows the imputation quality (INFO) score, p-value of the SNP association (two-sided and not corrected for multiple testing, derived from a fixed-effects inverse-variance weighted meta-analysis), allele frequency in cases with case sample size ($f_{ca}(n)$), allele frequency in controls and control sample size ($f_{co}(n)$), beta estimates ($\ln(OR)$) and standard error (STDerr) for each study as well as for the combined meta-analysis. Each cohort's sample size is detailed in supplementary table 1. Data on the right show the $\ln(OR)$ value as listed per cohort with 95% confidence intervals. In the regional association plot (B), the $-\log_{10}$ p-values (from two-sided association tests) are shown on the left y-axis. Recombination rates expressed in centimorgans (cM) per Mb (Megabase) (blue line) are shown on the right y-axis. Position in Mb is on the x-axis, with genes shown below the regional association plot. Only SNPs with an association p-value less than 0.1 were plotted. The most associated SNP is shown as a diamond and marked with the letter a.

A

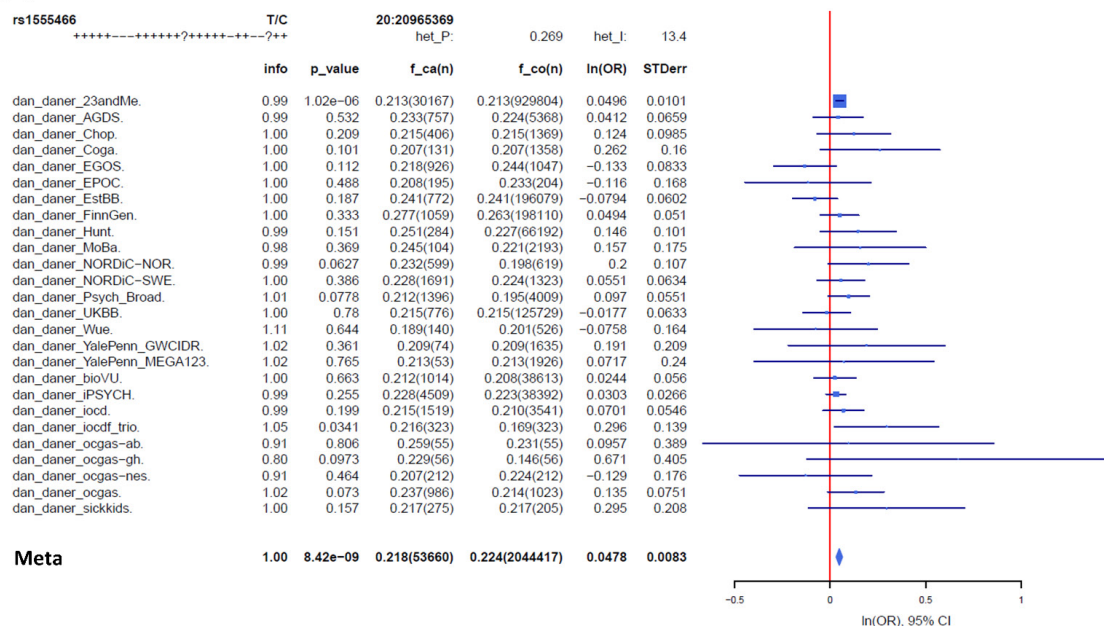


B

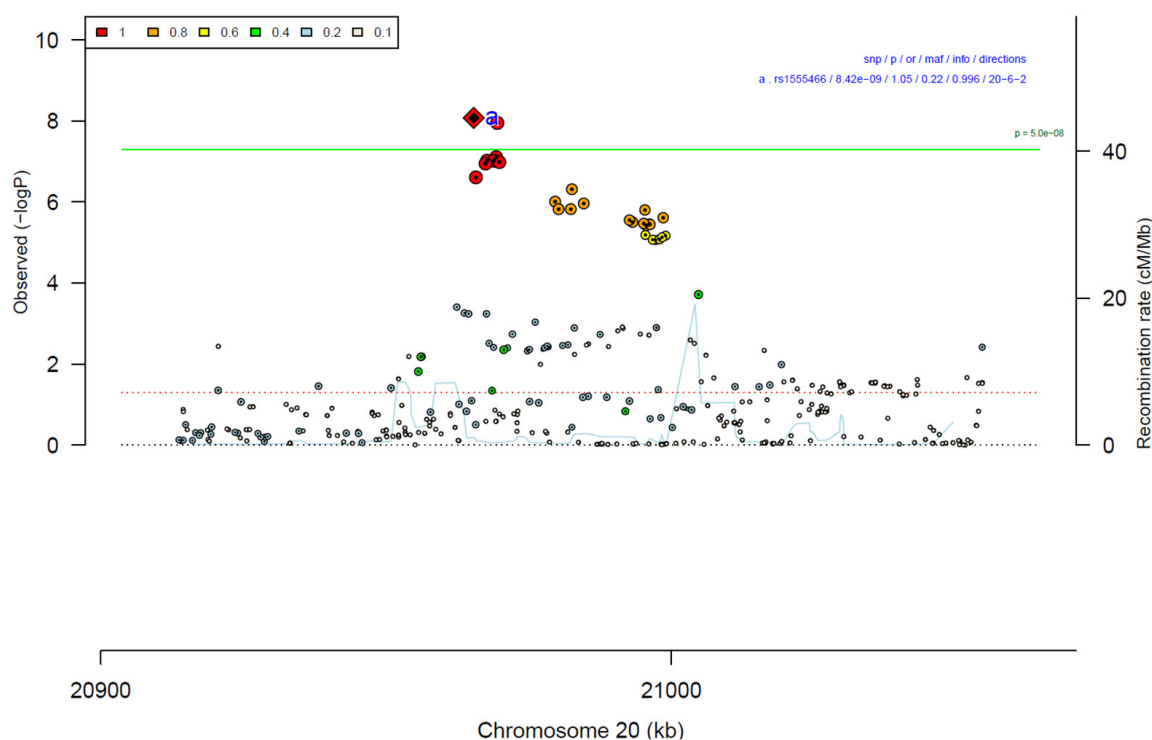


Supplementary Figure 18: Forest plot (A) and regional association plot (B) of SNP rs67839857 in the meta-analysis. The forest plot (A) shows the imputation quality (INFO) score, p-value of the SNP association (two-sided and not corrected for multiple testing, derived from a fixed-effects inverse-variance weighted meta-analysis), allele frequency in cases with case sample size ($f_{ca}(n)$), allele frequency in controls and control sample size ($f_{co}(n)$), beta estimates ($\ln(OR)$) and standard error (STDerr) for each study as well as for the combined meta-analysis. Each cohort's sample size is detailed in supplementary table 1. Data on the right show the $\ln(OR)$ value as listed per cohort with 95% confidence intervals. In the regional association plot (B), the $-\log_{10}$ p-values (from two-sided association tests) are shown on the left y-axis. Recombination rates expressed in centimorgans (cM) per Mb (Megabase) (blue line) are shown on the right y-axis. Position in Mb is on the x-axis, with genes shown below the regional association plot. Only SNPs with an association p-value less than 0.1 were plotted. The most associated SNP is shown as a diamond and marked with the letter a.

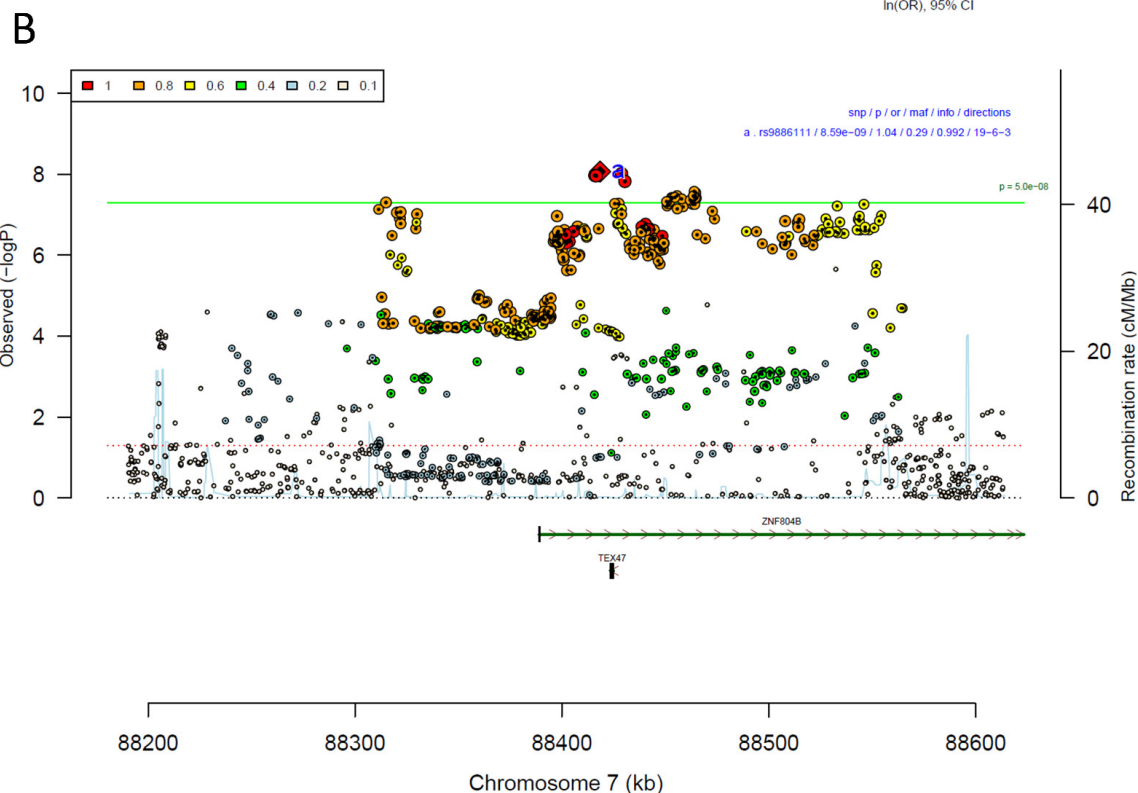
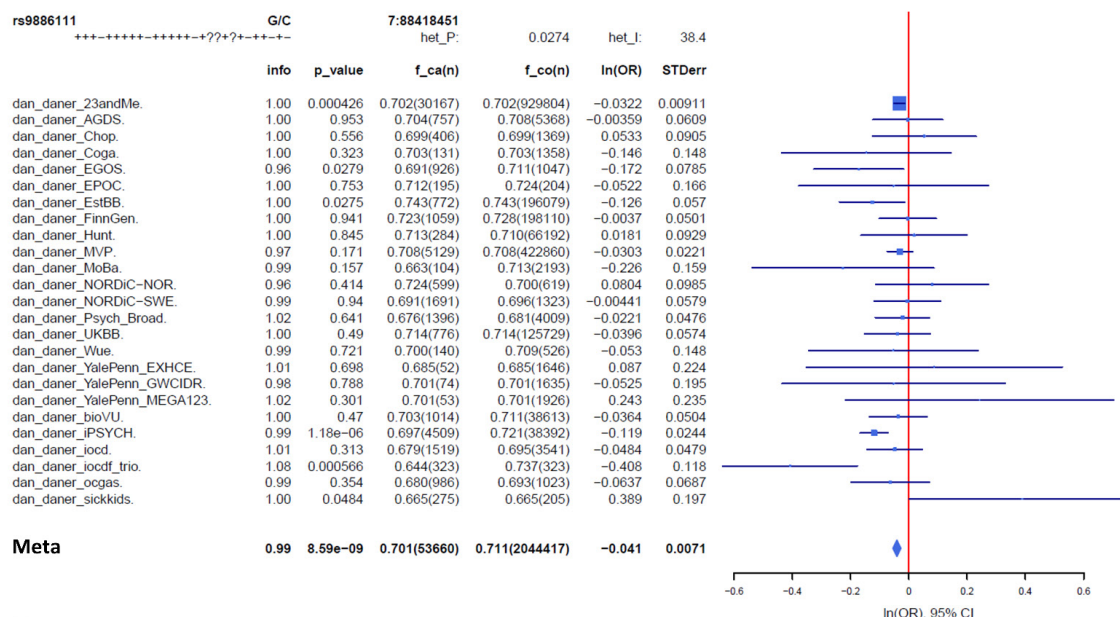
A

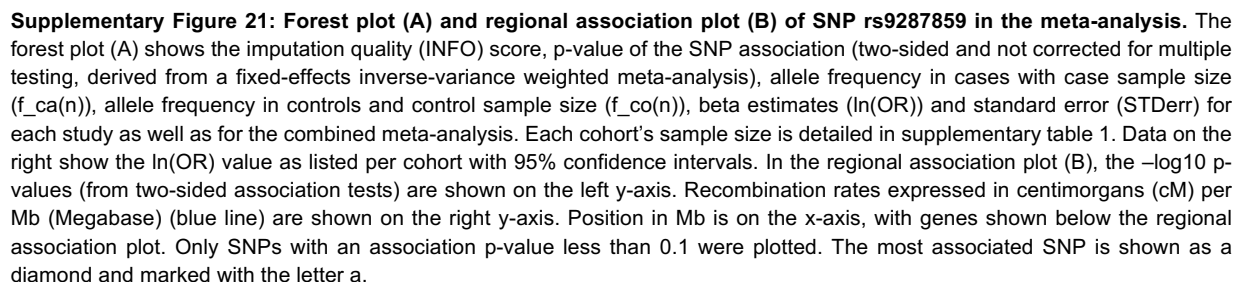


B

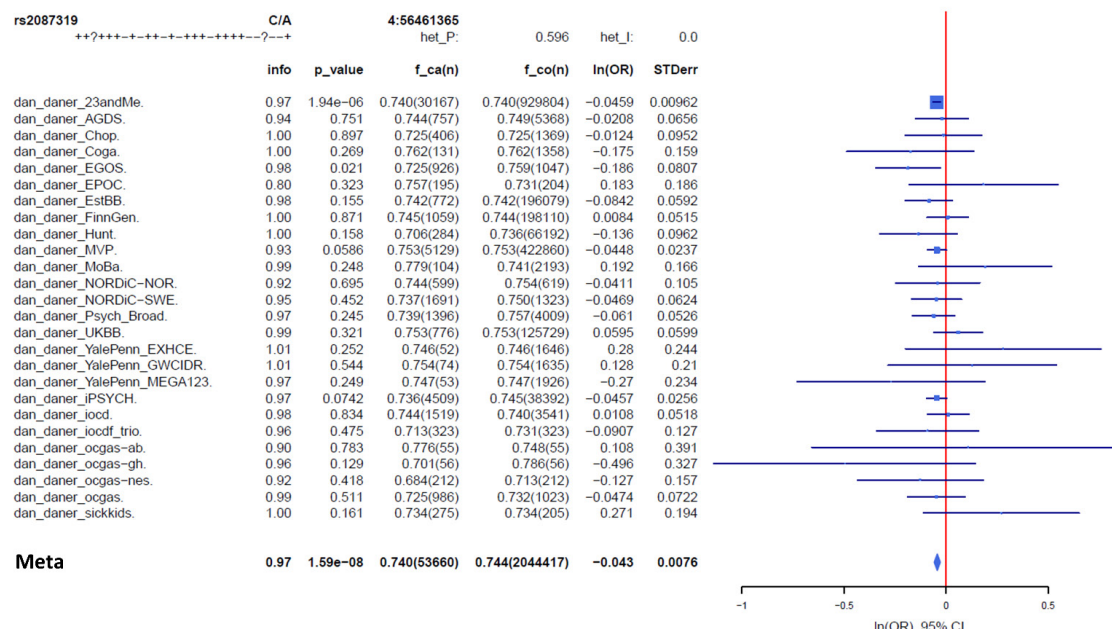


Supplementary Figure 19: Forest plot (A) and regional association plot (B) of SNP rs1555466 in the meta-analysis. The forest plot (A) shows the imputation quality (INFO) score, p-value of the SNP association (two-sided and not corrected for multiple testing, derived from a fixed-effects inverse-variance weighted meta-analysis), allele frequency in cases with case sample size ($f_{ca}(n)$), allele frequency in controls and control sample size ($f_{co}(n)$), beta estimates ($\ln(OR)$) and standard error (STDerr) for each study as well as for the combined meta-analysis. Each cohort's sample size is detailed in supplementary table 1. Data on the right show the $\ln(OR)$ value as listed per cohort with 95% confidence intervals. In the regional association plot (B), the $-\log_{10}$ p-values (from two-sided association tests) are shown on the left y-axis. Recombination rates expressed in centimorgans (cM) per Mb (Megabase) (blue line) are shown on the right y-axis. Position in Mb is on the x-axis, with genes shown below the regional association plot. Only SNPs with an association p-value less than 0.1 were plotted. The most associated SNP is shown as a diamond and marked with the letter a.

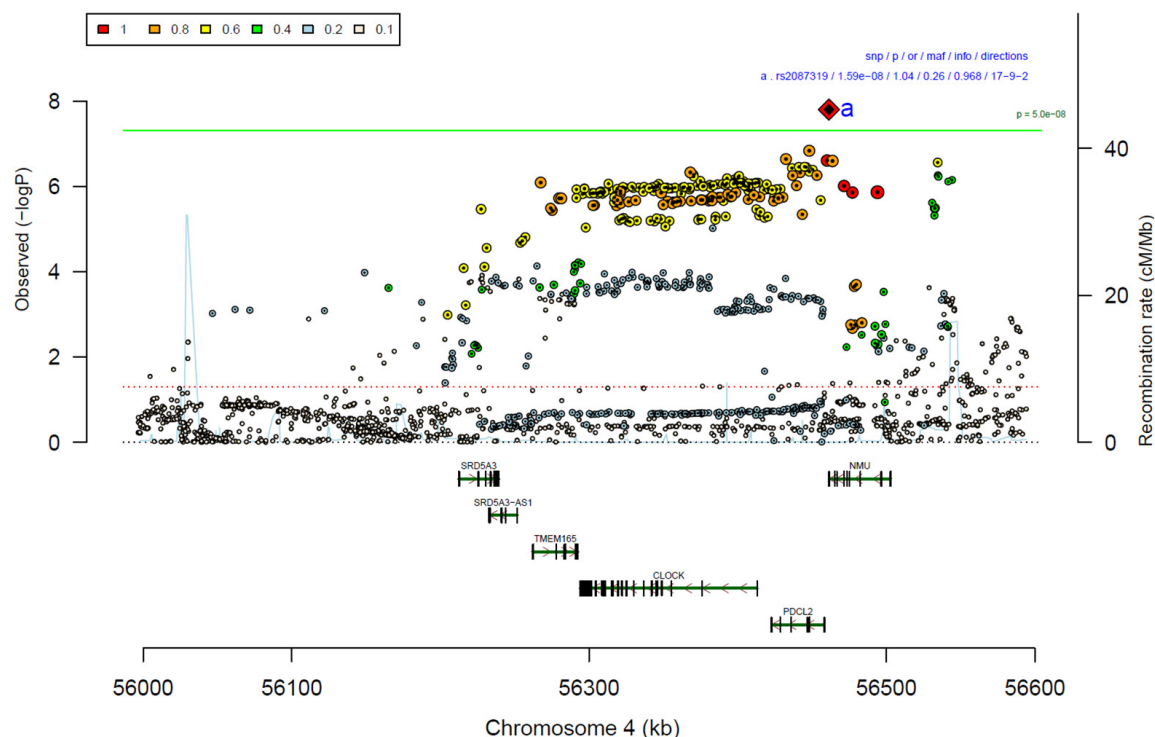




A

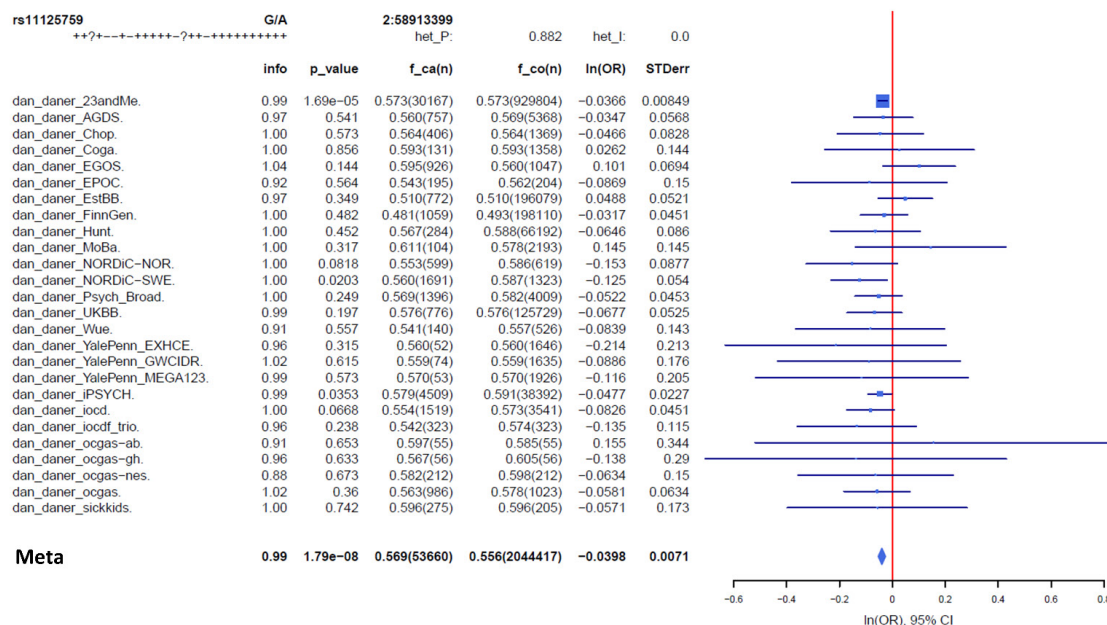


D

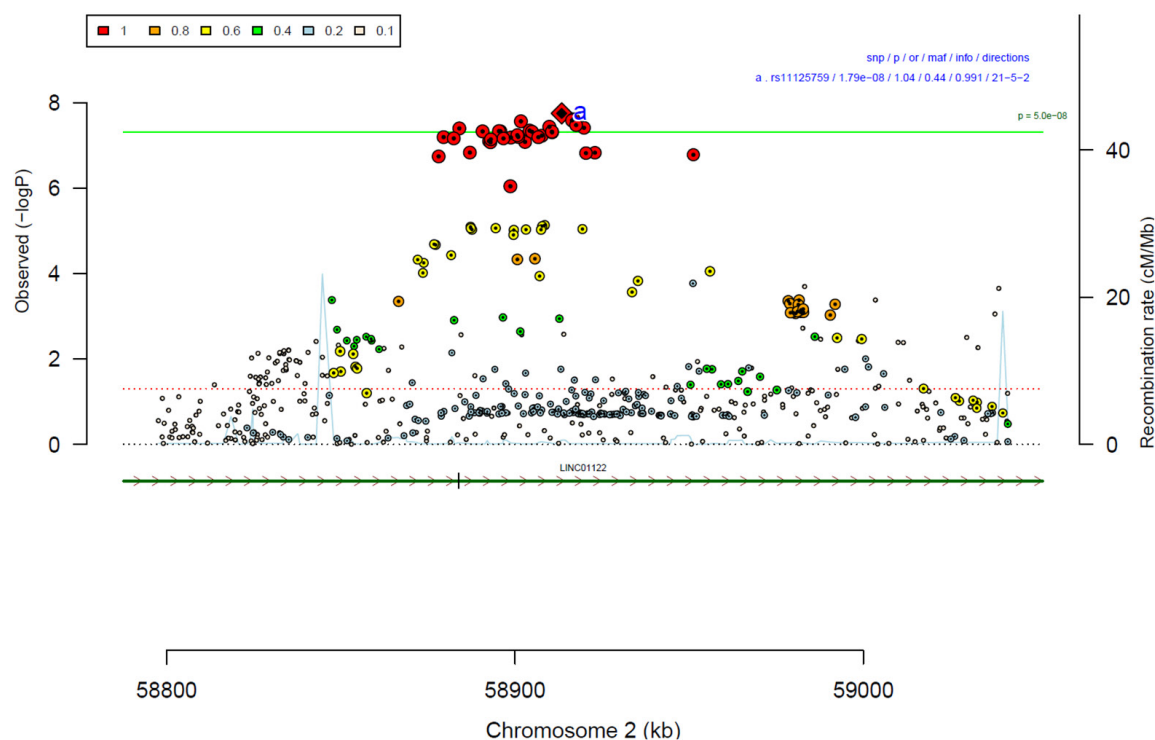


Supplementary Figure 22: Forest plot (A) and regional association plot (B) of SNP rs2087319 in the meta-analysis. The forest plot (A) shows the imputation quality (INFO) score, p-value of the SNP association (two-sided and not corrected for multiple testing, derived from a fixed-effects inverse-variance weighted meta-analysis), allele frequency in cases with case sample size ($f_{ca}(n)$), allele frequency in controls and control sample size ($f_{co}(n)$), beta estimates ($\ln(OR)$) and standard error (STDerr) for each study as well as for the combined meta-analysis. Each cohort's sample size is detailed in supplementary table 1. Data on the right show the $\ln(OR)$ value as listed per cohort with 95% confidence intervals. In the regional association plot (B), the $-\log_{10}$ p-values (from two-sided association tests) are shown on the left y-axis. Recombination rates expressed in centimorgans (cM) per Mb (Megabase) (blue line) are shown on the right y-axis. Position in Mb is on the x-axis, with genes shown below the regional association plot. Only SNPs with an association p-value less than 0.1 were plotted. The most associated SNP is shown as a diamond and marked with the letter a.

A

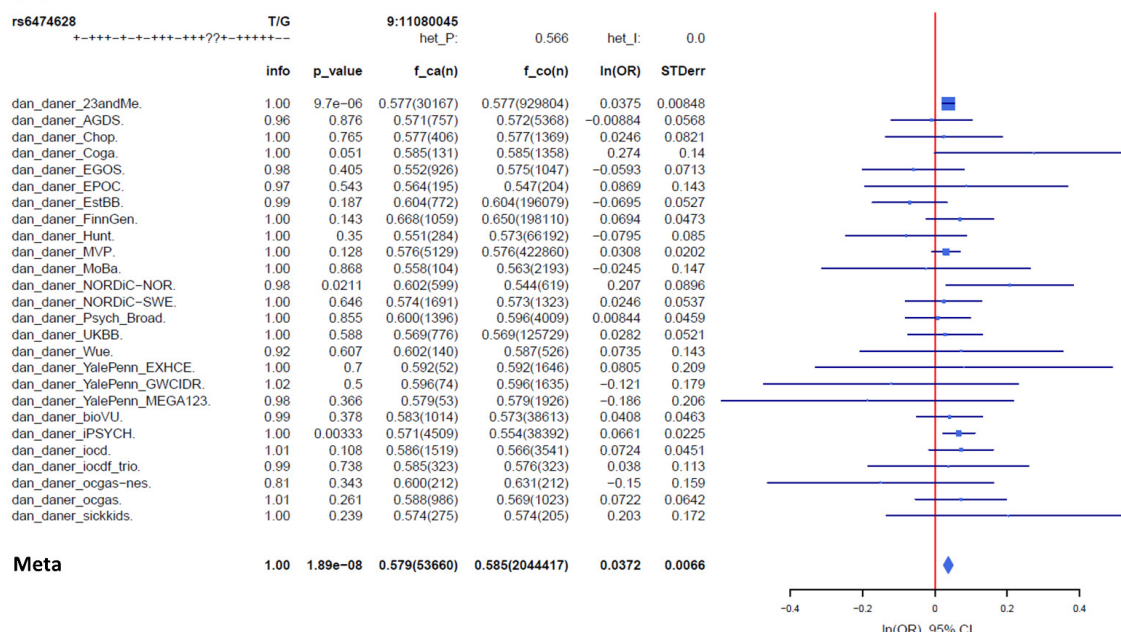


B

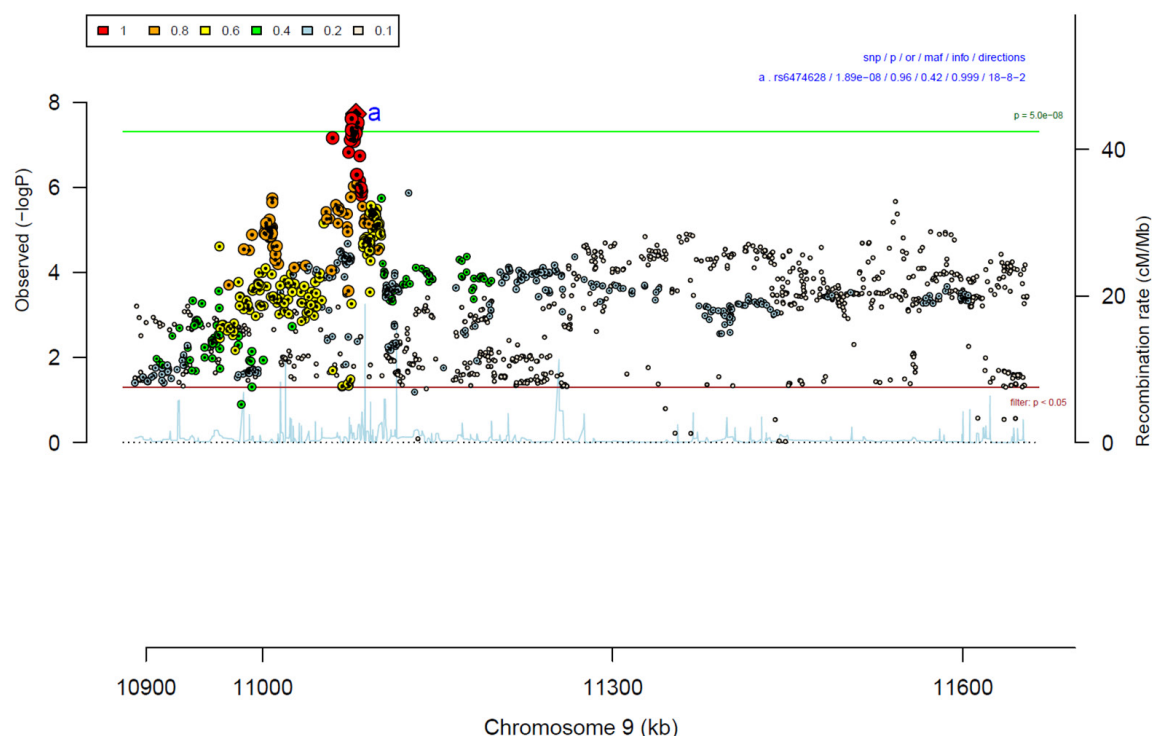


Supplementary Figure 23: Forest plot (A) and regional association plot (B) of SNP rs11125759 in the meta-analysis. The forest plot (A) shows the imputation quality (INFO) score, p-value of the SNP association (two-sided and not corrected for multiple testing, derived from a fixed-effects inverse-variance weighted meta-analysis), allele frequency in cases with case sample size ($f_{ca}(n)$), allele frequency in controls and control sample size ($f_{co}(n)$), beta estimates ($\ln(OR)$) and standard error (STDerr) for each study as well as for the combined meta-analysis. Each cohort's sample size is detailed in supplementary table 1. Data on the right show the $\ln(OR)$ value as listed per cohort with 95% confidence intervals. In the regional association plot (B), the $-\log_{10}$ p-values (from two-sided association tests) are shown on the left y-axis. Recombination rates expressed in centimorgans (cM) per Mb (Megabase) (blue line) are shown on the right y-axis. Position in Mb is on the x-axis, with genes shown below the regional association plot. Only SNPs with an association p-value less than 0.1 were plotted. The most associated SNP is shown as a diamond and marked with the letter a.

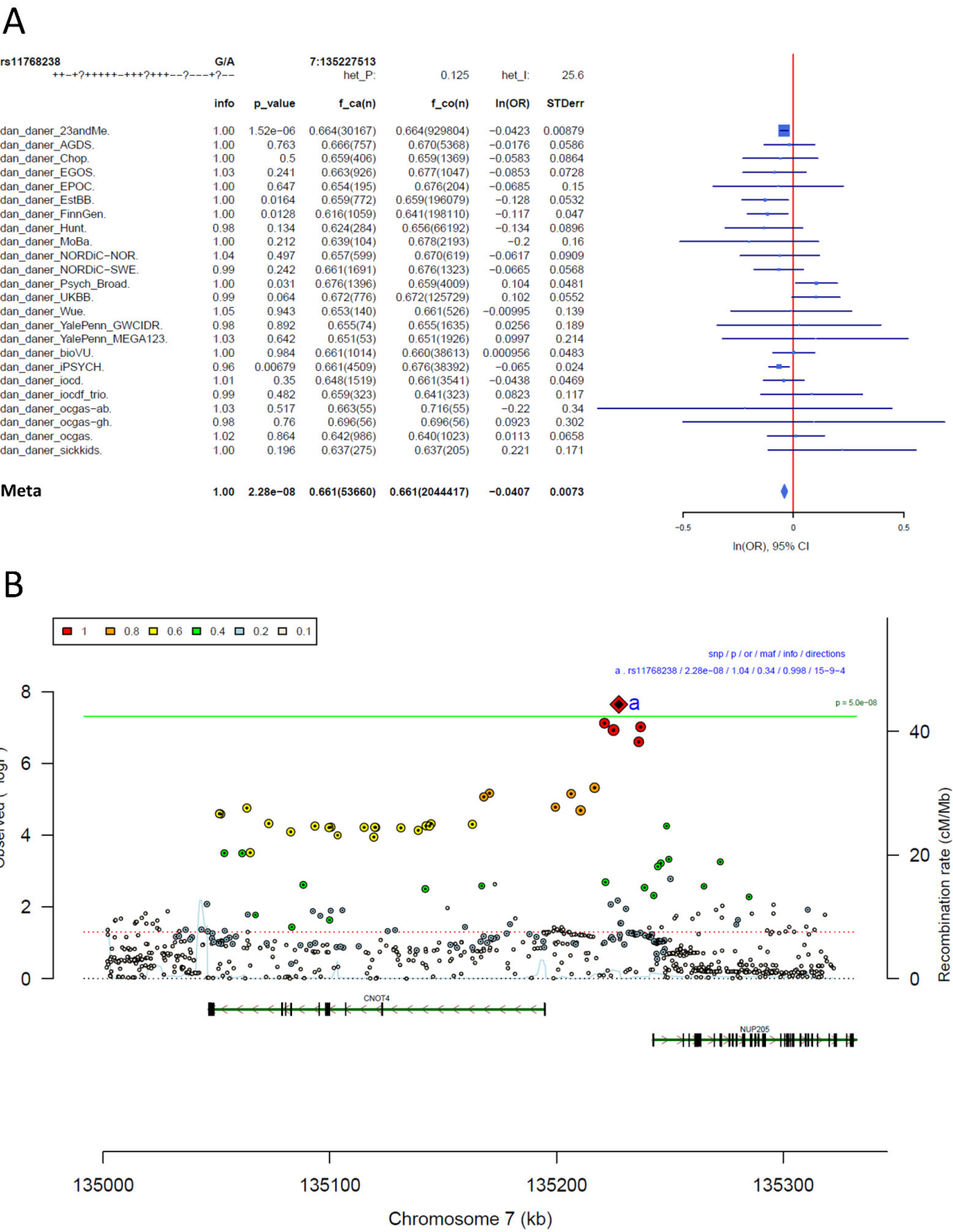
A



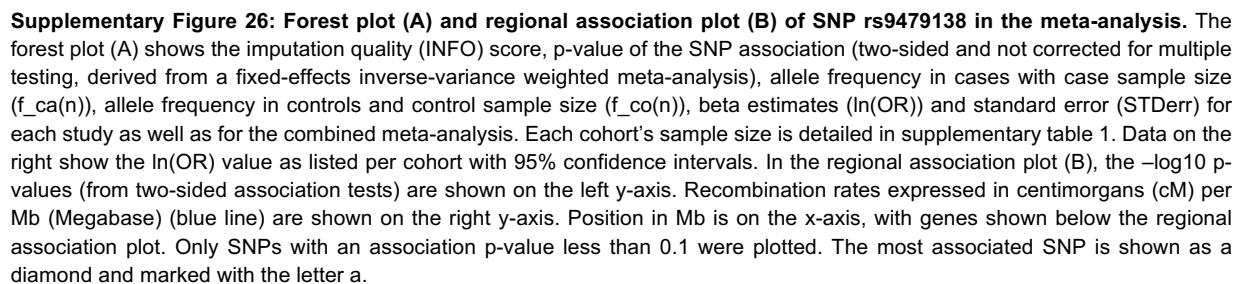
B

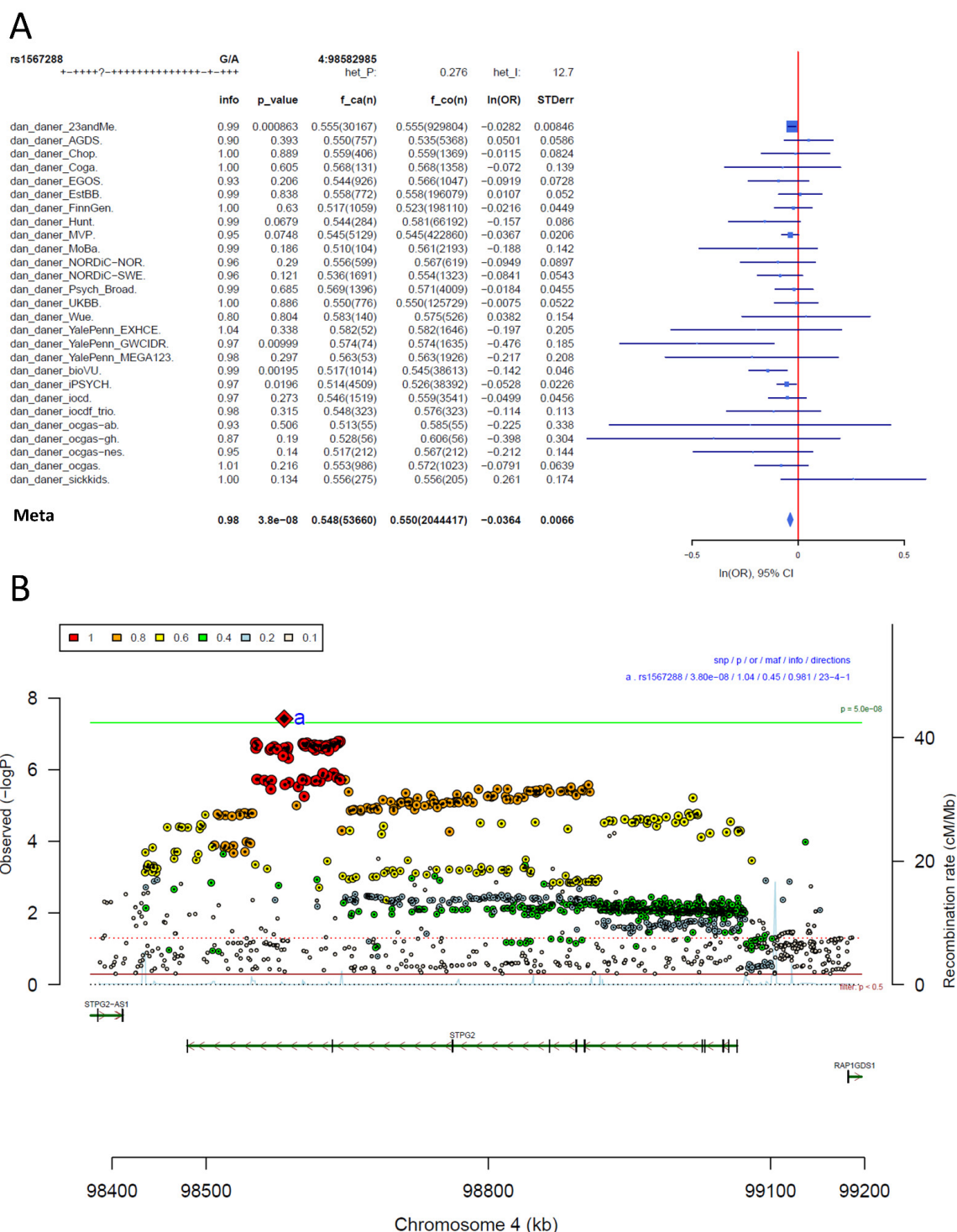


Supplementary Figure 24: Forest plot (A) and regional association plot (B) of SNP rs6474628 in the meta-analysis. The forest plot (A) shows the imputation quality (INFO) score, p-value of the SNP association (two-sided and not corrected for multiple testing, derived from a fixed-effects inverse-variance weighted meta-analysis), allele frequency in cases with case sample size ($f_{ca}(n)$), allele frequency in controls and control sample size ($f_{co}(n)$), beta estimates ($\ln(OR)$) and standard error (STDerr) for each study as well as for the combined meta-analysis. Each cohort's sample size is detailed in supplementary table 1. Data on the right show the $\ln(OR)$ value as listed per cohort with 95% confidence intervals. In the regional association plot (B), the $-\log_{10}$ p-values (from two-sided association tests) are shown on the left y-axis. Recombination rates expressed in centimorgans (cM) per Mb (Megabase) (blue line) are shown on the right y-axis. Position in Mb is on the x-axis, with genes shown below the regional association plot. Only SNPs with an association p-value less than 0.1 were plotted. The most associated SNP is shown as a diamond and marked with the letter a.



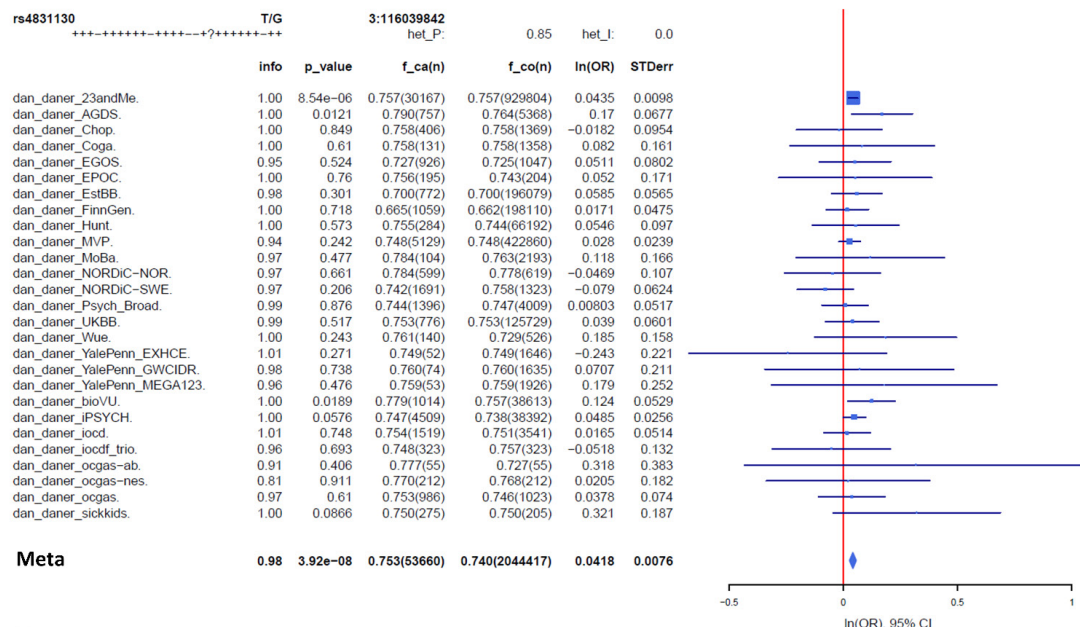
Supplementary Figure 25: Forest plot (A) and regional association plot (B) of SNP rs11768238 in the meta-analysis. The forest plot (A) shows the imputation quality (INFO) score, p-value of the SNP association (two-sided and not corrected for multiple testing, derived from a fixed-effects inverse-variance weighted meta-analysis), allele frequency in cases with case sample size ($f_{ca}(n)$), allele frequency in controls and control sample size ($f_{co}(n)$), beta estimates ($\ln(OR)$) and standard error (STDerr) for each study as well as for the combined meta-analysis. Each cohort's sample size is detailed in supplementary table 1. Data on the right show the $\ln(OR)$ value as listed per cohort with 95% confidence intervals. In the regional association plot (B), the $-\log_{10}$ p-values (from two-sided association tests) are shown on the left y-axis. Recombination rates expressed in centimorgans (cM) per Mb (Megabase) (blue line) are shown on the right y-axis. Position in Mb is on the x-axis, with genes shown below the regional association plot. Only SNPs with an association p-value less than 0.1 were plotted. The most associated SNP is shown as a diamond and marked with the letter a.



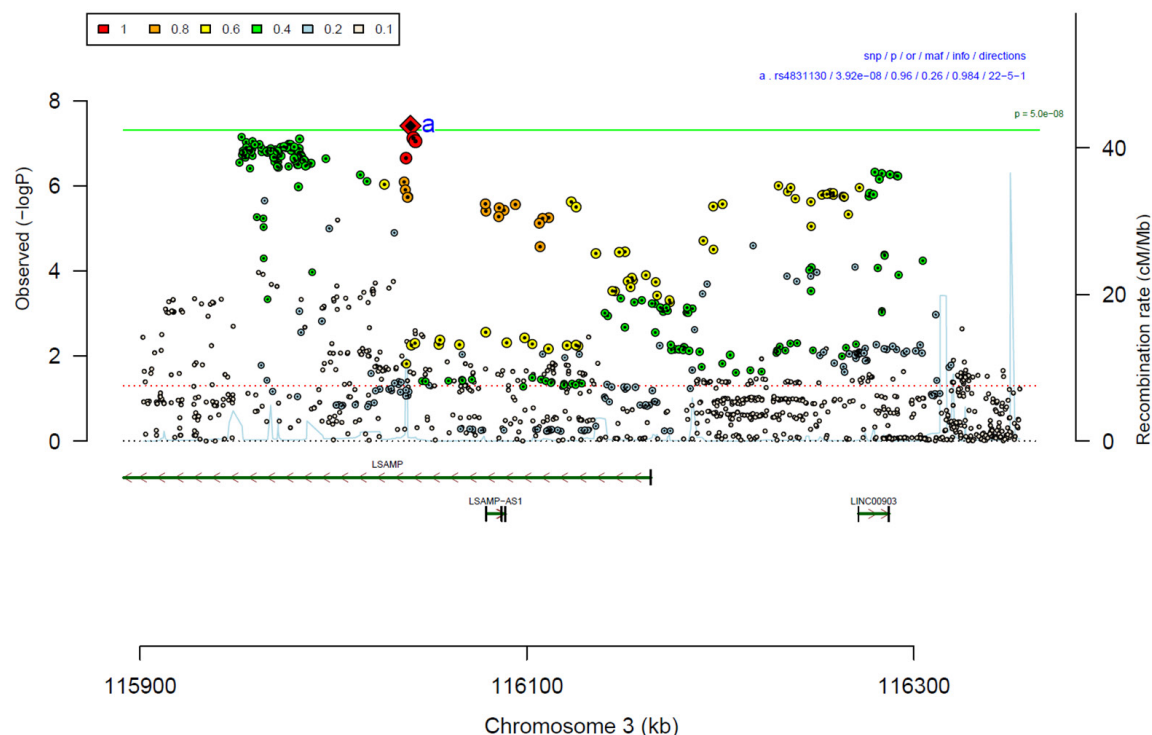


Supplementary Figure 27: Forest plot (A) and regional association plot (B) of SNP rs1567288 in the meta-analysis. The forest plot (A) shows the imputation quality (INFO) score, p-value of the SNP association (two-sided and not corrected for multiple testing, derived from a fixed-effects inverse-variance weighted meta-analysis), allele frequency in cases with case sample size (f_ca(n)), allele frequency in controls and control sample size (f_co(n)), beta estimates (ln(OR)) and standard error (STDerr) for each study as well as for the combined meta-analysis. Each cohort's sample size is detailed in supplementary table 1. Data on the right show the ln(OR) value as listed per cohort with 95% confidence intervals. In the regional association plot (B), the $-\log_{10}$ p-values (from two-sided association tests) are shown on the left y-axis. Recombination rates expressed in centimorgans (cM) per Mb (Megabase) (blue line) are shown on the right y-axis. Position in Mb is on the x-axis, with genes shown below the regional association plot. Only SNPs with an association p-value less than 0.1 were plotted. The most associated SNP is shown as a diamond and marked with the letter a.

A

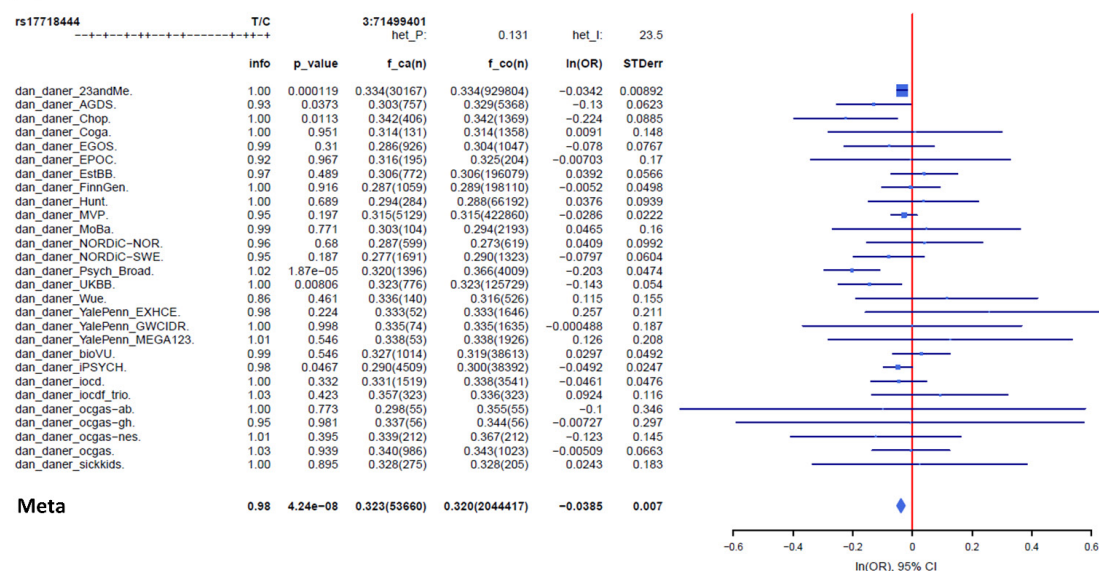


B

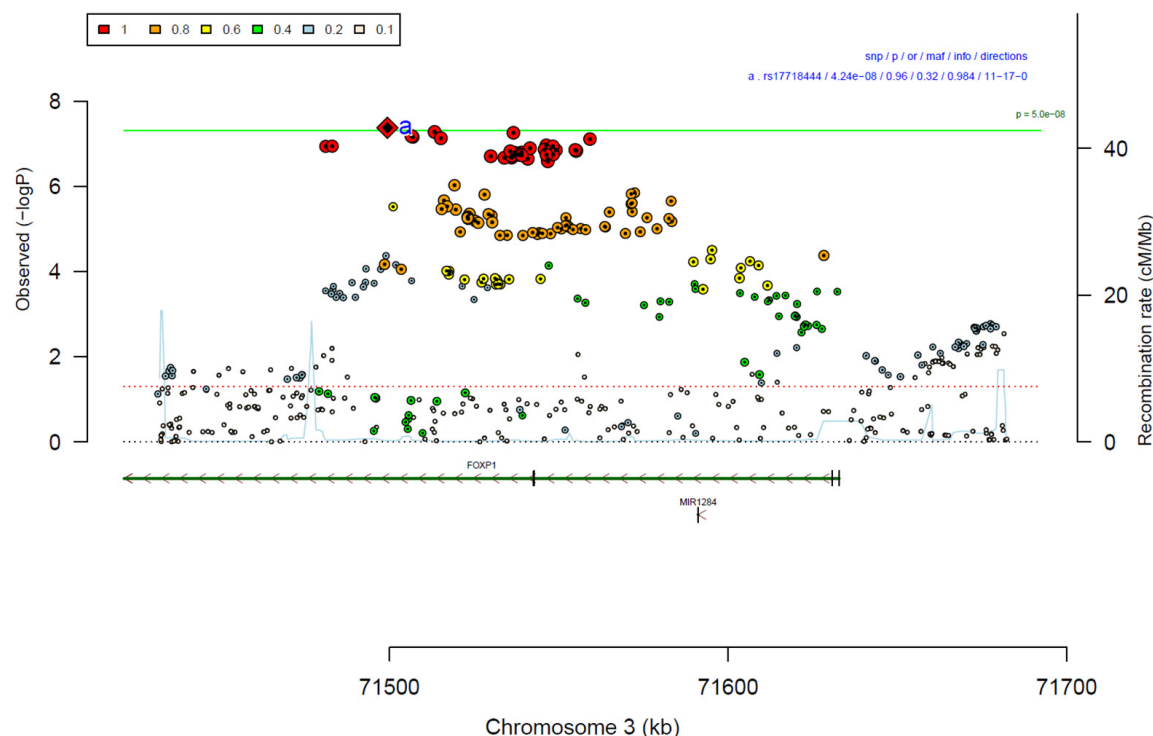


Supplementary Figure 28: Forest plot (A) and regional association plot (B) of SNP rs4831130 in the meta-analysis. The forest plot (A) shows the imputation quality (INFO) score, p-value of the SNP association (two-sided and not corrected for multiple testing, derived from a fixed-effects inverse-variance weighted meta-analysis), allele frequency in cases with case sample size ($f_{ca}(n)$), allele frequency in controls and control sample size ($f_{co}(n)$), beta estimates ($\ln(OR)$) and standard error (STDerr) for each study as well as for the combined meta-analysis. Each cohort's sample size is detailed in supplementary table 1. Data on the right show the $\ln(OR)$ value as listed per cohort with 95% confidence intervals. In the regional association plot (B), the $-\log_{10}$ p-values (from two-sided association tests) are shown on the left y-axis. Recombination rates expressed in centimorgans (cM) per Mb (Megabase) (blue line) are shown on the right y-axis. Position in Mb is on the x-axis, with genes shown below the regional association plot. Only SNPs with an association p-value less than 0.1 were plotted. The most associated SNP is shown as a diamond and marked with the letter a.

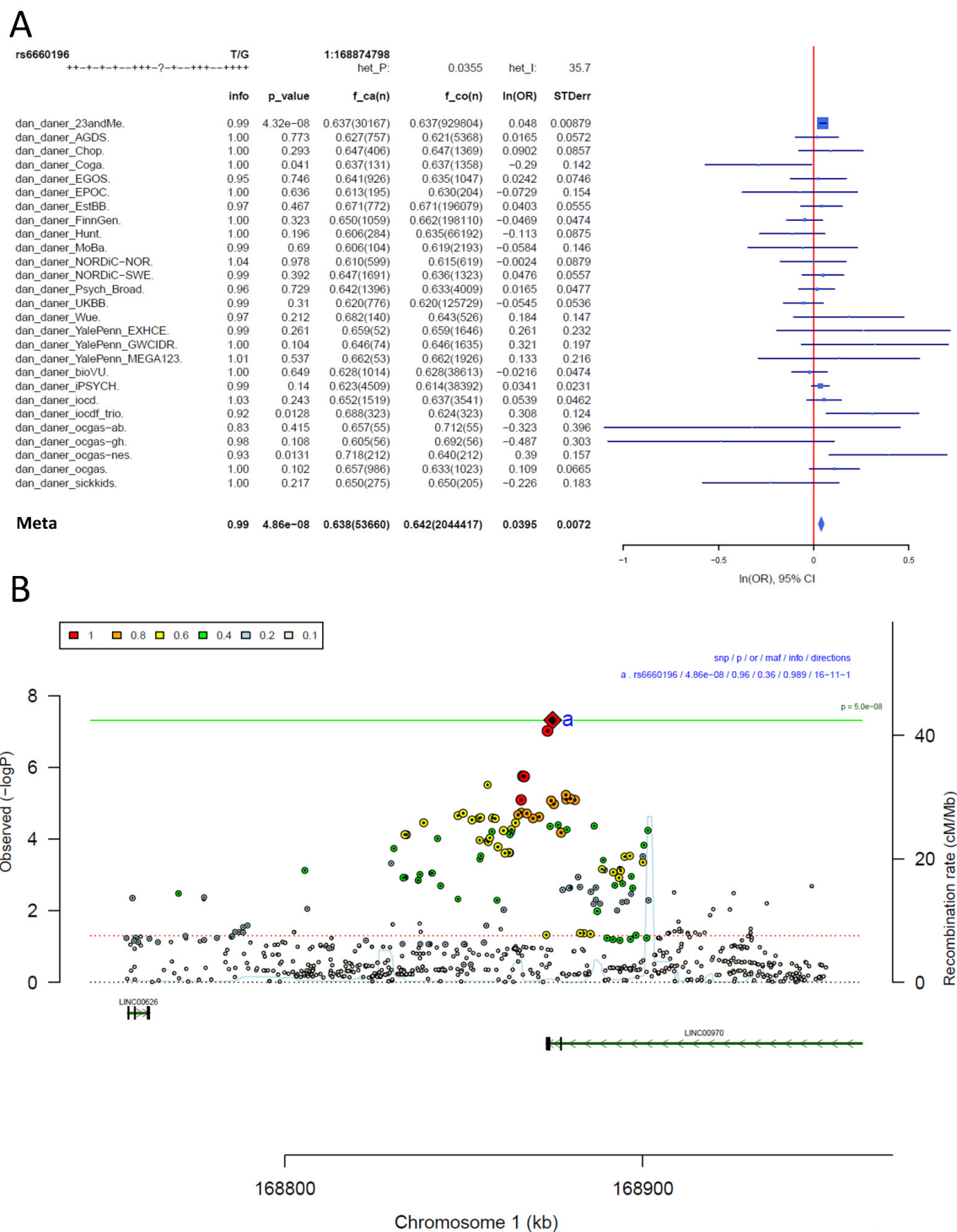
A



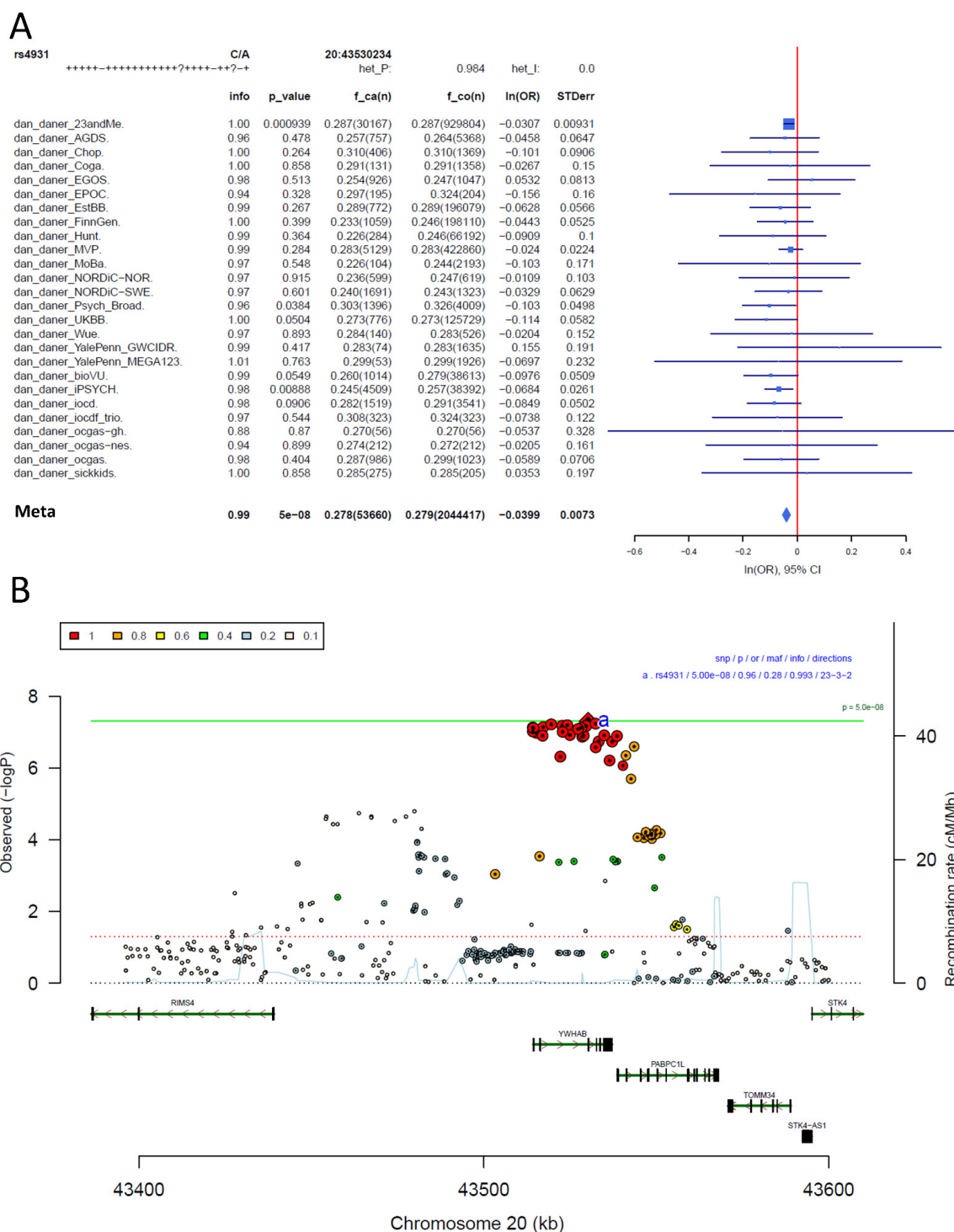
B



Supplementary Figure 29: Forest plot (A) and regional association plot (B) of SNP rs17718444 in the meta-analysis. The forest plot (A) shows the imputation quality (INFO) score, p-value of the SNP association (two-sided and not corrected for multiple testing, derived from a fixed-effects inverse-variance weighted meta-analysis), allele frequency in cases with case sample size (f_ca(n)), allele frequency in controls and control sample size (f_co(n)), beta estimates (ln(OR)) and standard error (STDerr) for each study as well as for the combined meta-analysis. Each cohort's sample size is detailed in supplementary table 1. Data on the right show the ln(OR) value as listed per cohort with 95% confidence intervals. In the regional association plot (B), the $-\log_{10}$ p-values (from two-sided association tests) are shown on the left y-axis. Recombination rates expressed in centimorgans (cM) per Mb (Megabase) (blue line) are shown on the right y-axis. Position in Mb is on the x-axis, with genes shown below the regional association plot. Only SNPs with an association p-value less than 0.1 were plotted. The most associated SNP is shown as a diamond and marked with the letter a.

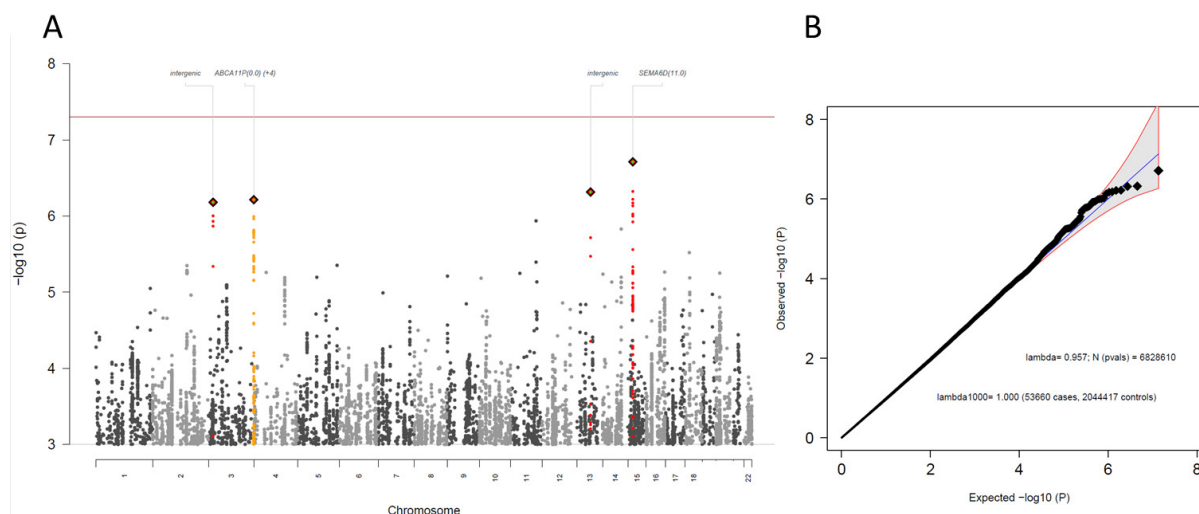


Supplementary Figure 30: Forest plot (A) and regional association plot (B) of SNP rs6660196 in the meta-analysis. The forest plot (A) shows the imputation quality (INFO) score, p-value of the SNP association (two-sided and not corrected for multiple testing, derived from a fixed-effects inverse-variance weighted meta-analysis), allele frequency in cases with case sample size ($f_{ca}(n)$), allele frequency in controls and control sample size ($f_{co}(n)$), beta estimates ($\ln(OR)$) and standard error (STDerr) for each study as well as for the combined meta-analysis. Each cohort's sample size is detailed in supplementary table 1. Data on the right show the $\ln(OR)$ value as listed per cohort with 95% confidence intervals. In the regional association plot (B), the $-\log_{10}$ p-values (from two-sided association tests) are shown on the left y-axis. Recombination rates expressed in centimorgans (cM) per Mb (Megabase) (blue line) are shown on the right y-axis. Position in Mb is on the x-axis, with genes shown below the regional association plot. Only SNPs with an association p-value less than 0.1 were plotted. The most associated SNP is shown as a diamond and marked with the letter a.

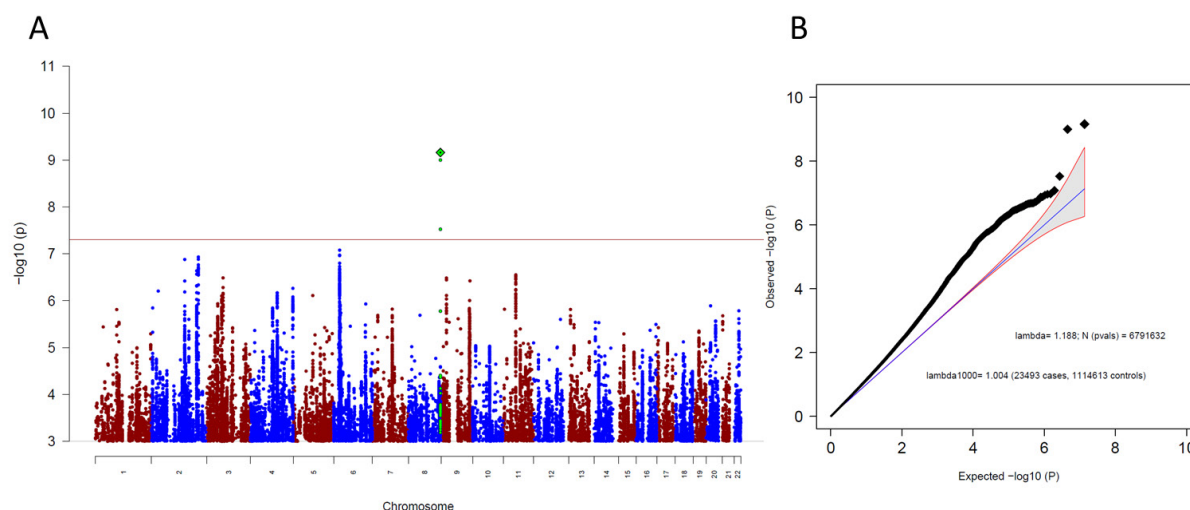


Supplementary Figure 31: Forest plot (A) and regional association plot (B) of SNP rs4931 in the meta-analysis. The forest plot (A) shows the imputation quality (INFO) score, p-value of the SNP association (two-sided and not corrected for multiple testing, derived from a fixed-effects inverse-variance weighted meta-analysis), allele frequency in cases with case sample size ($f_{ca}(n)$), allele frequency in controls and control sample size ($f_{co}(n)$), beta estimates ($\ln(OR)$) and standard error (STDerr) for each study as well as for the combined meta-analysis. Each cohort's sample size is detailed in supplementary table 1. Data on the right show the $\ln(OR)$ value as listed per cohort with 95% confidence intervals. In the regional association plot (B), the $-\log_{10}$ p-values (from two-sided association tests) are shown on the left y-axis. Recombination rates expressed in centimorgans (cM) per Mb (Megabase) (blue line) are shown on the right y-axis. Position in Mb is on the x-axis, with genes shown below the regional association plot. Only SNPs with an association p-value less than 0.1 were plotted. The most associated SNP is shown as a diamond and marked with the letter a.

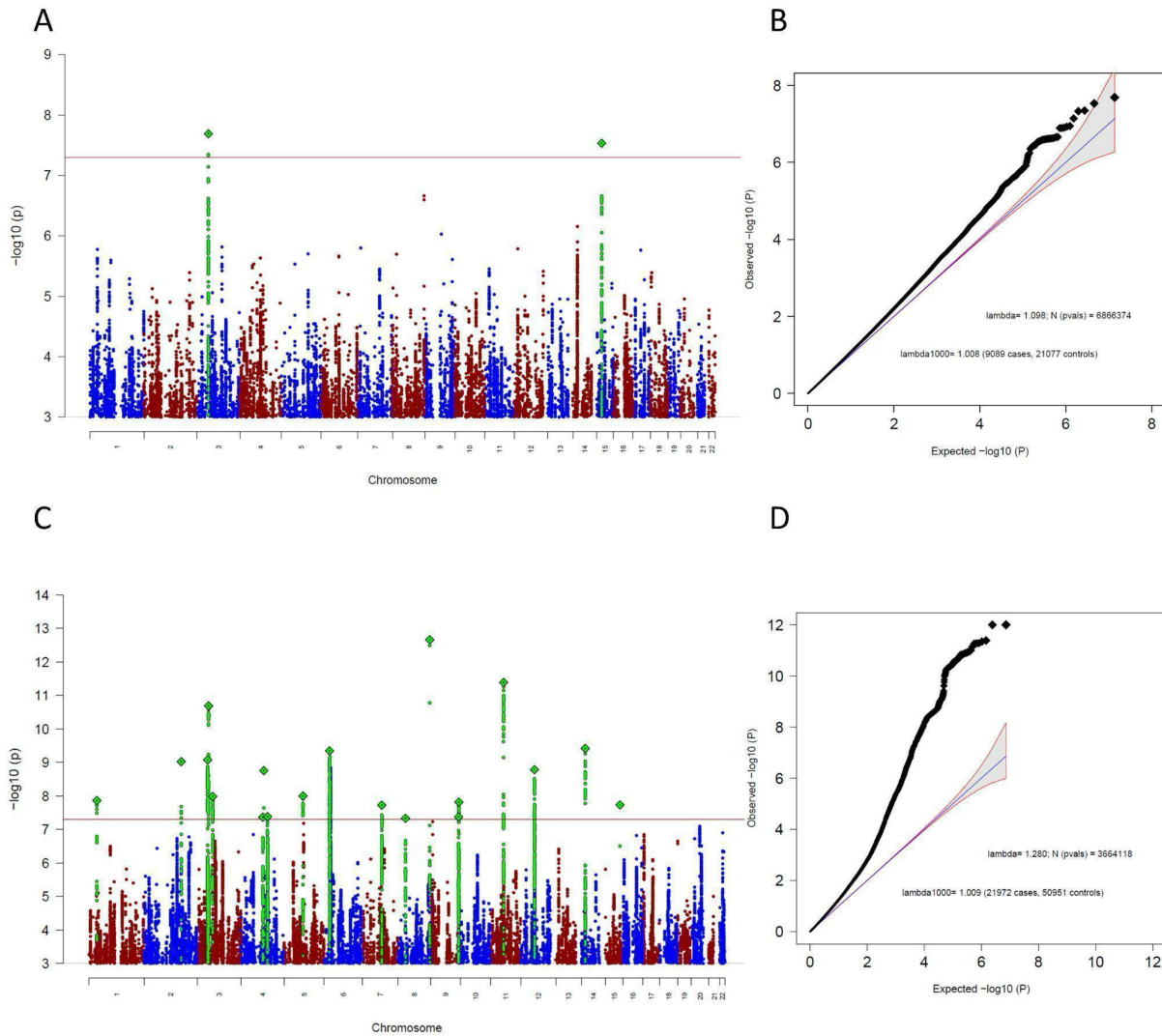
Manhattan-plots and QQ-plots of sub-analyses



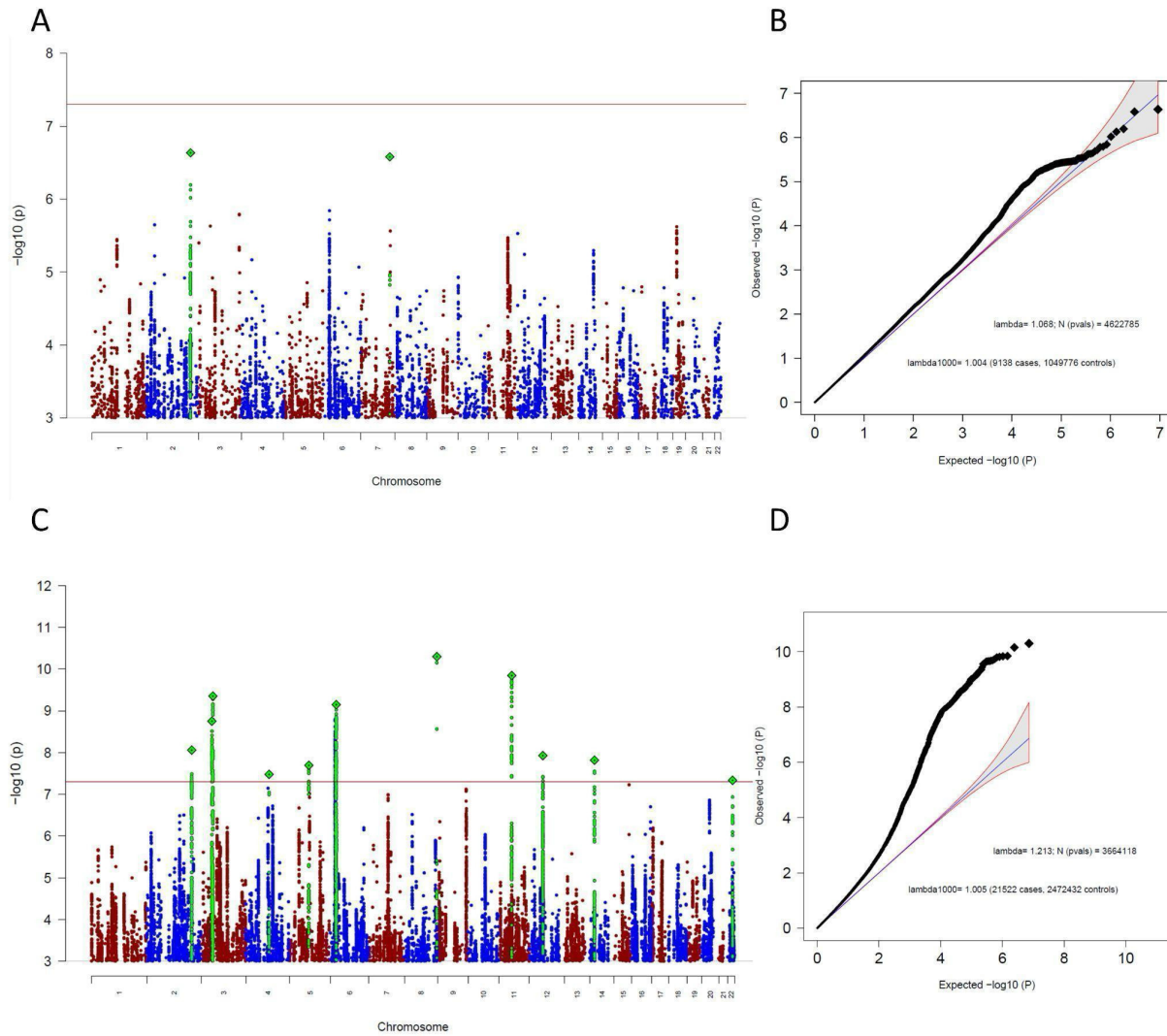
Supplementary Figure 32: Manhattan-plot (A) and QQ-plot (B) of heterogeneity test, indicating whether SNPs across the genome are significantly heterogeneously associated with some cohorts but not others. METAL implements Cochran's Q-test for heterogeneity based on a chi-square distribution; it generates a probability that, when large, indicates larger variation across studies rather than within subjects within a study.



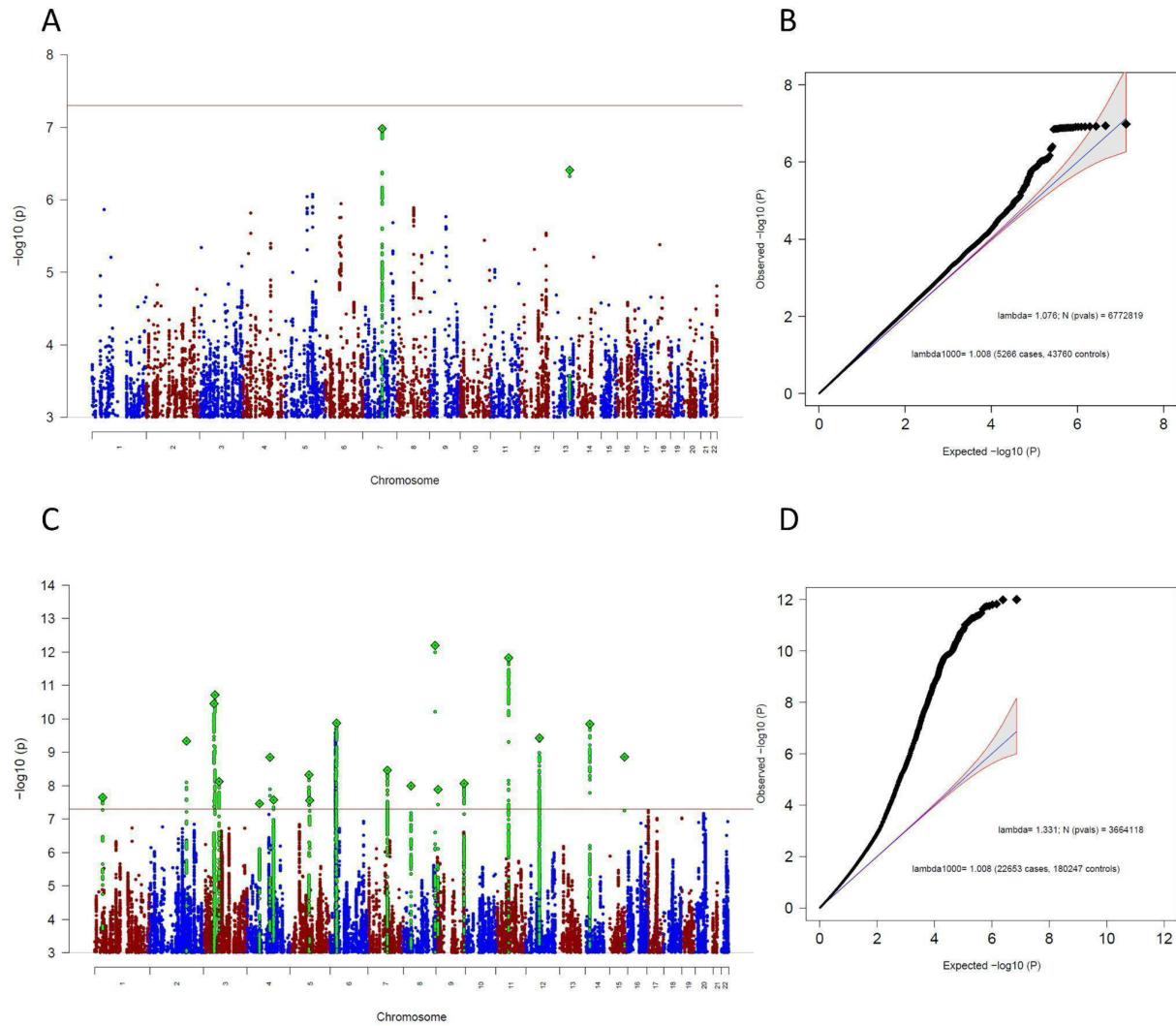
Supplementary Figure 33: Manhattan-plot (A) and QQ-plot (B) of the genome-wide association results for the OCD meta-analysis leaving out 23andMe. (A) The x-axis shows the position in the genome (chromosome 1 to 22), the y-axis represents $-\log_{10} p$ -values for the association of variants with OCD from metaanalysis using an inverse-variance weighted fixed effects model. The horizontal red line shows the threshold for genome-wide significance (5×10^{-8}). Each dot represents one SNP that was tested in the GWAS, with a green diamond indicating the lead SNP of a genome-wide significant locus, with green dots below belonging to that locus. (B) the expected $-\log_{10}(p)$ under the null is plotted against the observed $-\log_{10}(p)$. The shading indicates the 95% confidence region under the null. Lambda and Lambda1000 (which is the Lambda if the GWAS contained 1000 cases and 1000 controls) indicate genomic inflation factors. Number of SNPs (N (pvals)) and Number of cases and controls are given in parentheses.



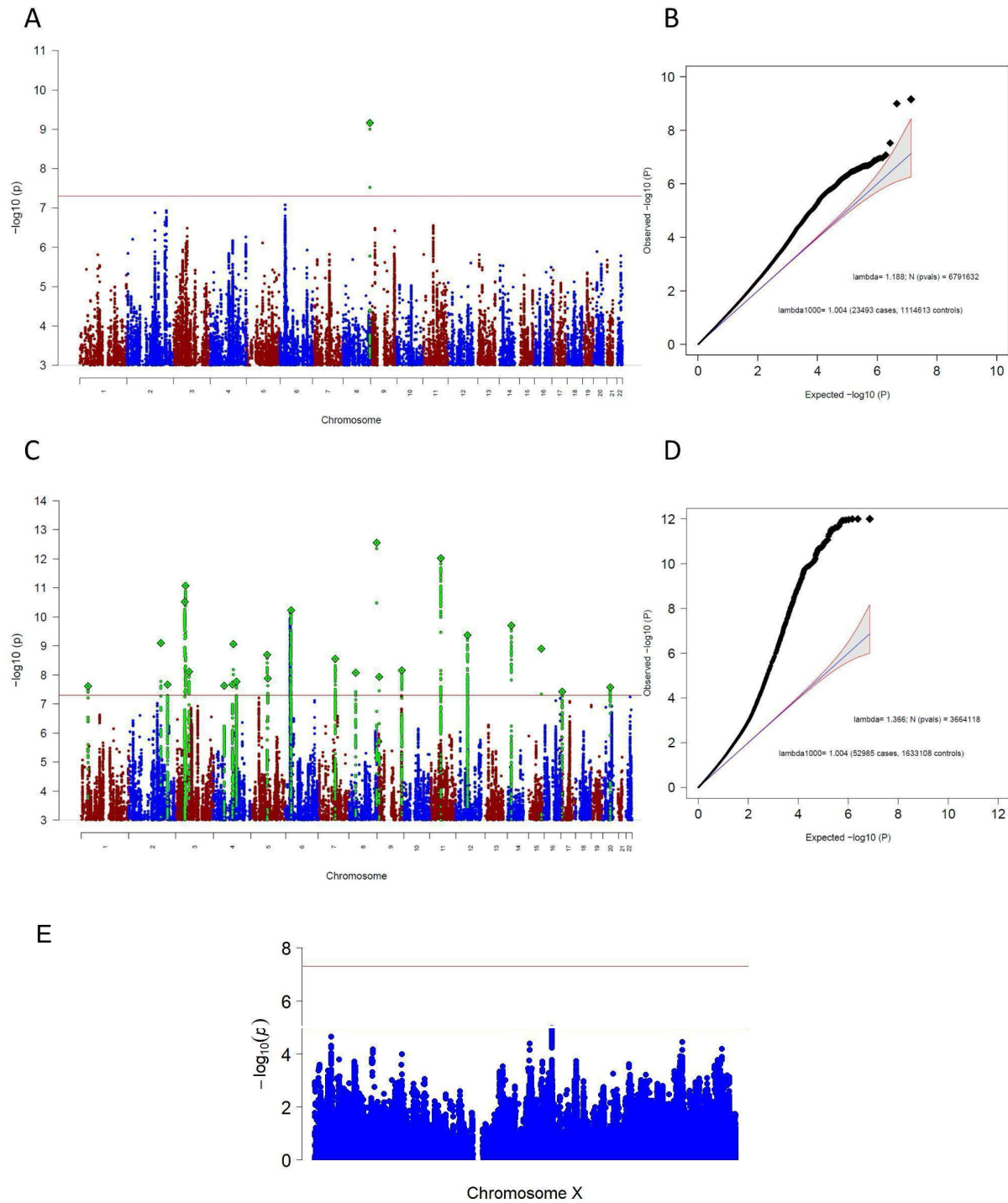
Supplementary Figure 34: Manhattan-plot (A) and QQ-plot (B) of the sub-group specific GWAS analysis, including only clinical cohorts and Manhattan-plot (C) and QQ-plot (D) of the MTAG analysis specific to clinical cohorts. Both analyses aimed to find ascertainment-specific signals associated with OCD specifically in clinical cohorts. (A,C) The x-axis shows the position in the genome (chromosome 1 to 22), the y-axis represents $-\log_{10}$ p-values for the association of variants with OCD from metaanalysis using an inverse-variance weighted fixed effects model (A) or MTAG (C). The horizontal red line shows the threshold for genome-wide significance (5×10^{-8}). Each dot represents one SNP that was tested in the GWAS, with a green diamond indicating the lead SNP of a genome-wide significant locus, with green dots below belonging to that locus. (B) the expected $-\log_{10}(p)$ under the null is plotted against the observed $-\log_{10}(p)$. The shading indicates the 95% confidence region under the null. Lambda and Lambda1000 (which is the Lambda if the GWAS contained 1000 cases and 1000 controls) indicate genomic inflation factors. Number of SNPs (N (pvals)) and Number of cases and controls are given in parentheses.



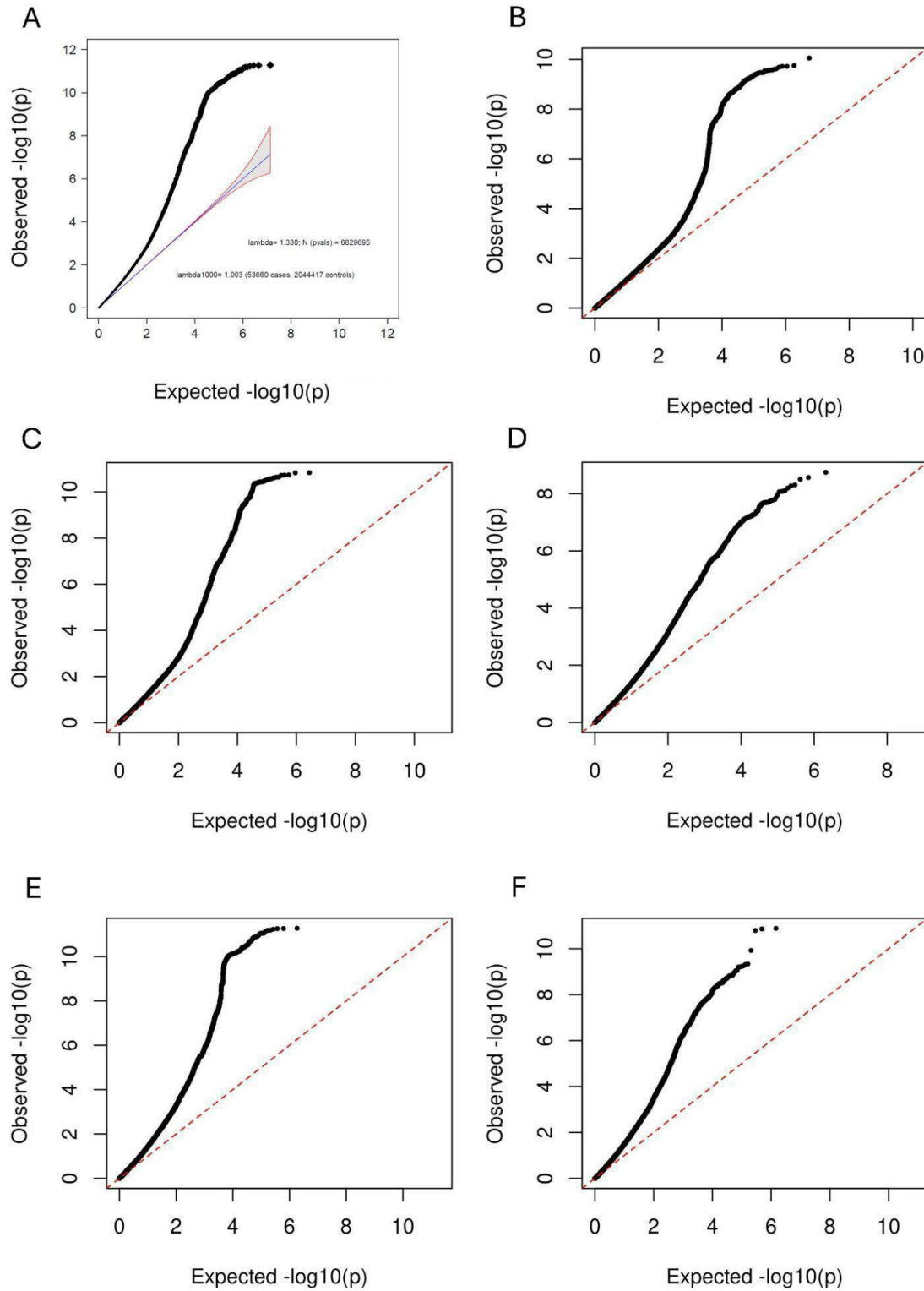
Supplementary Figure 35: Manhattan-plot (A) and QQ-plot (B) of the sub-group specific GWAS analysis, including only biobank cohorts and Manhattan-plot (C) and QQ-plot (D) of the MTAG analysis specific to biobank cohorts. Both analyses aimed to find ascertainment-specific signals associated with OCD specifically in biobank cohorts. (A,C) The x-axis shows the position in the genome (chromosome 1 to 22), the y-axis represents $-\log_{10}$ p-values for the association of variants with OCD from metaanalysis using an inverse-variance weighted fixed effects model (A) or MTAG (C). The horizontal red line shows the threshold for genome-wide significance (5×10^{-8}). Each dot represents one SNP that was tested in the GWAS, with a green diamond indicating the lead SNP of a genome-wide significant locus, with green dots below belonging to that locus. (B) the expected $-\log_{10}(p)$ under the null is plotted against the observed $-\log_{10}(p)$. The shading indicates the 95% confidence region under the null. Lambda and Lambda1000 (which is the Lambda if the GWAS contained 1000 cases and 1000 controls) indicate genomic inflation factors. Number of SNPs (N (pvals)) and Number of cases and controls are given in parentheses.



Supplementary Figure 36: Manhattan-plot (A) and QQ-plot (B) of the sub-group specific GWAS analysis, including only comorbid cohorts and Manhattan-plot (C) and QQ-plot (D) of the MTAG analysis specific to comorbid cohorts. Both analyses aimed to find ascertainment-specific signals associated with OCD specifically in comorbid cohorts. (A,C) The x-axis shows the position in the genome (chromosome 1 to 22), the y-axis represents $-\log_{10}$ p-values for the association of variants with OCD from metaanalysis using an inverse-variance weighted fixed effects model (A) or MTAG (C). The horizontal red line shows the threshold for genome-wide significance (5×10^{-8}). Each dot represents one SNP that was tested in the GWAS, with a green diamond indicating the lead SNP of a genome-wide significant locus, with green dots below belonging to that locus. (B) the expected $-\log_{10}(p)$ under the null is plotted against the observed $-\log_{10}(p)$. The shading indicates the 95% confidence region under the null. Lambda and Lambda1000 (which is the Lambda if the GWAS contained 1000 cases and 1000 controls) indicate genomic inflation factors. Number of SNPs (N (pvals)) and Number of cases and controls are given in parentheses.

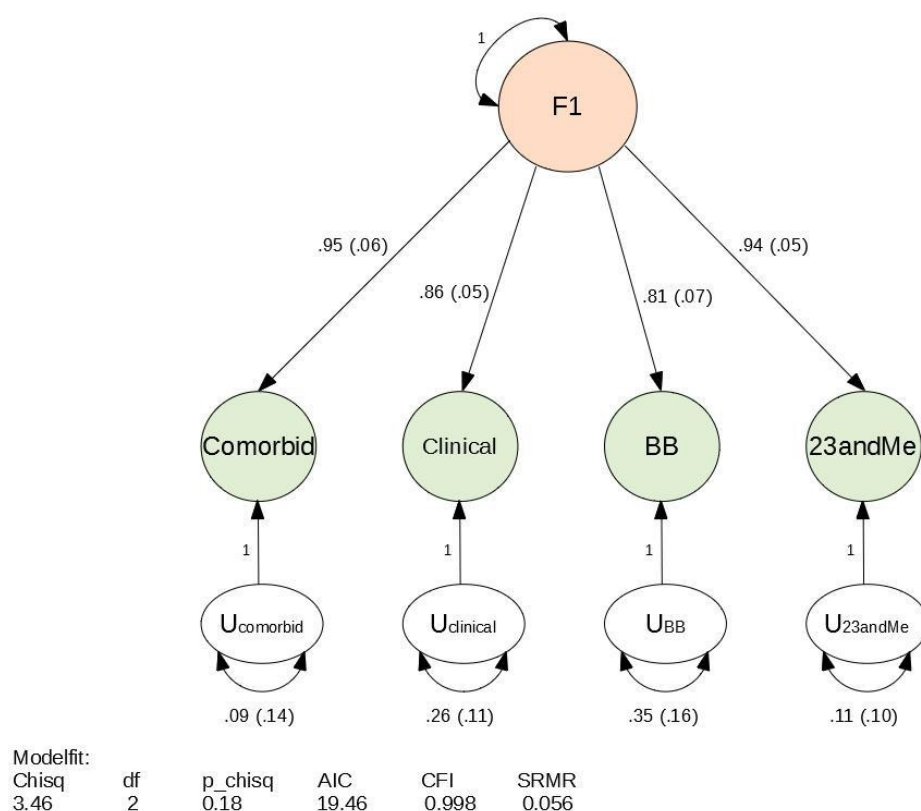


Supplementary Figure 37: Manhattan-plot (A) and QQ-plot (B) of the sub-group specific GWAS analysis, including only 23andMe, Manhattan-plot (C) and QQ-plot (D) of the MTAG analysis specific to 23andMe, and (E) a plot of the X-chromosome findings specific to 23andMe. Both analyses (A-D) aimed to find ascertainment-specific signals associated with OCD specifically in 23andMe. (A,C) The x-axis shows the position in the genome (chromosome 1 to 22), the y-axis represents $-\log_{10}$ p-values for the association of variants with OCD from metaanalysis using an inverse-variance weighted fixed effects model (A) or MTAG (C). The horizontal red line shows the threshold for genome-wide significance (5×10^{-8}). Each dot represents one SNP that was tested in the GWAS, with a green diamond indicating the lead SNP of a genome-wide significant locus, with green dots below belonging to that locus. (B) the expected $-\log_{10}(p)$ under the null is plotted against the observed $-\log_{10}(p)$. The shading indicates the 95% confidence region under the null. Lambda and Lambda1000 (which is the Lambda if the GWAS contained 1000 cases and 1000 controls) indicate genomic inflation factors. Number of SNPs (N (pvals)) and Number of cases and controls are given in parentheses.

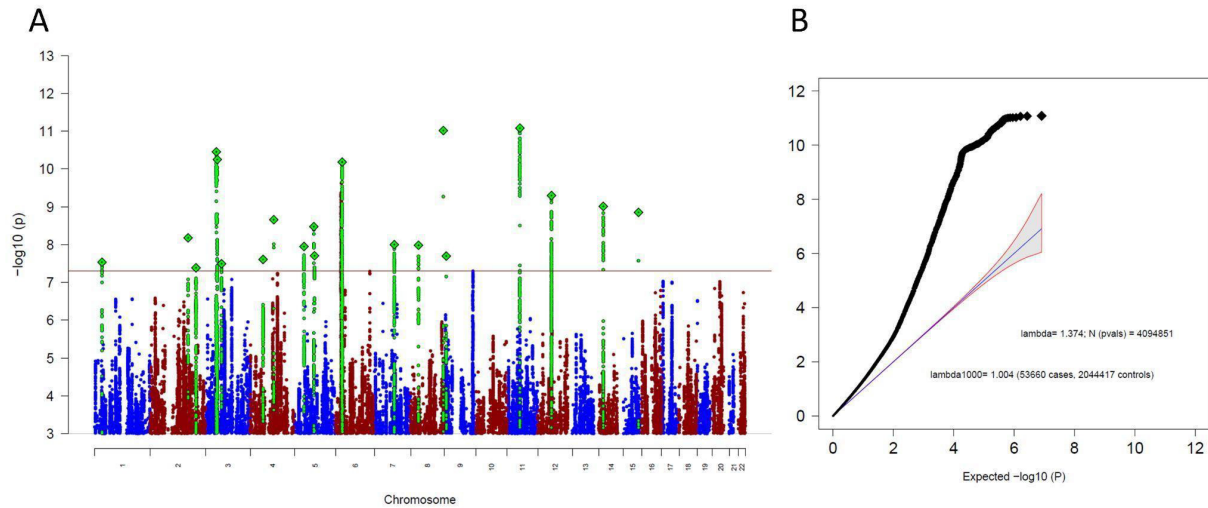


Supplementary Figure 38: QQ-plot for main GWAS results. (A) across the entire allele-frequency spectrum; (B)-(F) partitioned by MAF bins (B) 0.01-0.1, (C) 0.1-0.2, (D) 0.2-0.3, (E) 0.3-0.4, (F) 0.4-0.5. For each bin, SNPs were included if their MAF in the overall analysis fell within that range. The expected $-\log_{10}(p)$ values under the null hypothesis are plotted against the observed $-\log_{10}(p)$ values.

GenomicSEM one factor model of the four OCD ascertainment sub-groups

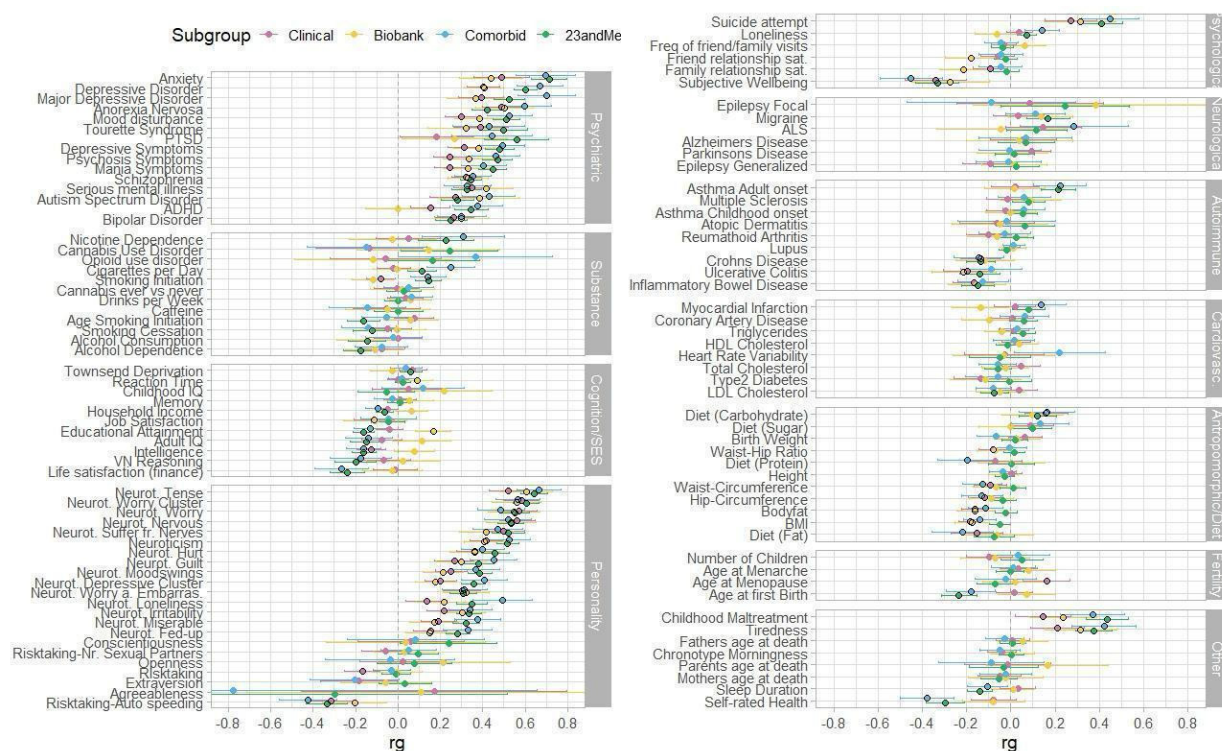


Supplementary Figure 39: A path diagram of the common-factor GenomicSEM model without SNP effects, specified with unit variance identification, fixing the variance of the common factor F1 to 1. All estimates are standardized. Model-fit indices below are Chi-square statistic (Chisq); degrees of freedom of the model (df); p-value of the Chi-square (p_chisq); Akaike Information Criterion (AIC), which is a comparative measure of fit with lower values indicating a better fit; comparative fit index (CFI) which assesses the relative improvement in fit compared with the baseline model, ranging between 0 and 1; and standardized root mean square residual (SRMR), which is an absolute measure of fit defined as the standardized difference between observed correlation and the predicted correlation with a value of 0 indicating perfect fit.

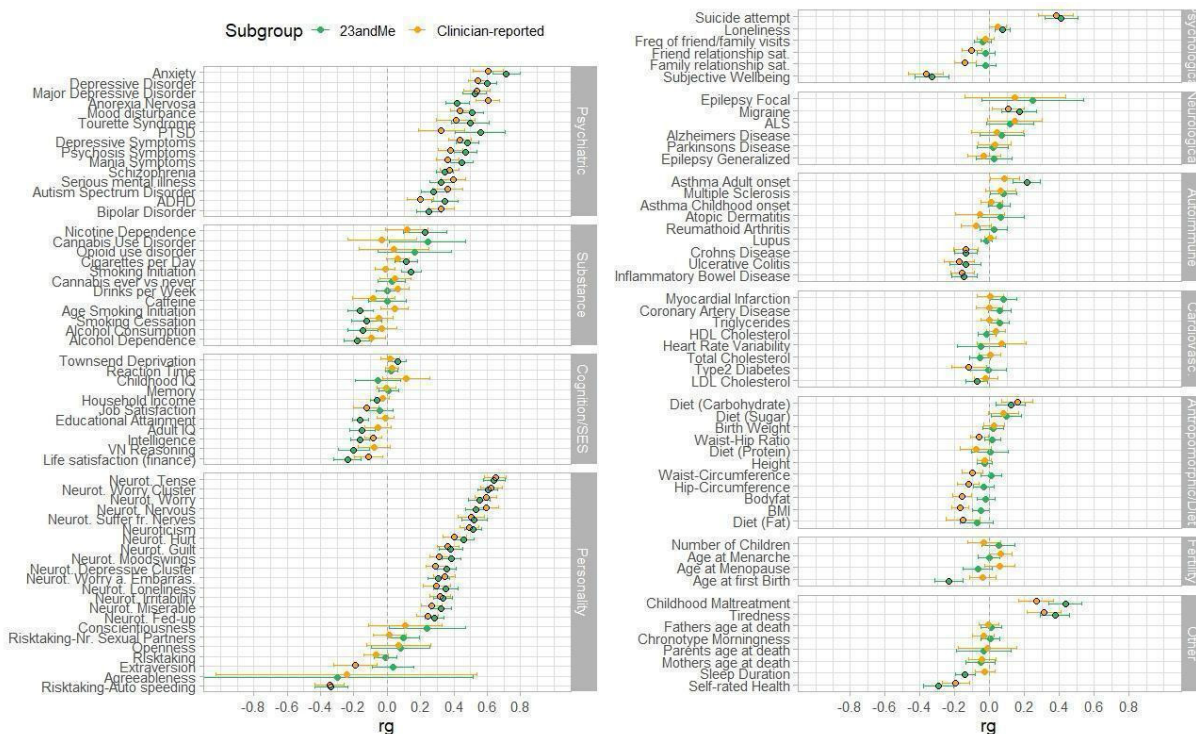


Supplementary Figure 40: Manhattan-plot (A) and QQ-plot (B) of the common-factor GenomicSEM GWAS as shown in Supplementary Figure 39. (A) The x-axis shows the position in the genome (chromosome 1 to 22), the y-axis represents $-\log_{10}$ p-values for the association of variants with OCD (A). The horizontal red line shows the threshold for genome-wide significance (5×10^{-8}). Each dot represents one SNP that was tested in the GWAS, with a green diamond indicating the lead SNP of a genome-wide significant locus, with green dots below belonging to that locus. (B) the expected $-\log_{10}(p)$ under the null is plotted against the observed $-\log_{10}(p)$. The shading indicates the 95% confidence region under the null. Lambda and Lambda1000 (which is the Lambda if the GWAS contained 1000 cases and 1000 controls) indicate genomic inflation factors. Number of SNPs (N (pvals)) and Number of cases and controls are given in parentheses.

Cross-trait genetic correlations of the ascertainment-specific subgroup analyses

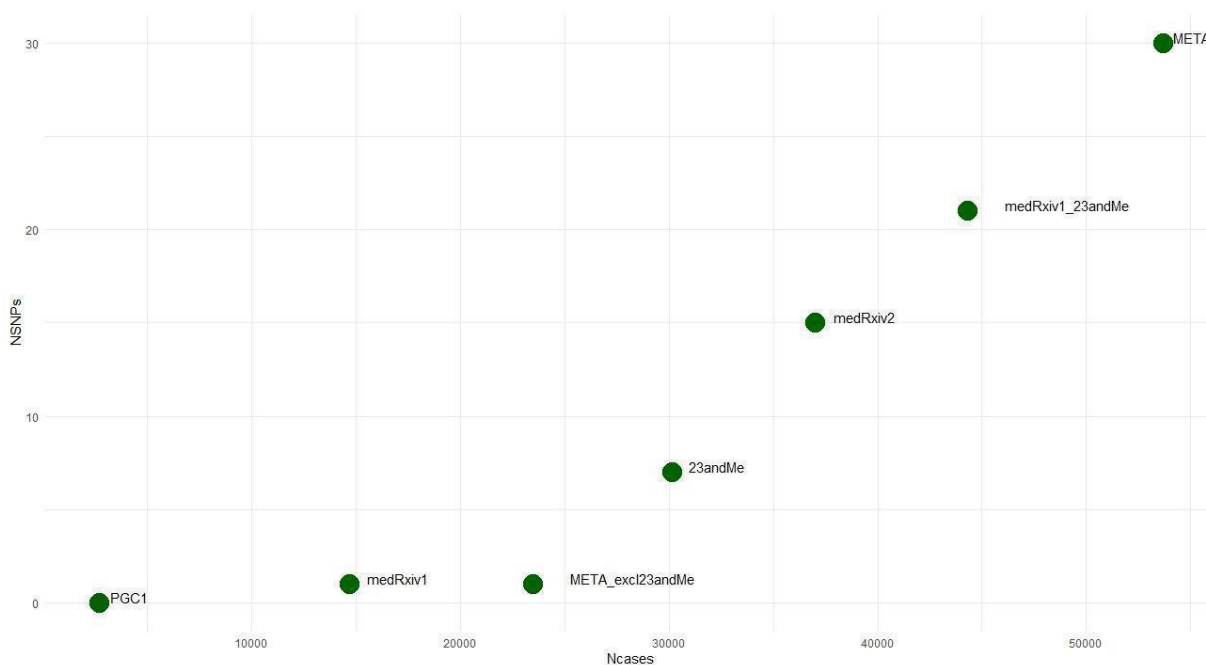


Supplementary Figure 41: Genetic correlations (rg) between the ascertainment-specific OCD GWASs (23andMe, Biobank, Clinical, comorbid) and 112 psychiatric, substance use, cognition/socioeconomic status (SES), personality, psychological, neurological, autoimmune, cardiovascular, anthropomorphic/diet, fertility, and other phenotypes. References and sample sizes of the corresponding summary statistics of the GWAS studies can be found in **Supplementary Table 19. The OCD summary statistics are of the main meta-analysis (Ncases = 53,660 and Ncontrols = 2,044,417). Genetic correlations of OCD specific to 23andMe are in green, specific to biobanks are in yellow, specific to clinical cohorts are in pink, and specific to comorbid cohorts are in blue. Error bars represent 95% confidence intervals for the genetic correlation estimates (rg), black encircled estimates indicate significant associations with a p-value adjusted for multiple testing with the Benjamini-Hochberg procedure to control the FDR (< 0.05), adjusted separately for each OCD ascertainment-specific subgroup (see full list of results in Supplementary Table 18).**



Supplementary Figure 42: Genetic correlations (rg) of the OCD GWASs including only 23andMe and the OCD GWAS leaving out 23andMe (Clinician-reported) with 112 psychiatric, substance use, cognition/socioeconomic status (SES), personality, psychological, neurological, autoimmune, cardiovascular, anthropomorphic/diet, fertility, and other phenotypes. References and sample sizes of the corresponding summary statistics of the GWAS studies can be found in **Supplementary Table 19**. The OCD summary statistics are of the main meta-analysis (Ncases = 53,660 and Ncontrols = 2,044,417). Genetic correlations of OCD specific to 23andMe are in green, specific to the GWAS excluding 23andMe (Clinician-reported) in orange. Error bars represent 95% confidence intervals for the genetic correlation estimates (rg), black encircled estimates indicate significant associations with a p-value adjusted for multiple testing with the Benjamini-Hochberg procedure to control the FDR (< 0.05), adjusted separately for each OCD subgroup.

Consistency of SNP findings across OCD GWAS versions



Supplementary Figure 43: Number of GWAS cases (x-axis) and the respective number of independent significant SNPs they identified (y-axis). Shown are seven GWAS studies - PGC1 (IOCDF-GC and OCGAS, 2018) ; PGC OCD2 GWAS (medRxiv1; Strom et al. 2021); the meta-analysis of the current paper (META) excluding 23andMe (META_excl23andMe); 23andMe, iPSYCH-PGC-23andMe GWAS (medRxiv2; Strom et al. 2024), PGC OCD2 GWAS and 23andMe (medRxiv1_23andMe), and the current meta-analysis (META). Please also see **Supplementary Table 20** for details. Note that these iterations of OCD GWAS meta-analyses are not independent of each other. Cohorts with fewer cases (to the left of the plot) are included in meta-analyses with more cases (to the right of the plot). An overview of the significant SNPs from the published medRxiv GWASs (medRxiv1 and medRxiv2) and their corresponding p-values in the current OCD GWAS can be found in **Supplementary Table 5**.

References

- Abraham, G., & Inouye, M. (2014). Fast Principal Component Analysis of Large-Scale Genome-Wide Data. *PLOS ONE*, 9(4), e93766. <https://doi.org/10.1371/journal.pone.0093766>
- Altshuler, D. M., Gibbs, R. A., Peltonen, L., Schaffner, S. F., Yu, F., Dermitzakis, E., Bonnen, P. E., De Bakker, P. I. W., Deloukas, P., Gabriel, S. B., Gwilliam, R., Hunt, S., Inouye, M., Jia, X., Aarno Palotie, Parkin, M., Whittaker, P., Chang, K., Hawes, A., ... McEwen, J. E. (2010). Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311), 52–58. <https://doi.org/10.1038/nature09298>
- Arnold, P. D., Askland, K. D., Barlassina, C., Bellodi, L., Bienvenu, O. J., Black, D., Bloch, M., Brentani, H., Burton, C. L., Camarena, B., Cappi, C., Cath, D., Cavallini, M., Conti, D., Cook, E., Coric, V., Cullen, B. A., Cusi, D., Davis, L. K., ... Zai, G. (2018). Revealing the complex genetic architecture of obsessive-compulsive disorder using meta-analysis. *Molecular Psychiatry*, 23(5). <https://doi.org/10.1038/mp.2017.154>
- Bosch, R., Pagerols, M., Rivas, C., Sixto, L., Bricollé, L., Español-Martín, G., Prat, R., Ramos-Quiroga, J. A., & Casas, M. (2021). Neurodevelopmental disorders among Spanish school-age children: Prevalence and sociodemographic correlates. *Psychological Medicine*, 1–11. <https://doi.org/10.1017/S0033291720005115>
- Browning, B. L. (2017). Beagle 4.1. https://faculty.washington.edu/Browning/Beagle/Beagle_4.1_21Jan17.Pdf. <https://doi.org/10.1086/521987>
- Browning, B. L., Zhou, Y., & Browning, S. R. (2018). A One-Penny Imputed Genome from Next-Generation Reference Panels. *American Journal of Human Genetics*, 103(3), 338–348. <https://doi.org/10.1016/j.ajhg.2018.07.015>
- Browning, S. R., & Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American Journal of Human Genetics*, 81(5). <https://doi.org/10.1086/521987>

- Burton, C. L., Lemire, M., Xiao, B., Corfield, E. C., Erdman, L., Bralten, J., Poelmans, G., Yu, D., Shaheen, S. M., Goodale, T., Sinopoli, V. M., Askland, K. D., Barlassina, C., Bienvenu, O. J., Black, D., Bloch, M., Brentani, H., Camarena, B., Cappi, C., ... Arnold, P. D. (2021). Genome-wide association study of pediatric obsessive-compulsive traits: Shared genetic risk between traits and disorder. *Translational Psychiatry* 2021 11:1, 11(1), 1–10. <https://doi.org/10.1038/s41398-020-01121-9>
- Bybjerg-Grauholm, J., Pedersen, C. B., Bækvad-Hansen, M., Pedersen, M. G., Adamsen, D., Hansen, C. S., Agerbo, E., Grove, J., Als, T. D., Schork, A. J., Buil, A., Mors, O., Nordentoft, M., Werge, T., Børglum, A. D., Hougaard, D. M., & Mortensen, P. B. (2020). *The iPSYCH2015 Case-Cohort sample: Updated directions for unravelling genetic and environmental architectures of severe mental disorders* (p. 2020.11.30.20237768). medRxiv. <https://doi.org/10.1101/2020.11.30.20237768>
- Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., Cortes, A., Welsh, S., Young, A., Effingham, M., McVean, G., Leslie, S., Allen, N., Donnelly, P., & Marchini, J. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature* 2018 562:7726, 562(7726), 203–209. <https://doi.org/10.1038/s41586-018-0579-z>
- Calkins, M. E., Merikangas, K. R., Moore, T. M., Burstein, M., Behr, M. A., Satterthwaite, T. D., Ruparel, K., Wolf, D. H., Roalf, D. R., Mentch, F. D., Qiu, H., Chiavacci, R., Connolly, J. J., Sleiman, P. M. A., Gur, R. C., Hakonarson, H., & Gur, R. E. (2015). The Philadelphia Neurodevelopmental Cohort: Constructing a deep phenotyping collaborative. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, 56(12), 1356–1369. <https://doi.org/10.1111/jcpp.12416>
- Calkins, M. E., Moore, T. M., Merikangas, K. R., Burstein, M., Satterthwaite, T. D., Bilker, W. B., Ruparel, K., Chiavacci, R., Wolf, D. H., Mentch, F., Qiu, H., Connolly, J. J., Sleiman, P. A., Hakonarson, H., Gur, R. C., & Gur, R. E. (2014). The psychosis spectrum in a young U.S. community sample: Findings from the Philadelphia Neurodevelopmental Cohort. *World*

<https://doi.org/10.1002/wps.20152>

- Chang, C. C., Chow, C. C., Tellier, L. C. A. M., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience*, 4(1), 7. <https://doi.org/10.1186/S13742-015-0047-8/2707533>
- Cheng, Z., Zhou, H., Sherva, R., Farrer, L. A., Kranzler, H. R., & Gelernter, J. (2018). Genome-wide Association Study Identifies a Regulatory Variant of RGMA Associated With Opioid Dependence in European Americans. *Biological Psychiatry*, 84(10), 762–770. <https://doi.org/10.1016/j.biopsych.2017.12.016>
- Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A. E., Kwong, A., Vrieze, S. I., Chew, E. Y., Levy, S., McGue, M., Schlessinger, D., Stambolian, D., Loh, P. R., Iacono, W. G., Swaroop, A., Scott, L. J., Cucca, F., Kronenberg, F., Boehnke, M., ... Fuchsberger, C. (2016). Next-generation genotype imputation service and methods. *Nature Genetics*, 48(10), 1284–1287. <https://doi.org/10.1038/NG.3656>
- de Klein, N., Tsai, E. A., Vochteloo, M., Baird, D., Huang, Y., Chen, C.-Y., van Dam, S., Oelen, R., Deelen, P., Bakker, O. B., El Garwany, O., Ouyang, Z., Marshall, E. E., Zavodszky, M. I., van Rheenen, W., Bakker, M. K., Veldink, J., Gaunt, T. R., Runz, H., ... Westra, H.-J. (2023). Brain expression quantitative trait locus and network analyses reveal downstream effects and putative drivers for brain-related diseases. *Nature Genetics*, 55(3), Article 3. <https://doi.org/10.1038/s41588-023-01300-6>
- Delaneau, O., Marchini, J., & Zagury, J. F. (2011). A linear complexity phasing method for thousands of genomes. *Nature Methods* 2011 9:2, 9(2), 179–181. <https://doi.org/10.1038/nmeth.1785>
- Delaneau, O., Zagury, J.-F., & Marchini, J. (2013). Improved whole-chromosome phasing for disease and population genetic studies. *Nature Methods*, 10(1), Article 1. <https://doi.org/10.1038/nmeth.2307>
- Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., Epstein, C. B., Fritze, S., Harrow, J., Kaul, R., Khatun, J., Lajoie, B. R., Landt, S. G., Lee, B. K., Pauli, F.,

- Rosenbloom, K. R., Sabo, P., Safi, A., Sanyal, A., ... Lochovsky, L. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414), 57–74. <https://doi.org/10.1038/nature11247>
- Durbin, R. M., Altshuler, D., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Collins, F. S., De La Vega, F. M., Donnelly, P., Egholm, M., Flicek, P., Gabriel, S. B., Gibbs, R. A., Knoppers, B. M., Lander, E. S., Lehrach, H., Mardis, E. R., McVean, G. A., ... The Translational Genomics Research Institute. (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467(7319), Article 7319. <https://doi.org/10.1038/nature09534>
- Evans, L. M., Tahmasbi, R., Vrieze, S. I., Abecasis, G. R., Das, S., Gazal, S., Bjelland, D. W., de Candia, T. R., Haplotype Reference Consortium, Goddard, M. E., Neale, B. M., Yang, J., Visscher, P. M., & Keller, M. C. (2018). Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits. *Nature genetics*, 50(5), 737–745. <https://doi.org/10.1038/s41588-018-0108-x>
- Gaziano, J. M., Concato, J., Brophy, M., Fiore, L., Pyarajan, S., Breeling, J., Whitbourne, S., Deen, J., Shannon, C., Humphries, D., Guarino, P., Aslan, M., Anderson, D., LaFleur, R., Hammond, T., Schaa, K., Moser, J., Huang, G., Muralidhar, S., ... O'Leary, T. J. (2016). Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *Journal of Clinical Epidemiology*, 70, 214–223. <https://doi.org/10.1016/j.jclinepi.2015.09.016>
- Gazzellone, M. J., Zarrei, M., Burton, C. L., Walker, S., Uddin, M., Shaheen, S. M., Coste, J., Rajendram, R., Schachter, R. J., Colasanto, M., Hanna, G. L., Rosenberg, D. R., Soreni, N., Fitzgerald, K. D., Marshall, C. R., Buchanan, J. A., Merico, D., Arnold, P. D., & Scherer, S. W. (2016). Uncovering obsessive-compulsive disorder risk genes in a pediatric cohort by high-resolution analysis of copy number variation. *Journal of Neurodevelopmental Disorders*, 8(1), 1–10. <https://doi.org/10.1186/s11689-016-9170-9>
- Goldstein, J. I., Crenshaw, A., Carey, J., Grant, G. B., Maguire, J., Fromer, M., O'Dushlaine, C., Moran, J. L., Chambert, K., Stevens, C., Swedish Schizophrenia Consortium, ARRA

- Autism Sequencing Consortium, Sklar, P., Hultman, C. M., Purcell, S., McCarroll, S. A., Sullivan, P. F., Daly, M. J., & Neale, B. M. (2012). zCall: A rare variant caller for array-based genotyping: genetics and population analysis. *Bioinformatics (Oxford, England)*, 28(19), 2543–2545. <https://doi.org/10.1093/bioinformatics/bts479>
- Grotzinger, A.D., Rhemtulla, M., de Vlaming, R. et al. Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. *Nat Hum Behav* 3, 513–525 (2019). <https://doi.org/10.1038/s41562-019-0566-x>
- Guo, Y., He, J., Zhao, S., Wu, H., Zhong, X., Sheng, Q., Samuels, D. C., Shyr, Y., & Long, J. (2014). Illumina human exome genotyping array clustering and quality control. *Nature Protocols*, 9(11), 2643–2662. <https://doi.org/10.1038/nprot.2014.174>
- Gur, R. C., Richard, J., Calkins, M. E., Chiavacci, R., Hansen, J. A., Bilker, W. B., Loughhead, J., Connolly, J. J., Qiu, H., Mentch, F. D., Abou-Sleiman, P. M., Hakonarson, H., & Gur, R. E. (2012). Age group and sex differences in performance on a computerized neurocognitive battery in children age 8-21. *Neuropsychology*, 26(2), 251–265. <https://doi.org/10.1037/a0026712>
- Halvorsen M, Samuels J, Wang Y, Greenberg BD, Fyer AJ, McCracken JT, *et al.* (2021): Exome sequencing in obsessive-compulsive disorder reveals a burden of rare damaging coding variants. *Nat Neurosci* 24: 1071–1076.
- Harrington, K. M., Nguyen, X.-M. T., Song, R. J., Hannagan, K., Quaden, R., Gagnon, D. R., Cho, K., Deen, J. E., Muralidhar, S., O’Leary, T. J., Gaziano, J. M., Whitbourne, S. B., & VA Million Veteran Program. (2019). Gender Differences in Demographic and Health Characteristics of the Million Veteran Program Cohort. *Women’s Health Issues: Official Publication of the Jacobs Institute of Women’s Health*, 29 Suppl 1, S56–S66. <https://doi.org/10.1016/j.whi.2019.04.012>
- Helgeland, Ø., Vaudel, M., Juliusson, P. B., Lingaas Holmen, O., Juodakis, J., Bacelis, J., Jacobsson, B., Lindekleiv, H., Hveem, K., Lie, R. T., Knudsen, G. P., Stoltenberg, C., Magnus, P., Sagen, J. V., Molven, A., Johansson, S., & Njølstad, P. R. (2019). Genome-wide association study reveals dynamic role of genetic variation in infant and early

- childhood growth. *Nature Communications*, 10(1), 1–10. <https://doi.org/10.1038/s41467-019-12308-0>
- Howie, B. N., Donnelly, P., & Marchini, J. (2009). A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLOS Genetics*, 5(6), e1000529. <https://doi.org/10.1371/journal.pgen.1000529>
- Hunter-Zinck, H., Shi, Y., Li, M., Gorman, B. R., Ji, S.-G., Sun, N., Webster, T., Liem, A., Hsieh, P., Devineni, P., Karnam, P., Gong, X., Radhakrishnan, L., Schmidt, J., Assimes, T. L., Huang, J., Pan, C., Humphries, D., Brophy, M., ... Pyarajan, S. (2020). Genotyping Array Design and Data Quality Control in the Million Veteran Program. *American Journal of Human Genetics*, 106(4), 535–548. <https://doi.org/10.1016/j.ajhg.2020.03.004>
- Hyde, C., Nagle, M., Tian, C., Chen, X., Paciga, S., Wendland, J., Tung, J., Hinds, D., Perlis, R., & Winslow, A. (2016). Identification of 15 genetic loci associated with risk of major depression in individuals of European descent. *Nature Genetics*, 48(9), 1031–1036. <https://doi.org/10.1038/NG.3623>
- International Obsessive Compulsive Disorder Foundation Genetics Collaborative (IOCDF-GC) and OCD Collaborative Genetics Association Studies (OCGAS) (2018). Revealing the complex genetic architecture of obsessive-compulsive disorder using meta-analysis. *Molecular psychiatry*, 23(5), 1181–1188. <https://doi.org/10.1038/mp.2017.154>
- Jun, G., Flickinger, M., Hetrick, K. N., Romm, J. M., Doheny, K. F., Abecasis, G. R., Boehnke, M., & Kang, H. M. (2012). Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *American Journal of Human Genetics*, 91(5), 839–848. <https://doi.org/10.1016/j.ajhg.2012.09.004>
- Kaufman, J., Birmaher, B., Brent, D. A., Ryan, N. D., & Rao, U. (2000). K-SADS-PL. *Journal of the American Academy of Child & Adolescent Psychiatry*, 39(10), 1208–1208. <https://doi.org/10.1097/00004583-200010000-00002>
- Korn, J. M., Kuruvilla, F. G., McCarroll, S. A., Wysoker, A., Nemesh, J., Cawley, S., Hubbell, E., Veitch, J., Collins, P. J., Darvishi, K., Lee, C., Nizzari, M. M., Gabriel, S. B., Purcell, S., Daly, M. J., & Altshuler, D. (2008). Integrated genotype calling and association analysis of

- SNPs, common copy number polymorphisms and rare CNVs. *Nature Genetics*, 40(10), Article 10. <https://doi.org/10.1038/ng.237>
- Krokstad, S., Langhammer, A., Hveem, K., Holmen, T. L., Midthjell, K., Stene, T. R., Bratberg, G., Heggland, J., & Holmen, J. (2013). Cohort profile: The HUNT study, Norway. *International Journal of Epidemiology*. <https://doi.org/10.1093/ije/dys095>
- Lai, D., Wetherill, L., Bertelsen, S., Carey, C. E., Kamarajan, C., Kapoor, M., Meyers, J. L., Anokhin, A. P., Bennett, D. A., Bucholz, K. K., Chang, K. K., De Jager, P. L., Dick, D. M., Hesselbrock, V., Kramer, J., Kuperman, S., Nurnberger Jr, J. I., Raj, T., Schuckit, M., ... Foroud, T. (2019). Genome-wide association studies of alcohol dependence, DSM-IV criterion count and individual criteria. *Genes, Brain and Behavior*, 18(6), e12579. <https://doi.org/10.1111/gbb.12579>
- Lai, D., Wetherill, L., Kapoor, M., Johnson, E. C., Schwandt, M., Ramchandani, V. A., Goldman, D., Joslyn, G., Rao, X., Liu, Y., Farris, S., Mayfield, R. D., Dick, D., Hesselbrock, V., Kramer, J., McCutcheon, V. V., Nurnberger, J., Tischfield, J., Goate, A., ... Schuckit, M. (2020). Genome-wide association studies of the self-rating of effects of ethanol (SRE). *Addiction Biology*, 25(2), e12800. <https://doi.org/10.1111/adb.12800>
- Lam, M., Awasthi, S., Watson, H. J., Goldstein, J., Panagiotaropoulou, G., Trubetskoy, V., Karlsson, R., Frei, O., Fan, C. C., De Witte, W., Mota, N. R., Mullins, N., Brügger, K., Hong Lee, S., Wray, N. R., Skarabis, N., Huang, H., Neale, B., Daly, M. J., ... Ripke, S. (2020). RICOPIIL: Rapid Imputation for COnsortias PIpeLIne. *Bioinformatics*, 36(3), 930–933. <https://doi.org/10.1093/bioinformatics/btz633>
- Leckman, J. F., Riddle, M. A., Hardin, M. T., Ort, S. I., Swartz, K. L., Stevenson, J., & Cohen, D. J. (1989). The Yale Global Tic Severity Scale: Initial Testing of a Clinician-Rated Scale of Tic Severity. *Journal of the American Academy of Child and Adolescent Psychiatry*, 28(4), 566–573. <https://doi.org/10.1097/00004583-198907000-00015>
- Lee, J. H., Cheng, R., Graff-Radford, N., Foroud, T., & Mayeux, R. (2008). Analyses of the national institute on aging late-onset Alzheimer's disease family study: Implication of additional

- loci. *Archives of Neurology*, 65(11), 1518–1526.
<https://doi.org/10.1001/archneur.65.11.1518>
- Leitsalu, L., Alavere, H., Tammesoo, M. L., Leego, E., & Metspalu, A. (2015). Linking a population biobank with national health registries—The Estonian experience. *Journal of Personalized Medicine*, 5(2). <https://doi.org/10.3390/jpm5020096>
- Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, *et al.* (2016): Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536: 285–291.
- Li, J. Z., Absher, D. M., Tang, H., Southwick, A. M., Casto, A. M., Ramachandran, S., Cann, H. M., Barsh, G. S., Feldman, M., Cavalli-Sforza, L. L., & Myers, R. M. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. *Science*, 319(5866), 1100–1104. <https://doi.org/10.1126/science.1153717>
- Loh, P. R., Danecek, P., Palamara, P. F., Fuchsberger, C., Reshef, Y. A., Finucane, H. K., Schoenherr, S., Forer, L., McCarthy, S., Abecasis, G. R., Durbin, R., & Price, A. L. (2016). Reference-based phasing using the Haplotype Reference Consortium panel. *Nature Genetics* 2016 48:11, 48(11), 1443–1448. <https://doi.org/10.1038/ng.3679>
- Loh, P. R., Palamara, P. F., & Price, A. L. (2016). Fast and accurate long-range phasing in a UK Biobank cohort. *Nature Genetics* 2016 48:7, 48(7), 811–816. <https://doi.org/10.1038/ng.3571>
- Mahjani, B., Dellenvall, K., Grahnat, A. C. S., Karlsson, G., Tuuliainen, A., Reichert, J., Mahjani, C. G., Klei, L., De Rubeis, S., Reichenberg, A., Devlin, B., Hultman, C. M., Buxbaum, J. D., Sandin, S., & Grice, D. E. (2020). Cohort profile: Epidemiology and Genetics of Obsessive–compulsive disorder and chronic tic disorders in Sweden (EGOS). *Social Psychiatry and Psychiatric Epidemiology*, 55(10), 1383–1393. <https://doi.org/10.1007/s00127-019-01822-7>
- Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., & Chen, W. M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics*, 26(22), 2867–2873. <https://doi.org/10.1093/BIOINFORMATICS/BTQ559>

- Marchini, J., Howie, B., Myers, S., McVean, G., & Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics*, 39(7), Article 7. <https://doi.org/10.1038/ng2088>
- Margraf, J., Cwik, J. C., Pflug, V., & Schneider, S. (2017). Strukturierte klinische Interviews zur Erfassung psychischer Störungen über die Lebensspanne. *Zeitschrift Für Klinische Psychologie Und Psychotherapie*, 46(3), 176–186. <https://doi.org/10.1026/1616-3443/a000430>
- Mars, N., Koskela, J. T., Ripatti, P., Kiiskinen, T. T. J., Havulinna, A. S., Lindbohm, J. V., Ahola-Olli, A., Kurki, M., Karjalainen, J., Palta, P., Neale, B. M., Daly, M., Salomaa, V., Palotie, A., Widén, E., & Ripatti, S. (2020). Polygenic and clinical risk scores and their impact on age at onset and prediction of cardiometabolic diseases and common cancers. *Nature Medicine*, 26(4), 549–557. <https://doi.org/10.1038/s41591-020-0800-0>
- Mataix-Cols, D., Hansen, B., Mattheisen, M., Karlsson, E. K., Addington, A. M., Boberg, J., Djurfeldt, D. R., Halvorsen, M., Lichtenstein, P., Solem, S., Lindblad-Toh, K., Nordic OCD and Related Disorders Consortium (NORDiC), Haavik, J., Kvale, G., Rück, C., & Crowley, J. J. (2020). Nordic OCD & Related Disorders Consortium: Rationale, design, and methods. *American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics: The Official Publication of the International Society of Psychiatric Genetics*, 183(1), 38–50. <https://doi.org/10.1002/ajmg.b.32756>
- Mattheisen, M., Samuels, J. F., Wang, Y., Greenberg, B. D., Fyer, A. J., Mccracken, J. T., Geller, D. A., Murphy, D. L., Knowles, J. A., Grados, M. A., Riddle, M. A., Rasmussen, S. A., McLaughlin, N. C., Nurmi, E. L., Askland, K. D., Qin, H. D., Cullen, B. A., Piacentini, J., Pauls, D. L., ... Nestadt, G. (2015). Genome-wide association study in obsessive-compulsive disorder: Results from the OCGAS. *Molecular Psychiatry*, 20(3), 337–344. <https://doi.org/10.1038/mp.2014.43>
- Mbatchou, J., Barnard, L., Backman, J., Marcketta, A., Kosmicki, J. A., Ziyatdinov, A., Benner, C., O'Dushlaine, C., Barber, M., Boutkov, B., Habegger, L., Ferreira, M., Baras, A., Reid, J., Abecasis, G., Maxwell, E., & Marchini, J. (2021). Computationally efficient whole-genome

- regression for quantitative and binary traits. *Nature Genetics*, 53(7), Article 7.
<https://doi.org/10.1038/s41588-021-00870-7>
- McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., Teumer, A., Kang, H. M., Fuchsberger, C., Danecek, P., Sharp, K., Luo, Y., Sidore, C., Kwong, A., Timpson, N., Koskinen, S., Vrieze, S., Scott, L. J., Zhang, H., Mahajan, A., ... Marchini, J. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics* 2016 48:10, 48(10), 1279–1283. <https://doi.org/10.1038/ng.3643>
- Mitt, M., Kals, M., Pärn, K., Gabriel, S. B., Lander, E. S., Palotie, A., Ripatti, S., Morris, A. P., Metspalu, A., Esko, T., Mägi, R., & Palta, P. (2017). Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage WGS-based imputation reference panel. *European Journal of Human Genetics*, 25(7). <https://doi.org/10.1038/ejhg.2017.51>
- O'Connell, J. R., & Weeks, D. E. (1998). PedCheck: A program for identification of genotype incompatibilities in linkage analysis. *American Journal of Human Genetics*, 63(1), 259–266. <https://doi.org/10.1086/301904>
- Pato, M. T., Sobell, J. L., Medeiros, H., Abbott, C., Sklar, B. M., Buckley, P. F., Bromet, E. J., Escamilla, M. A., Fanous, A. H., Lehrer, D. S., Macciardi, F., Malaspina, D., McCarroll, S. A., Marder, S. R., Moran, J., Morley, C. P., Nicolini, H., Perkins, D. O., Purcell, S. M., ... Pato, C. N. (2013). The genomic psychiatry cohort: Partners in discovery. *American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics: The Official Publication of the International Society of Psychiatric Genetics*, 162B(4), 306–312. <https://doi.org/10.1002/ajmg.b.32160>
- Pedersen, B. S., & Quinlan, A. R. (2017). Who's Who? Detecting and Resolving Sample Anomalies in Human DNA Sequencing Studies with Peddy. *The American Journal of Human Genetics*, 100(3), 406–413. <https://doi.org/10.1016/j.ajhg.2017.01.017>
- Pedersen, C. B., Bybjerg-Grauholm, J., Pedersen, M. G., Grove, J., Agerbo, E., Bækvad-Hansen, M., Poulsen, J. B., Hansen, C. S., McGrath, J. J., Als, T. D., Goldstein, J. I., Neale, B. M., Daly, M. J., Hougaard, D. M., Mors, O., Nordentoft, M., Børglum, A. D., Werge, T., &

- Mortensen, P. B. (2018). The iPSYCH2012 case-cohort sample: New directions for unravelling genetic and environmental architectures of severe mental disorders. *Mol Psychiatry*, 23(1), 6–14. <https://doi.org/10.1038/mp.2017.196>
- Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8), 904–909. <https://doi.org/10.1038/NG1847>
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., De Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, 81(3), 559–575. <https://doi.org/10.1086/519795>
- Rovira, P., Demontis, D., Sánchez-Mora, C., Zayats, T., Klein, M., Mota, N. R., Weber, H., Garcia-Martínez, I., Pagerols, M., Vilar-Ribó, L., Arribas, L., Richarte, V., Corrales, M., Fadeuilhe, C., Bosch, R., Martin, G. E., Almos, P., Doyle, A. E., Grevet, E. H., ... Ribasés, M. (2020). Shared genetic background between children and adults with attention deficit/hyperactivity disorder. *Neuropsychopharmacology*, 45(10), Article 10. <https://doi.org/10.1038/s41386-020-0664-5>
- Satterthwaite, T. D., Connolly, J. J., Ruparel, K., Calkins, M. E., Jackson, C., Elliott, M. A., Roalf, D. R., Hopson, R., Prabhakaran, K., Behr, M., Qiu, H., Mentch, F. D., Chiavacci, R., Sleiman, P. M. A., Gur, R. C., Hakonarson, H., & Gur, R. E. (2016). The Philadelphia Neurodevelopmental Cohort: A publicly available resource for the study of normal and abnormal brain development in youth. *NeuroImage*, 124(Pt B), 1115–1119. <https://doi.org/10.1016/j.neuroimage.2015.03.056>
- Scahill, L., Riddle, M. A., McSwiggin-Hardin, M., Ort, S. I., King, R. A., Goodman, W. K., Cicchetti, D., & Leckman, J. F. (1997). Children's Yale-Brown Obsessive Compulsive Scale: Reliability and validity. *Journal of the American Academy of Child and Adolescent Psychiatry*, 36(6), 844–852. <https://doi.org/10.1097/00004583-199706000-00023>

- Schmidt, M. H., Döpfner, M., Berner, W., Flechtner, H., Lehmkuhl, G., & Steinhausen, H.-C. E. (2000). *Psychopathologisches Befund-System für Kinder und Jugendliche (CASCAP-D)*. 28(1):64-64. <https://doi.org/10.1024//1422-4917.28.1.64>
- Schumacher, F. R., Berndt, S. I., Siddiq, A., Jacobs, K. B., Wang, Z., Lindstrom, S., Stevens, V. L., Chen, C., Mondul, A. M., Travis, R. C., Stram, D. O., Eeles, R. A., Easton, D. F., Giles, G., Hopper, J. L., Neal, D. E., Hamdy, F. C., Donovan, J. L., Muir, K., ... Kraft, P. (2011). Genome-wide association study identifies new prostate cancer susceptibility loci. *Human Molecular Genetics*, 20(19), 3867–3875. <https://doi.org/10.1093/hmg/ddr295>
- Sherva, R., Wang, Q., Kranzler, H., Zhao, H., Koesterer, R., Herman, A., Farrer, L. A., & Gelernter, J. (2016). Genome-wide Association Study of Cannabis Dependence Severity, Novel Risk Variants, and Shared Genetic Risks. *JAMA Psychiatry*, 73(5), 472–480. <https://doi.org/10.1001/jamapsychiatry.2016.0036>
- Solovieff, N., Hartley, S. W., Baldwin, C. T., Perls, T. T., Steinberg, M. H., & Sebastiani, P. (2010). Clustering by genetic ancestry using genome-wide SNP data. *BMC Genetics*, 11, 108. <https://doi.org/10.1186/1471-2156-11-108>
- Stewart, S. E., Mayerfeld, C., Arnold, P. D., Crane, J. R., O'Dushlaine, C., Fagerness, J. A., Yu, D., Scharf, J. M., Chan, E., Kassam, F., Moya, P. R., Wendland, J. R., Delorme, R., Richter, M. A., Kennedy, J. L., Veenstra-Vanderweele, J., Samuels, J., Greenberg, B. D., McCracken, J. T., ... Mathews, C. A. (2013). Meta-analysis of association between obsessive-compulsive disorder and the 3' region of neuronal glutamate transporter gene SLC1A1. *American Journal of Medical Genetics, Part B: Neuropsychiatric Genetics*, 162(4), 367–379. <https://doi.org/10.1002/ajmg.b.32137>
- Strom, N. I., Verhulst, B., Bacanu, S.-A., Cheesman, R., Purves, K. L., Gedik, H., ... Hettema, J. (2024). Genome-wide association study of major anxiety disorders in 122,341 European-ancestry cases identifies 58 loci and highlights GABAergic signaling. medRxiv 2024.07.03.24309466; doi: <https://doi.org/10.1101/2024.07.03.24309466>
- The 1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. *Nature*, 526(7571), 68–74. <https://doi.org/10.1038/nature15393>

- Võsa, U., Claringbould, A., Westra, H.-J., Bonder, M. J., Deelen, P., Zeng, B., Kirsten, H., Saha, A., Kreuzhuber, R., Yazar, S., Brugge, H., Oelen, R., de Vries, D. H., van der Wijst, M. G. P., Kasela, S., Pervjakova, N., Alves, I., Favé, M.-J., Agbessi, M., ... Franke, L. (2021). Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nature Genetics*, 53(9), 1300–1310. <https://doi.org/10.1038/s41588-021-00913-z>
- Wainberg, M., Merico, D., Keller, M. C., Fauman, E. B., & Tripathy, S. J. (2022). Predicting causal genes from psychiatric genome-wide association studies using high-level etiological knowledge. *Molecular Psychiatry*, 27(7), Article 7. <https://doi.org/10.1038/s41380-022-01542-6>
- Wang, C., Zhan, X., Bragg-Gresham, J., Kang, H. M., Stambolian, D., Chew, E. Y., Branham, K. E., Heckenlively, J., Fulton, R., Wilson, R. K., Mardis, E. R., Lin, X., Swaroop, A., Zöllner, S., & Abecasis, G. R. (2014). Ancestry estimation and control of population stratification for sequence-based association studies. *Nature Genetics*, 46(4), 409–415. <https://doi.org/10.1038/ng.2924>
- Wang S, Mandell JD, Kumar Y, Sun N, Morris MT, Arbelaez J, *et al.* (2018): De Novo Sequence and Copy Number Variants Are Strongly Associated with Tourette Disorder and Implicate Cell Polarity in Pathogenesis. *Cell Rep* 25: 3544.
- Warren, H. R., Evangelou, E., Cabrera, C. P., Gao, H., Ren, M., Mifsud, B., Ntalla, I., Surendran, P., Liu, C., Cook, J. P., Kraja, A. T., Drenos, F., Loh, M., Verweij, N., Marten, J., Karaman, I., Segura Lepe, M. P., O'Reilly, P. F., Knight, J., ... Morris, A. P. (2017). Genome-wide association analysis identifies novel blood pressure loci and offers biological insights into cardiovascular risk. *Nature Genetics*, 49(3), 403–415. <https://doi.org/10.1038/ng.3768>
- Wray, N.R., Ripke, S., Mattheisen, M. et al. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat Genet* 50, 668–681 (2018). <https://doi.org/10.1038/s41588-018-0090-3>

- Yang, J., Lee, S. H., Goddard, M. E., & Visscher, P. M. (2011). GCTA: a tool for genome-wide complex trait analysis. *American journal of human genetics*, 88(1), 76–82. <https://doi.org/10.1016/j.ajhg.2010.11.011>
- Zhou, W., Nielsen, J. B., Fritsche, L. G., Dey, R., Gabrielsen, M. E., Wolford, B. N., LeFaive, J., VandeHaar, P., Gagliano, S. A., Gifford, A., Bastarache, L. A., Wei, W. Q., Denny, J. C., Lin, M., Hveem, K., Kang, H. M., Abecasis, G. R., Willer, C. J., & Lee, S. (2018). Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nature Genetics*, 50(9), 1335–1341. <https://doi.org/10.1038/s41588-018-0184-y>