A contribution of novel CNVs to schizophrenia from a genome-wide study of 41,321 subjects

CNV Analysis Group and the Schizophrenia Working Group of the Psychiatric Genomics Consortium

Authors:

Christian R. Marshall[1]*, Daniel P. Howrigan[2,3]*, Daniele Merico[1]*, Bhooma Thiruvahindrapuram[1], Wenting Wu[4,5], Douglas S. Greer[4,5], Danny Antaki[4,5], Aniket Shetty[4,5], Peter A. Holmans[6,7], Dalila Pinto[8,9], Madhusudan Gujral[4,5], William M. Brandler[4,5], Dheeraj Malhotra[4,5,10], Zhouzhi Wang[1], Karin V. Fuentes Fajarado[4,5], Stephan Ripke[2,3], Ingrid Agartz[11,12,13], Esben Agerbo[14,15,16], Margot Albus[17], Madeline Alexander[18], Farooq Amin[19,20], Joshua Atkins[21,22], Silviu A. Bacanu[23] ,Richard A. Belliveau Jr[3], Sarah E. Bergen[3,24], Marcelo Bertalan[16,25], Elizabeth Bevilacqua[3], Tim B. Bigdeli[23], Donald W. Black[26], Richard Bruggeman[27], Nancy G. Buccola[28], Randy L. Buckner[29,30,31], Brendan Bulik-Sullivan[2,3], William Byerley[32], Wiepke Cahn[33], Guiqing Cai[8,34], Murray J. Cairns[21,35,36], Dominique Campion[37], Rita M. Cantor[38], Vaughan J. Carr[35,39], Noa Carrera[6], Stanley V. Catts[35,40], Kimberley D. Chambert[3], Wei Cheng[41], C. Robert Cloninger[42], David Cohen[43], Paul Cormican[44], Nick Craddock[6,7], Benedicto Crespo-Facorro[45,46], James J. Crowley[47], David Curtis[48,49], Michael Davidson[50], Kenneth L, Davis[8], Franziska Degenhardt[51,52], Jurgen Del Favero[53], Lynn E. DeLisi[54,55], Ditte Demontis[16,56,57], Dimitris Dikeos[58], Timothy Dinan[59], Srdjan Djurovic[11,60], Gary Donohoe[44,61], Elodie Drapeau[8], Jubao Duan[62,63], Frank Dudbridge[64], Peter Eichhammer[65], Johan Eriksson[66,67,68], Valentina Escott-Price[6], Laurent Essioux[69], Ayman H. Fanous[70,71,72,73], Kai-How Farh[2], Martilias S. Farrell[47], Josef Frank[74], Lude Franke[75], Robert Freedman[76], Nelson B. Freimer[77], Joseph I. Friedman[8], Andreas J. Forstner[51,52], Menachem Fromer[2,3,78,79], Giulio Genovese[3], Lyudmila Georgieva[6], Elliot S. Gershon[80], Ina Giegling[81,82], Paola Giusti-Rodríguez[47], Stephanie Godard[83], Jacqueline I. Goldstein[2,84], Jacob Gratten[85], Lieuwe de Haan[86], Marian L. Hamshere[6], Mark Hansen[87], Thomas Hansen[16,25], Vahram Haroutunian[8,88,89], Annette M. Hartmann[81], Frans A. Henskens[35,36,90], Stefan Herms[51,52,91], Joel N. Hirschhorn[84,92,93], Per Hoffmann[51,52,91], Andrea Hofman[51,52], Mads V. Hollegaard[94], David M. Hougaard[94], Hailiang Huang[2,84], Masashi Ikeda[95], Inge Joa[96], Anna K Kähler[24], René S Kahn[33], Luba Kalaydjieva[97,167], Juha Karjalainen[75], David Kavanagh[6], Matthew C. Keller[99], Brian J. Kelly[36], James L. Kennedy[100,101,102], Yunjung Kim[47], James A. Knowles[103], Bettina Konte[81], Claudine Laurent[18,104], Phil Lee[2,3,79], S. Hong Lee[85], Sophie E. Legge[6], Bernard Lerer[105], Deborah L. Levy[55,106], Kung-Yee Liang[107], Jeffrey Lieberman[108], Jouko Lönnqvist[109], Carmel M. Loughland[35,36], Patrik K.E. Magnusson[24], Brion S. Maher[110], Wolfgang Maier[111], Jacques Mallet[112], Manuel Mattheisen[16,56,57,113], Morten Mattingsdal[11,114], Robert W McCarley[54,55], Colm McDonald[115], Andrew M. McIntosh[116,117], Sandra Meier[74], Carin J. Meijer[86], Ingrid Melle[11,118], Raquelle I. Mesholam-Gately[55,119], Andres Metspalu[120], Patricia T. Michie[35,121], Lili Milani[120], Vihra Milanova[122], Younes Mokrab[123], Derek W. Morris[44,61], Ole Mors[16,57,124], Bertram Müller-Myhsok[125,126,127], Kieran C. Murphy[128], Robin M. Murray[129], Inez Myin-Germeys[130], Igor Nenadic[131], Deborah A. Nertney[132], Gerald Nestadt[133], Kristin K. Nicodemus[134], Laura Nisenbaum[135], Annelie Nordin[136], Eadbhard O'Callaghan[137], Colm O'Dushlaine[3], Sang-Yun Oh[138], Ann Olincy[76], Line Olsen[16,25], F. Anthony O'Neill[139], Jim Van Os[130,140], Christos Pantelis[35,141],

George N. Papadimitriou[58], Elena Parkhomenko[8], Michele T. Pato[103], Tiina Paunio[142], Psychosis Endophenotypes International Consortium, Diana O. Perkins[143], Tune H. Pers[84,93,144], Olli Pietiläinen[142,145], Jonathan Pimm[49], Andrew J. Pocklington[6], John Powell[129], Alkes Price[84,146], Ann E. Pulver[133], Shaun M. Purcell[78], Digby Quested[147], Henrik B. Rasmussen[16,25], Abraham Reichenberg[8,89], Mark A. Reimers[23], Alexander L. Richards[6,7], Joshua L. Roffman[30,31], Panos Roussos[78,148], Douglas M. Ruderfer[6,78], Veikko Salomaa[67], Alan R. Sanders[62,63], Adam Savitz[149], Ulrich Schall[35,36], Thomas G. Schulze[74,150], Sibylle G. Schwab[151], Edward M. Scolnick[3], Rodney J. Scott[21,35,152], Larry J. Seidman[55,119], Jianxin Shi[153], Jeremy M. Silverman[8,154], Jordan W. Smoller[3,79], Erik Söderman[13], Chris C.A. Spencer[155], Eli A. Stahl[78,84], Eric Strengman[33,156], Jana Strohmaier[74], T. Scott Stroup[108], Jaana Suvisaari[109], Dragan M. Svrakic[42], Jin P. Szatkiewicz[47], Srinivas Thirumalai[157], Paul A. Tooney[21,35,36], Juha Veijola[158,159], Peter M. Visscher[85], John Waddington[160], Dermot Walsh[161], Bradley T. Webb[23], Mark Weiser[50], Dieter B. Wildenauer[98], Nigel M. Williams[6], Stephanie Williams[47], Stephanie H. Witt[74], Aaron R. Wolen[23], Brandon K. Wormley[23], Naomi R Wray[85], Jing Qin Wu[21,35], Clement C. Zai[100,101], Wellcome Trust Case-Control Consortium 2, Rolf Adolfsson[136], Ole A. Andreassen[11,118], Douglas H.R. Blackwood[116], Anders D. Børglum[16,56,57,124], Elvira Bramon[162], Joseph D. Buxbaum[8,34,89,163], Sven Cichon[51,52,91,164], David A .Collier[123,165], Aiden Corvin[44], Mark J. Daly[2,3,84], Ariel Darvasi[166], Enrico Domenici[10], Tõnu Esko[84,92,93,120], Pablo V. Gejman[62,63], Michael Gill[44], Hugh Gurling[49], Christina M. Hultman[24], Nakao Iwata[95], Assen V. Jablensky[35,98,167,168], Erik G Jönsson[11,13], Kenneth S Kendler[23], George Kirov[6], Jo Knight[100,101,102], Douglas F. Levinson[18], Qingqin S Li[149], Steven A McCarroll[3,92], Andrew McQuillin[49], Jennifer L. Moran[3], Preben B. Mortensen[14,15,16], Bryan J. Mowry[85,132], Markus M. Nöthen[51,52], Roel A. Ophoff[33,38,77], Michael J. Owen[6,7], Aarno Palotie[3,79,145], Carlos N. Pato[103], Tracey L. Petryshen[3,55,169], Danielle Posthuma[170,171,172], Marcella Rietschel[74], Brien P. Riley[23], Dan Rujescu[81,82], Pamela Sklar[78,89,148], David St. Clair[173], James T.R. Walters[6], Thomas Werge[16,25,174], Patrick F. Sullivan[24,47,143], Michael C O'Donovan[6,7]†, Stephen W. Scherer[1,175]†, Benjamin M. Neale[2,3,79,84]†, Jonathan Sebat[4,5,176]†

*these authors contributed equally
†these authors co-supervised the study
Correspondence: jsebat@ucsd.edu

[1]The Centre for Applied Genomics and Program in Genetics and Genome Biology, The Hospital for Sick Children, Toronto, ON, Canada
[2]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts 02114, USA
[3]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA
[4]Beyster Center for Psychiatric Genomics, University of California, San Diego, La Jolla, CA 92093, USA
[5]Department of Psychiatry, University of California, San Diego, La Jolla, CA 92093, USA
[6]MRC Centre for Neuropsychiatric Genetics and Genomics, Institute of Psychological Medicine and Clinical Neurosciences, School of Medicine, Cardiff University, Cardiff, CF24 4HQ, UK
[7]National Centre for Mental Health, Cardiff University, Cardiff, CF24 4HQ, UK

[8]Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, New York 10029, USA
[9]Department of Genetics and Genomic Sciences, Seaver Autism Center, The Mindich Child Health & Development Institute, Icahn School of Medicine at Mount Sinai, New York, New York 10029, USA
[10]Neuroscience Discovery and Translational Area, Pharma Research & Early Development, F. Hoffmann-La Roche Ltd, CH-4070 Basel, Switzerland
[11]NORMENT, KG Jebsen Centre for Psychosis Research, Institute of Clinical Medicine, University of Oslo, 0424 Oslo, Norway
[12]Department of Psychiatry, Diakonhjemmet Hospital, 0319 Oslo, Norway
[13]Department of Clinical Neuroscience, Psychiatry Section, Karolinska Institutet, SE-17176 Stockholm, Sweden
[14]National Centre for Register-based Research, Aarhus University, DK-8210 Aarhus, Denmark
[15]Centre for Integrative Register-based Research, CIRRAU, Aarhus University, DK-8210 Aarhus, Denmark
[16]The Lundbeck Foundation Initiative for Integrative Psychiatric Research, iPSYCH, Denmark
[17]State Mental Hospital, 85540 Haar, Germany
[18]Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, California 94305, USA
[19]Department of Psychiatry and Behavioral Sciences, Emory University, Atlanta, Georgia 30322, USA
[20]Department of Psychiatry and Behavioral Sciences, Atlanta Veterans Affairs Medical Center, Atlanta, Georgia 30033, USA
[21]School of Biomedical Sciences and Pharmacy, University of Newcastle, Callaghan NSW 2308, Australia
[22]Hunter Medical Research Institute, New Lambton, New South Wales, Australia
[23]Virginia Institute for Psychiatric and Behavioral Genetics, Department of Psychiatry, Virginia Commonwealth University, Richmond, Virginia 23298, USA
[24]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm SE-17177, Sweden
[25]Institute of Biological Psychiatry, Mental Health Centre Sct. Hans, Mental Health Services Copenhagen, DK-4000, Denmark
[26]Department of Psychiatry, University of Iowa Carver College of Medicine, Iowa City, Iowa 52242, USA
[27]University Medical Center Groningen, Department of Psychiatry, University of Groningen, NL-9700 RB, The Netherlands
[28]School of Nursing, Louisiana State University Health Sciences Center, New Orleans, Louisiana 70112, USA
[29]Center for Brain Science, Harvard University, Cambridge, Massachusetts 02138, USA
[30]Department of Psychiatry, Massachusetts General Hospital, Boston, Massachusetts 02114, USA
[31]Athinoula A. Martinos Center, Massachusetts General Hospital, Boston, Massachusetts 02129, USA

[32]Department of Psychiatry, University of California at San Francisco, San Francisco, California, 94143 USA
[33]University Medical Center Utrecht, Department of Psychiatry, Rudolf Magnus Institute of Neuroscience, 3584 Utrecht, The Netherlands
[34]Department of Human Genetics, Icahn School of Medicine at Mount Sinai, New York, New York 10029, USA
[35]Schizophrenia Research Institute, Sydney NSW 2010, Australia
[36]Priority Centre for Translational Neuroscience and Mental Health, University of Newcastle, Newcastle NSW 2300, Australia
[37]Centre Hospitalier du Rouvray and INSERM U1079 Faculty of Medicine, 76301 Rouen, France
[38]Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, California 90095, USA
[39]School of Psychiatry, University of New South Wales, Sydney NSW 2031, Australia
[40]Royal Brisbane and Women's Hospital, University of Queensland, Brisbane QLD 4072, Australia
[41]Department of Computer Science, University of North Carolina, Chapel Hill, North Carolina 27514, USA
[42]Department of Psychiatry, Washington University, St. Louis, Missouri 63110, USA
[43]Department of Child and Adolescent Psychiatry, Assistance Publique Hospitaux de Paris, Pierre and Marie Curie Faculty of Medicine and Institute for Intelligent Systems and Robotics, Paris, 75013, France
[44]Neuropsychiatric Genetics Research Group, Department of Psychiatry, Trinity College Dublin, Dublin 8, Ireland
[45]University Hospital Marqués de Valdecilla, Instituto de Formación e Investigación Marqués de Valdecilla, University of Cantabria, E-39008 Santander, Spain
[46]Centro Investigación Biomédica en Red Salud Mental, Madrid, Spain
[47]Department of Genetics, University of North Carolina, Chapel Hill, North Carolina 27599-7264, USA
[48]Department of Psychological Medicine, Queen Mary University of London, London E1 1BB, UK
[49]Molecular Psychiatry Laboratory, Division of Psychiatry, University College London, London WC1E 6JJ, UK
[50]Sheba Medical Center, Tel Hashomer 52621, Israel
[51]Institute of Human Genetics, University of Bonn, D-53127 Bonn, Germany
[52]Department of Genomics, Life and Brain Center, D-53127 Bonn, Germany
[53]Applied Molecular Genomics Unit, VIB Department of Molecular Genetics, University of Antwerp, B-2610 Antwerp, Belgium
[54]VA Boston Health Care System, Brockton, Massachusetts 02301, USA
[55]Department of Psychiatry, Harvard Medical School, Boston, Massachusetts 02115, USA
[56]Department of Biomedicine, Aarhus University, DK-8000 Aarhus C, Denmark
[57]Centre for Integrative Sequencing, iSEQ, Aarhus University, DK-8000 Aarhus C, Denmark
[58]First Department of Psychiatry, University of Athens Medical School, Athens 11528, Greece
[59]Department of Psychiatry, University College Cork, Co. Cork, Ireland
[60]Department of Medical Genetics, Oslo University Hospital, 0424 Oslo, Norway

[61]Cognitive Genetics and Therapy Group, School of Psychology and Discipline of Biochemistry, National University of Ireland Galway, Co. Galway, Ireland
[62]Department of Psychiatry and Behavioral Sciences, NorthShore University HealthSystem, Evanston, Illinois 60201, USA
[63]Department of Psychiatry and Behavioral Neuroscience, University of Chicago, Chicago, Illinois 60637, USA
[64]Department of Non-Communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, London WC1E 7HT, UK
[65]Department of Psychiatry, University of Regensburg, 93053 Regensburg, Germany
[66]Folkhälsan Research Center, Helsinki, Finland, Biomedicum Helsinki 1, Haartmaninkatu 8, FI-00290, Helsinki, Finland
[67]National Institute for Health and Welfare, P.O. BOX 30, FI-00271 Helsinki, Finland
[68]Department of General Practice, Helsinki University Central Hospital, University of Helsinki P.O. BOX 20, Tukholmankatu 8 B, FI-00014, Helsinki, Finland
[69]Translational Technologies and Bioinformatics, Pharma Research and Early Development, F.Hoffman-La Roche, CH-4070 Basel, Switzerland
[70]Mental Health Service Line, Washington VA Medical Center, Washington DC 20422, USA
[71]Department of Psychiatry, Georgetown University, Washington DC 20057, USA
[72]Department of Psychiatry, Virginia Commonwealth University, Richmond, Virginia 23298, USA
[73]Department of Psychiatry, Keck School of Medicine at University of Southern California, Los Angeles, California 90033, USA
[74]Department of Genetic Epidemiology in Psychiatry, Central Institute of Mental Health, Medical Faculty Mannheim, University of Heidelberg, Heidelberg, D-68159 Mannheim, Germany
[75]Department of Genetics, University of Groningen, University Medical Centre Groningen, 9700 RB Groningen, The Netherlands
[76]Department of Psychiatry, University of Colorado Denver, Aurora, Colorado 80045, USA
[77]Center for Neurobehavioral Genetics, Semel Institute for Neuroscience and Human Behavior, University of California, Los Angeles, California 90095, USA
[78]Division of Psychiatric Genomics, Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, New York 10029, USA
[79]Psychiatric and Neurodevelopmental Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts 02114, USA
[80]Departments of Psychiatry and Human Genetics, University of Chicago, Chicago, Illinois 60637 USA
[81]Department of Psychiatry, University of Halle, 06112 Halle, Germany
[82]Department of Psychiatry, University of Munich, 80336, Munich, Germany
[83]Departments of Psychiatry and Human and Molecular Genetics, INSERM, Institut de Myologie, Hôpital de la Pitiè-Salpêtrière, Paris, 75013, France
[84]Medical and Population Genetics Program, Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA
[85]Queensland Brain Institute, The University of Queensland, Brisbane, QLD 4072, Australia
[86]Academic Medical Centre University of Amsterdam, Department of Psychiatry, 1105 AZ Amsterdam, The Netherlands

[87]Illumina, La Jolla, California, California 92122, USA
[88]J.J. Peters VA Medical Center, Bronx, New York, New York 10468, USA
[89]Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, New York, New York 10029, USA
[90]School of Electrical Engineering and Computer Science, University of Newcastle, Newcastle NSW 2308, Australia
[91]Division of Medical Genetics, Department of Biomedicine, University of Basel, Basel, CH-4058, Switzerland
[92]Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA
[93]Division of Endocrinology and Center for Basic and Translational Obesity Research, Boston Children's Hospital, Boston, Massachusetts 02115, USA
[94]Section of Neonatal Screening and Hormones, Department of Clinical Biochemistry, Immunology and Genetics, Statens Serum Institut, Copenhagen, DK-2300, Denmark
[95]Department of Psychiatry, Fujita Health University School of Medicine, Toyoake, Aichi, 470-1192, Japan
[96]Regional Centre for Clinical Research in Psychosis, Department of Psychiatry, Stavanger University Hospital, 4011 Stavanger, Norway
[97]Centre for Medical Research, The University of Western Australia, Perth, WA 6009, Australia
[98]School of Psychiatry and Clinical Neurosciences, The University of Western Australia, Perth, WA 6009, Australia
[99]Department of Psychology, University of Colorado Boulder, Boulder, Colorado 80309, USA
[100]Campbell Family Mental Health Research Institute, Centre for Addiction and Mental Health, Toronto, Ontario, M5T 1R8, Canada
[101]Department of Psychiatry, University of Toronto, Toronto, Ontario, M5T 1R8, Canada
[102]Institute of Medical Science, University of Toronto, Toronto, Ontario, M5S 1A8, Canada
[103]Department of Psychiatry and Zilkha Neurogenetics Institute, Keck School of Medicine at University of Southern California, Los Angeles, California 90089, USA
[104]Department of Child and Adolescent Psychiatry, Pierre and Marie Curie Faculty of Medicine, Paris 75013, France
[105]Department of Psychiatry, Hadassah-Hebrew University Medical Center, Jerusalem 91120, Israel
[106]Psychology Research Laboratory, McLean Hospital, Belmont, MA
[107]Department of Biostatistics, Johns Hopkins University Bloomberg School of Public Health, Baltimore, Maryland 21205, USA
[108]Department of Psychiatry, Columbia University, New York, New York 10032, USA
[109]Department of Mental Health and Substance Abuse Services, National Institute for Health and Welfare, P.O. BOX 30, FI-00271 Helsinki, Finland
[110]Department of Mental Health, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland 21205, USA
[111]Department of Psychiatry, University of Bonn, D-53127 Bonn, Germany
[112]Centre National de la Recherche Scientifique, Laboratoire de Génétique Moléculaire de la Neurotransmission et des Processus Neurodégénératifs, Hôpital de la Pitié Salpêtrière, 75013, Paris, France
[113]Department of Genomics Mathematics, University of Bonn, D-53127 Bonn, Germany

[114]Research Unit, Sørlandet Hospital, 4604 Kristiansand, Norway
[115]Department of Psychiatry, National University of Ireland Galway, Co. Galway, Ireland
[116]Division of Psychiatry, University of Edinburgh, Edinburgh EH16 4SB, UK
[117]Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh, Edinburgh EH16 4SB, UK
[118]Division of Mental Health and Addiction, Oslo University Hospital, 0424 Oslo, Norway
[119]Massachusetts Mental Health Center Public Psychiatry Division of the Beth Israel Deaconess Medical Center, Boston, Massachusetts 02114, USA
[120]Estonian Genome Center, University of Tartu, Tartu 50090, Estonia
[121]School of Psychology, University of Newcastle, Newcastle NSW 2308, Australia
[122]First Psychiatric Clinic, Medical University, Sofia 1431, Bulgaria
[123]Eli Lilly and Company Limited, Erl Wood Manor, Sunninghill Road, Windlesham, Surrey, GU20 6PH UK
[124]Department P, Aarhus University Hospital, DK-8240 Risskov, Denmark
[125]Max Planck Institute of Psychiatry, 80336 Munich, Germany
[126]Institute of Translational Medicine, University of Liverpool, Liverpool L69 3BX, UK127Munich
[127]Cluster for Systems Neurology (SyNergy), 80336 Munich, Germany
[128]Department of Psychiatry, Royal College of Surgeons in Ireland, Dublin 2, Ireland
[129]King's College London, London SE5 8AF, UK
[130]Maastricht University Medical Centre, South Limburg Mental Health Research and Teaching Network, EURON, 6229 HX Maastricht, The Netherlands
[131]Department of Psychiatry and Psychotherapy, Jena University Hospital, 07743 Jena, Germany
[132]Queensland Centre for Mental Health Research, University of Queensland, Brisbane QLD 4076, Australia
[133]Department of Psychiatry and Behavioral Sciences, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA
[134]Department of Psychiatry, Trinity College Dublin, Dublin 2, Ireland
[135]Eli Lilly and Company, Lilly Corporate Center, Indianapolis, 46285 Indiana, USA
[136]Department of Clinical Sciences, Psychiatry, Umeå University, SE-901 87 Umeå, Sweden
[137]DETECT Early Intervention Service for Psychosis, Blackrock, Co. Dublin, Ireland
[138]Lawrence Berkeley National Laboratory, University of California at Berkeley, Berkeley, California 94720, USA
[139]Centre for Public Health, Institute of Clinical Sciences, Queen's University Belfast, Belfast BT12 6AB, UK
[140]Institute of Psychiatry, King's College London, London SE5 8AF, UK
[141]Melbourne Neuropsychiatry Centre, University of Melbourne & Melbourne Health, Melbourne VIC 3053, Australia
[142]Public Health Genomics Unit, National Institute for Health and Welfare, P.O. BOX 30, FI-00271 Helsinki, Finland
[143]Department of Psychiatry, University of North Carolina, Chapel Hill, North Carolina 27599-7160, USA
[144]Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, DK-2800, Denmark

[145]Institute for Molecular Medicine Finland, FIMM, University of Helsinki, P.O. BOX 20 FI-00014, Helsinki, Finland
[146]Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts 02115, USA
[147]Department of Psychiatry, University of Oxford, Oxford, OX3 7JX, UK
[148]Institute for Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, New York 10029, USA
[149]Neuroscience Therapeutic Area, Janssen Research and Development, Raritan, New Jersey 08869, USA
[150]Department of Psychiatry and Psychotherapy, University of Göttingen, 37073 Göttingen, Germany
[151]Psychiatry and Psychotherapy Clinic, University of Erlangen, 91054 Erlangen, Germany
[152]Hunter New England Health Service, Newcastle NSW 2308, Australia
[153]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland 20892, USA
[154]Research and Development, Bronx Veterans Affairs Medical Center, New York, New York 10468, USA
[155]Wellcome Trust Centre for Human Genetics, Oxford, OX3 7BN, UK
[156]Department of Medical Genetics, University Medical Centre Utrecht, Universiteitsweg 100, 3584 CG, Utrecht, The Netherlands
[157]Berkshire Healthcare NHS Foundation Trust, Bracknell RG12 1BQ, UK
[158]Department of Psychiatry, University of Oulu, P.O. BOX 5000, 90014, Finland
[159]University Hospital of Oulu, P.O. BOX 20, 90029 OYS, Finland
[160]Molecular and Cellular Therapeutics, Royal College of Surgeons in Ireland, Dublin 2, Ireland
[161]Health Research Board, Dublin 2, Ireland
[162]University College London, London WC1E 6BT, UK
[163]Department of Neuroscience, Icahn School of Medicine at Mount Sinai, New York, New York 10029, USA
[164]Institute of Neuroscience and Medicine (INM-1), Research Center Juelich, 52428 Juelich, Germany
[165]Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, King's College London, London, SE5 8AF, UK
[166]Department of Genetics, The Hebrew University of Jerusalem, 91905 Jerusalem, Israel
[167]The Perkins Institute for Medical Research, The University of Western Australia, Perth, WA 6009, Australia
[168]Centre for Clinical Research in Neuropsychiatry, School of Psychiatry and Clinical Neurosciences, The University of Western Australia, Medical Research Foundation Building, Perth WA 6000, Australia
[169]Center for Human Genetic Research and Department of Psychiatry, Massachusetts General Hospital, Boston, Massachusetts 02114, USA
[170]Department of Functional Genomics, Center for Neurogenomics and Cognitive Research, Neuroscience Campus Amsterdam, VU University, Amsterdam 1081, The Netherlands
[171]Department of Complex Trait Genetics, Neuroscience Campus Amsterdam, VU University Medical Center Amsterdam, Amsterdam 1081, The Netherlands

[172]Department of Child and Adolescent Psychiatry, Erasmus University Medical Centre, Rotterdam 3000, The Netherlands
[173]University of Aberdeen, Institute of Medical Sciences, Aberdeen, AB25 2ZD, UK
[174]Department of Clinical Medicine, University of Copenhagen, Copenhagen 2200, Denmark
[175]Department of Molecular Genetics and McLaughlin Centre, University of Toronto, Toronto, Ontario, Canada
[176]Department of Cellular and Molecular Medicine, University of California, San Diego, La Jolla, CA 92093, USA

Abstract

Genomic copy number variants (CNVs) have been strongly implicated in the etiology of schizophrenia (SCZ). However, apart from a small number of risk variants, elucidation of the CNV contribution to risk has been difficult due to the rarity of risk alleles, all occurring in less than 1% of cases. We sought to address this obstacle through a collaborative effort in which we applied a centralized analysis pipeline to a SCZ cohort of 21,094 cases and 20,227 controls. We observed a global enrichment of CNV burden in cases (OR=1.11, P=5.7x10^{-15}) which persisted after excluding loci implicated in previous studies (OR=1.07, P=1.7x10^{-6}). CNV burden is also enriched for genes associated with synaptic function (OR = 1.68, P = 2.8e-11) and neurobehavioral phenotypes in mouse (OR = 1.18, P= 7.3e-5). We identified genome-wide significant support for eight loci, including 1q21.1, 2p16.3 (NRXN1), 3q29, 7q11.2, 15q13.3, distal 16p11.2, proximal 16p11.2 and 22q11.2. We find support at a suggestive level for eight additional candidate susceptibility and protective loci, which consist predominantly of CNVs mediated by non-allelic homologous recombination (NAHR).

Introduction

Studies of genomic copy number variation (CNV) have established a role for rare genetic variants in the etiology of SCZ [1]. There are three lines of evidence that CNVs contribute to risk for SCZ: genome-wide enrichment of rare deletions and duplications in SCZ cases relative to controls [2,3], a higher rate of *de novo* CNVs in cases relative to controls[4-6], and association evidence implicating a small number of specific loci (**Extended data table 1**). All CNVs that have been implicated in SCZ are rare in the population, but confer significant risk (odds ratios 2-60).

To date, CNVs associated with SCZ have largely emerged from mergers of summary data for specific candidate loci [7-9]; yet even the largest genome-wide scans (sample sizes typically <10,000) remain under-powered to robustly confirm genetic association for the majority of pathogenic CNVs reported so far, particularly for those with low frequencies (<0.5% in cases) or intermediate effect sizes (odds ratios 2-10). It is important to address the low power of systematic CNV studies with larger samples given that this type of mutation has already proven useful for highlighting some aspects of SCZ related biology [6,10-13].

The limited statistical power provided by small samples is a significant obstacle in studies of rare and common genetic variation. In response, global collaborations have been formed in order to attain large sample sizes, as exemplified by the study of the Schizophrenia Working Group of the Psychiatric Genomics Consortium (PGC) in which 108 independent schizophrenia associated loci were identified [14]. Recognizing the need for similarly large samples in studies of CNVs for psychiatric disorders, we formed the PGC CNV Analysis Group. Our goal was to enable large-scale analyses of CNVs in psychiatry using centralized and uniform methodologies for CNV calling, quality control, and statistical analysis. Here, we report the largest genome-wide analysis of CNVs for any psychiatric disorder to date, using datasets assembled by the Schizophrenia Working Group of the PGC.

Data processing and meta-analytic methods

Raw intensity data were obtained from 57,577 subjects from 43 separate datasets (**Extended data table 2**). After CNV calling and quality control (QC), 41,321 subjects were retained for analysis. In large datasets derived from multiple studies, variability in CNV detection between studies and array platforms presents a significant challenge. To minimize the technical variability across different studies, we developed a centralized pipeline for systematic calling of CNVs for Affymetrix and Illumina platforms. (**Methods** and **Extended data figure 1**). The pipeline included multiple CNV callers run in parallel. Data from Illumina platforms were processed using PennCNV [15] and iPattern [16]. Data from Affymetrix platforms were analyzed using PennCNV and Birdsuite [17].Two additional methods, iPattern and C-score [18], were applied to data from the Affymetrix 6.0 platform. The CNV calls from each program were converted to a standardized format and a consensus call set was constructed by merging CNV outputs at the sample level. Only CNV segments that were detected by all algorithms were retained. We performed rigorous QC at the platform level to exclude samples with poor probe intensity and/or an excessive CNV load (number and length). Larger CNVs that appeared to be fragmented were merged and retained. CNVs spanning centromeres or those with >50% overlap with segmental duplications or regions prone to VDJ recombination (e.g.,

immunoglobulin or T cell receptor loci) were excluded. A final set of rare, high quality CNVs was defined as those >20kb in length, at least 10 probes, and <1% MAF.

Genetic associations were investigated by case-control tests of CNV burden at four levels: (1) genome-wide (2) pathways, (3) genes, and (4) probes. Analyses controlled for SNP-derived principal components, sex, genotyping platform, and individual-level probe intensity. Multiple-testing thresholds for genome-wide significance were estimated from family-wise error rates drawn from permutation

Genome wide analysis of CNV burden reveals an enrichment of ultra-rare variants

An elevated burden of rare CNVs has been well established among SCZ cases[2]. We applied our meta-analytic framework to measure the consistency of overall CNV burden across the genotyping platforms, and whether a measurable amount of CNV burden persists outside of previously implicated CNV regions. Consistent with previous estimates, the overall CNV burden is significantly greater among SCZ cases when measured as total Kb covered (OR=1.12, p = 5.7e-15), genes affected (OR=1.21, p = 6.6e-21), or CNV number (OR=1.03, p = 1e-3). Focusing on genes affected by CNV, our strongest signal of enrichment, the effect size is consistent across all genotyping platforms (Figure 1A). When we split by CNV type, the effect size for copy number losses (OR=1.40, p = 4e-16) is greater than for gains (OR=1.12, p = 2e-7) (**Extended data Figures 2-3**). Partitioning by CNV frequency (based on 50% reciprocal overlap with the full call set, **Methods**), CNV burden is enriched among cases across a range of frequencies, up to counts of 80 (MAF = 0.1%) in the combined sample (**Figure 1B**).

A primary question in this study is the contribution of novel loci to the excess CNV burden in cases. After removing nine previously implicated CNV loci (where reported *p*-values exceed our designated multiple testing threshold, **Extended data table 1**), excess CNV burden in SCZ remains significantly enriched (genes affected OR=1.11, p = 1.3e-7**, Extended data table 3**). CNV burden also remained significantly enriched after removal of all reported loci from **Extended data table 1**, but the effect-size was greatly reduced (OR = 1.08) compared to the enrichment overall (OR = 1.21). When we partition CNV burden by frequency, we find that

much of the previously unexplained signal is restricted to comparatively rare events (i.e., MAF < 0.1%, **Figure 1B**).

Gene-set (pathway) burden

We assessed whether CNV burden was concentrated within defined sets of genes involved in neurodevelopment or neurological function. A total of 36 gene-sets were evaluated (for a description see **Extended data table 3**), consisting of gene-sets representing neuronal function, synaptic components and neurological and neurodevelopmental phenotypes in human (19 sets), gene-sets based on brain expression patterns (7 sets), and human orthologs of mouse genes whose disruption causes phenotypic abnormalities, including neurobehavioral and nervous system abnormality (10 sets). Some gene-sets can be considered "negative controls", including genes not expressed in brain (1 set) or associated with abnormal phenotypes in mouse organ systems unrelated to brain (7 sets). We mapped CNVs to genes if they overlapped by at least one exonic bp.

Gene-set burden was tested using logistic regression deviance test [6]. In addition to using the same covariates included in genome-wide burden analysis, we controlled for the total number of genes per subject spanned by rare CNVs to account for signal that merely reflects the global enrichment of CNV burden in cases [19]. Multiple-testing correction (Benjamini-Hochberg False Discovery Rate, BH-FDR) was performed separately for each gene-set group and CNV type (gains, losses). After multiple test correction (Benjamini-Hochberg FDR ≤ 10%) 15 gene-sets were enriched for rare loss burden in cases and 4 for rare gains in cases, all of which are brain-related gene sets (**Figure 2**).

Of the 15 sets significant for losses, the majority consist of synaptic or other neuronal components (9 sets) from gene-set group (a); in particular, "GO synaptic" (GO:0045202) and "ARC complex" rank first based on statistical significance and effect-size respectively ("GO synaptic" deviance test p-value = 2.8e-11, "ARC complex" regression odds-ratio > 1.8, **Figure 2a**). Losses in cases were also significantly enriched for genes involved in nervous system or behavioral phenotypes in mouse but not for gene-sets related to other organ system phenotypes (**Figure 2c**). To account for dependency between synaptic and neuronal gene-sets,

we re-tested loss burden following a step-down logistic regression approach, ranking gene-sets based on significance or effect size (**Extended data table 4**). Only GO synaptic and ARC complex were significant in at least one of the two step-down analyses, suggesting that burden enrichment in the other neuronal categories is mostly accounted by the overlap with synaptic genes. Following the same approach, the mouse neurological/neurobehavioral phenotype set remained nominally significant, pointing to the existence of additional signal not captured by the synaptic set. Pathway enrichment was less pronounced for duplications, consistent with the smaller burden effects for this class of CNV. Duplication burden was significantly enriched for NMDA receptor complex, highly brain-expressed genes, medium/low brain-expressed genes and prenatally expressed brain genes (**Figure 2b**).

Given that synaptic gene sets were robustly enriched for deletions in cases, and with an appreciable contribution from loci that have not been strongly associated with SCZ previously, pathway-level interactions of these sets were further investigated. A protein-interaction network was seeded using the synaptic and ARC complex genes that were intersected by rare deletions in this study (**Figure 3**). A graph of the network highlights multiple subnetworks of synaptic proteins including pre-synaptic adhesion molecules (NRXN1, NRXN3), post-synaptic scaffolding proteins (DLG1, DLG2, DLGAP1, SHANK1, SHANK2), glutamatergic ionotropic receptors (GRID1, GRID2, GRIN1, GRIA4), and complexes such as Dystrophin and its synaptic interacting proteins (DMD, DTNB, SNTB1, UTRN). A subsequent test of the Dystrophin glycoprotein complex (DGC) revealed that deletion burden of the synaptic DGC proteins (intersection of “GO DGC” GO:0016010 and “GO synapse” GO:0045202) was enriched in cases (Deviance test P = 0.05), but deletion burden of the full DGC was not significant (P = 0.69).

Gene CNV burden

To define specific loci that confer risk for SCZ, we tested CNV burden at the level of individual genes, using logistic regression deviance test and the same covariates included in genome-wide burden analysis. To correctly account for large CNVs that affect multiple genes, we aggregated adjacent genes into single loci if their copy number was highly correlated across subjects. CNVs were mapped to genes if they overlapped one or more exons. The criterion for genome-wide

significance used the Family-Wise Error Rate (FWER) < 0.05. The criterion for suggestive evidence used a Benjamini-Hochberg False Discovery Rate (BH-FDR) < 0.05.

Of 18 independent CNV loci with gene-based BH-FDR < 0.05, two were excluded based on CNV calling accuracy or evidence of a batch effect (**Supplementary Information**). The sixteen loci that remained after these additional QC steps are listed in **Table 1**. P-values for this summary table were obtained by re-running our statistical model across the entire region (**Supplementary Results**). These 16 loci represent a set of novel (n=8) and previously implicated (n=8) loci. Manhattan plots of the gene association for losses and gains are provided in **Figure 4**. A permutation-based false discovery rate and yielded similar estimates to the BH-FDR.

Eight loci attain genome-wide significance, including copy number losses at 1q21.1, 2p16.3 (NRXN1), 3q29, 15q13.3, 16p11.2 (distal) and 22q11.2 along with gains at 7q11.23 and 16p11.2 (proximal). An additional eight loci meet criterion for suggestive association. Based on our estimation of False Discovery Rates (BH and permutations), we expect to observe less than two associations meeting suggestive criteria by chance.

Probe level CNV burden

With our current sample size and uniform CNV calling, many individual CNV loci can be tested with adequate power at the probe level, potentially facilitating discovery at a finer grain than locus-wide tests. Tests for association were performed at each CNV breakpoint using the residuals of case-control status after controlling for analysis covariates, with significance determined through permutation. Results for losses and gains are shown in **Extended data figure 4**. Four independent CNV loci surpass genome-wide significance, all of which were also identified in the gene-based test, including the 15q13.2-13.3 and 22q11.21 deletions, 16p11.2 duplication, and 1q21.1 deletion and duplication. While these loci represent less than half of the previously implicated SCZ loci, we do find support for all loci where the association originally reported meets the criteria for genome-wide correction in this study. We examined association among all previously reported loci showing association to SCZ, including 12 CNV losses and 20 CNV gains (**Extended data table 5**), and 14 of the 33 loci were associated with SCZ at $p < .05$.

When a probe-level test is applied, associations at some loci become better delineated. For instance, The *NRXN1* gene at 2p16.3 is a CNV hotspot, and exonic deletions of this gene are significantly enriched in SCZ[9,20]. In this large sample, we observe a high density of "non-recurrent" deletion breakpoints in cases and controls. The probe-level Manhattan plot reveals a saw tooth pattern of association, where peaks correspond to transcriptional start sites and exons of *NRXN1* (**Figure 5**). This example highlights how, with high diversity of alleles at a single locus, the association peak may become more refined, and in some cases converge toward individual functional elements. Similarly, a high density of duplication breakpoints at previously reported SCZ risk loci on 16p13.2 (http://bit.ly/1NPgIuq) and 8q11.23 (http://bit.ly/1PwdYTt) exhibit patterns of association that better delineate genes in these regions.

[the above URLs link to a PGC CNV browser display of the respective genomic regions. The browser can also be accessed directly at the following URL http://pgc.tcag.ca/gb2/gbrowse/pgc_hg18/]

Novel risk loci are predominantly NAHR-mediated CNVs

Many CNV loci that have been strongly implicated in human disease are hotspots for non-allelic homologous recombination (NAHR), a process which in most cases is mediated by flanking segmental duplications [21]. Consistent with the importance of NAHR in generating CNV risk alleles for schizophrenia, most of the loci in **Table 1** are flanked by segmental duplications. After excluding loci that have been implicated in previous studies, we investigated whether NAHR mutational mechanisms were also enriched among novel associated CNVs. We defined a CNV as "NAHR" when both the start and end breakpoint is located within a segmental duplication. Across all loci with FDR < 0.05 in the gene-base burden test, NAHR-mediated CNVs were significantly enriched, 6.03-fold (P=0.008; **Extended data figure 5**), when compared to a null distribution determined by randomizing the genomic positions of associated genes (**Supplemental Material**). These results suggest that novel SCZ CNVs tend to occur in regions prone to high rates of recurrent mutation.

Discussion

The present study of the PGC SCZ CNV dataset includes the majority of all microarray data that has been generated in genetic studies of SCZ to date. In this, the best body of evidence to date with which to evaluate CNV associations, we find definitive evidence for eight loci and we find significant evidence for a contribution from novel CNVs conferring both risk and protection. The complete results, including CNV calls and statistical evidence at the gene or probe level, can be viewed using the PGC CNV browser (URLs). Our data suggest that the novel risk loci that can be detected with current genotyping platforms lie at the ultra-rare end of the frequency spectrum and still larger samples will be needed to identify them at convincing levels of statistical evidence.

Collectively, the eight SCZ risk loci that surpass genome-wide significance are carried by a small fraction (1.4%) of SCZ cases in the PGC sample. We estimate 0.85% of the variance in SCZ liability is explained by carrying a CNV risk allele within these loci (**Supplementary Results**). As a comparison, 3.4% of the variance in SCZ liability is explained by the 108 genome-wide significant loci identified in the companion PGC GWAS analysis. Combined, the CNV and SNP loci that have been identified to date explain a small proportion (<5%) of heritability.

The large dataset here provides an opportunity to evaluate the strength of evidence for a variety of loci where an association with schizophrenia has been reported previously. Of 33 published findings from the recent literature, we find evidence for 14 loci ($P < 0.05$, **Extended data table 5**); thus, nearly half of the existing candidate loci are supported by our data. However we also find a lack of evidence for many. A lack of strong evidence in this dataset (which includes samples that overlap with many of the previous studies) may in some cases simply reflect that statistical power is limited for very rare variants, even in large samples. However, it is likely that some of these original findings represent spurious associations. Indeed, the loci that are not supported by our data consist largely of loci for which the original statistical evidence was modest (**Extended data table 5**). Thus, our results help to refine the list of promising candidate CNVs. Continued efforts to evaluate the growing number of candidate variants has considerable value for directing future research efforts focused on specific loci.

Novel candidate loci meeting suggestive criteria in this study highlight strong candidate loci that have not been previously implicated in SCZ. Two such associations are located on the X chromosome in a region of Xq28 that is highly prone to recurrent rearrangements [22-24] (**Extended data figure 6**). Gains at the distal Xq28 locus are enriched in cases in this study; similar duplications have been reported in association with intellectual disability, while reciprocal deletions of this region are associated with embryonic lethality in males [25]. Duplications at the proximal Xq28 locus, including a single gene *MAGEA11*, are enriched in controls in this study, and to our knowledge have not been documented in other disorders.

We observed multiple "protective" CNVs that showed a suggestive enrichment in controls, including duplications of 22q11.2, *MAGEA11,* and *ZMYM5* along with deletions and duplications of *ZNF92*. No protective effects were significant after genome-wide correction. Moreover, a rare CNV that confers reduced risk for SCZ may not confer a general protection from neurodevelopmental disorders. For example, microduplications of 22q11.2 appear to confer protection from SCZ [26]; however, such duplications have been shown to increase risk for developmental delay and a variety of congenital anomalies in pediatric clinical populations [27]. It is probable that some of the undiscovered rare alleles in SCZ are variants that confer protection but larger sample sizes are needed to determine this unequivocally. If true, our estimates of the excess CNV burden in cases may not fully account for the variation SCZ liability that is explained by rare CNVs.

Our results provide strong evidence that deletions in SCZ are enriched within a highly connected network of synaptic proteins, consistent with previous studies [2,6,10,28]. The large CNV dataset here allows a more detailed view of the synaptic network and highlights subsets of genes account for the excess deletion burden in SCZ, including synaptic cell adhesion and scaffolding proteins, glutamatergic ionotropic receptors and protein complexes such as the ARC complex and DGC. Modest CNV evidence implicating Dystrophin (DMD) and its binding partners is intriguing given that the involvement of certain components of the DGC have been postulated [29, 30] and disputed[31] previously. Larger studies of CNV are needed to define a role for this and other synaptic subnetworks in SCZ.

This study represents a milestone. Large-scale collaborations in psychiatric genetics have greatly advanced discovery through genome-wide association studies. Here we have extended this framework to rare CNVs. Our knowledge of the contribution from lower frequency variants gives us confidence that the application of this framework to large newly acquired datasets has the potential to further the discovery of loci and identification of the relevant genes and functional elements. The PGC CNV Resource is now publicly available through a custom browser at http://pgc.tcag.ca/gb2/gbrowse/pgc_hg18/.

Author Contributions

Management of the study, core analyses and content of the manuscript was the responsibility of the CNV Analysis Group chaired by J.S. and jointly supervised by S.W.S. and B.M.N. together with the Schizophrenia Working Group chaired by M.C.O'D. Core analyses were carried out by D.H., D.M., and C.R.M. Data Processing pipeline was implemented by C.R.M., B.T., W.W., D.G., M.G., A.S. and W.B. The A custom PGC CNV browser was developed by C.R.M and B.T. Additional analyses and interpretations were contributed by W.W., D.A and P.A.H. The individual studies or consortia contributing to the CNV meta-analysis were led by R.A.,O.A.A., D.H.R.B., A.D.B., E. Bramon, J.D.B., A.C., D.A.C., S.C., A.D., E. Domenici, H.E., T.E., P.V.G., M.G., H.G., C.M.H., N.I., A.V.J., E.G.J., K.S.K., G.K., J. Knight, T. Lencz, D.F.L., Q.S.L., J. Liu, A.K.M., S.A.M., A. McQuillin, J.L.M., P.B.M., B.J.M., M.M.N., M.C.O'D., R.A.O., M.J.O., A. Palotie, C.N.P., T.L.P., M.R., B.P.R., D.R., P.C.S, P. Sklar. D.St.C., P.F.S., D.R.W., J.R.W., J.T.R.W. and T.W. The remaining authors contributed to the recruitment, genotyping, or data processing for the contributing components of the meta-analysis. J.S., B.M.N, C.R.M, D.H., and D.M. drafted the manuscript which was shaped by the management group. All other authors saw, had the opportunity to comment on, and approved the final draft.

Competing Financial Interest

Several of the authors are employees of the following pharmaceutical companies: F.Hoffman-La Roche (E.D., L.E.), Eli Lilly (D.A.C., Y.M., L.N.) and Janssen (A.S., Q.S.L). None of these companies influenced the design of the study, the interpretation of the data, or the amount of data reported, or financially profit by publication of the results which are pre-competitive. The other authors declare no competing interests.

Acknowledgements

Core funding for the Psychiatric Genomics Consortium is from the US National Institute of Mental Health (U01 MH094421). We thank T. Lehner and Anjene Addington (NIMH). The work of the contributing groups was supported by numerous grants from governmental and charitable bodies as well as philanthropic donation. Details are provided in the Supplementary Notes. Membership of the Wellcome Trust Case Control Consortium and of the Psychosis Endophenotype International Consortium are provided in the Supplementary Notes.

URLs

PGC CNV browser, http://pgc.tcag.ca/gb2/gbrowse/pgc_hg18.

References

1. Malhotra, D. & Sebat, J. CNVs: harbingers of a rare variant revolution in psychiatric genetics. *Cell* **148**, 1223-41 (2012).
2. Walsh, T. *et al.* Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science* **320**, 539-43 (2008).
3. The International Schizophrenia, C. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature.* **455**, 237-241 (2008).
4. Malhotra, D. *et al.* High frequencies of de novo CNVs in bipolar disorder and schizophrenia. *Neuron* **72**, 951-63 (2011).
5. Xu, B. *et al.* Strong association of de novo copy number mutations with sporadic schizophrenia. *Nat Genet* **40**, 880-5 (2008).
6. Kirov, G. *et al.* De novo CNV analysis implicates specific abnormalities of postsynaptic signalling complexes in the pathogenesis of schizophrenia. *Molecular psychiatry* **17**, 142-53 (2012).
7. McCarthy, S.E. *et al.* Microduplications of 16p11.2 are associated with schizophrenia. *Nat Genet* **41**, 1223-7 (2009).
8. Mulle, J.G. *et al.* Microdeletions of 3q29 confer high risk for schizophrenia. *Am J Hum Genet* **87**, 229-36 (2010).
9. Rujescu, D. *et al.* Disruption of the neurexin 1 gene is associated with schizophrenia. *Hum Mol Genet* (2008).
10. Pocklington, A.J. *et al.* Novel Findings from CNVs Implicate Inhibitory and Excitatory Signaling Complexes in Schizophrenia. *Neuron* **86**, 1203-14 (2015).
11. Horev, G. *et al.* Dosage-dependent phenotypes in models of 16p11.2 lesions found in autism. *Proc Natl Acad Sci U S A* **108**, 17076-81 (2011).
12. Golzio, C. *et al.* KCTD13 is a major driver of mirrored neuroanatomical phenotypes of the 16p11.2 copy number variant. *Nature* **485**, 363-7 (2012).
13. Holmes, A.J. *et al.* Individual differences in amygdala-medial prefrontal anatomy link negative affect, impaired social functioning, and polygenic depression risk. *J Neurosci* **32**, 18087-100 (2012).
14. Schizophrenia Working Group of the Psychiatric Genomics, C. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421-7 (2014).
15. Wang, K. *et al.* PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res* **17**, 1665-74 (2007).
16. Pinto, D. *et al.* Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* **466**, 368-72 (2010).
17. Korn, J.M. *et al.* Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat.Genet.* **40**, 1253-1260 (2008).
18. Vacic, V. *et al.* Duplications of the neuropeptide receptor gene VIPR2 confer significant risk for schizophrenia. *Nature* **471**, 499-503 (2011).
19. Raychaudhuri, S. *et al.* Accurately assessing the risk of schizophrenia conferred by rare copy-number variation affecting genes with brain function. *PLoS Genet* **6**(2010).
20. Kirov, G. *et al.* Comparative genome hybridization suggests a role for NRXN1 and APBA2 in schizophrenia. *Hum Mol Genet* **17**, 458-65 (2008).
21. Lupski, J.R. Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends Genet* **14**, 417-22 (1998).

22. Calhoun, A.R. & Raymond, G.V. Distal Xq28 microdeletions: clarification of the spectrum of contiguous gene deletions involving ABCD1, BCAP31, and SLC6A8 with a new case and review of the literature. *Am J Med Genet A* **164A**, 2613-7 (2014).
23. El-Hattab, A.W. *et al.* Clinical characterization of int22h1/int22h2-mediated Xq28 duplication/deletion: new cases and literature review. *BMC Med Genet* **16**, 12 (2015).
24. Ravn, K. *et al.* Large genomic rearrangements in MECP2. *Hum Mutat* **25**, 324 (2005).
25. El-Hattab, A.W. *et al.* Int22h-1/int22h-2-mediated Xq28 rearrangements: intellectual disability associated with duplications and in utero male lethality with deletions. *J Med Genet* **48**, 840-50 (2011).
26. Rees, E. *et al.* Evidence that duplications of 22q11.2 protect against schizophrenia. *Mol Psychiatry* **19**, 37-40 (2014).
27. Van Campenhout, S. *et al.* Microduplication 22q11.2: a description of the clinical, developmental and behavioral characteristics during childhood. *Genet Couns* **23**, 135-48 (2012).
28. Fromer, M. *et al.* De novo mutations in schizophrenia implicate synaptic networks. *Nature* **506**, 179-84 (2014).
29. Zatz, M. *et al.* Cosegregation of schizophrenia with Becker muscular dystrophy: susceptibility locus for schizophrenia at Xp21 or an effect of the dystrophin gene in the brain? *J Med Genet* **30**, 131-4 (1993).
30. Straub, R.E. *et al.* Genetic variation in the 6p22.3 gene DTNBP1, the human ortholog of the mouse dysbindin gene, is associated with schizophrenia. *Am J Hum Genet* **71**, 337-48 (2002).
31. Mutsuddi, M. *et al.* Analysis of high-resolution HapMap of DTNBP1 (Dysbindin) suggests no consistency between reported common variant associations and schizophrenia. *Am J Hum Genet* **79**, 903-9 (2006).
32. Zuberi, K. *et al.* GeneMANIA prediction server 2013 update. *Nucleic Acids Res* **41**, W115-22 (2013).

Figure Legends

Figure 1. CNV Burden. (A) Forest plot of CNV burden (measured here as genes affected by CNV), partitioned by genotyping platform, with the full PGC sample at the bottom. CNV burden is calculated by combining CNV gains and losses. Case and control counts are listed, and "genes" is the rate of genes affected by CNV in controls. Burden tests use a logistic regression model predicting SCZ case/control status by CNV burden along with covariates (see methods). The odds ratio is the exponential of the logistic regression coefficient, and odds ratios above one predict increased SCZ risk. **(B)** CNV burden partitioned by CNV frequency. For reference, a CNV with MAF 0.1% in the PGC sample would have ~41 CNVs. Using the same model as above, each CNV was placed into a single CNV frequency category based on a 50% reciprocal overlap with other CNVs. CNV burden with inclusion of all CNVs are shown in green, whereas CNV burden excluding previously implicated CNV loci are shown in blue

Figure 2: Gene-set Burden

Gene-set burden test results for rare losses (a, c) and gains (b, d); frames a-b display gene-sets for neuronal function, synaptic components, neurological and neurodevelopmental phenotypes in human; frames c-d display gene-sets for human homologs of mouse genes implicated in abnormal phenotypes (organized by organ systems); both are sorted by –log 10 of the logistic regression deviance test p-value multiplied by the beta coefficient sign, obtained for rare losses when including known loci. Gene-sets passing the 10% BH-FDR threshold are marked with "*". Gene-sets representing brain expression patterns were omitted from the figure because only a few were significant (losses: 1, gains: 3).

Figure 3: Protein Interaction Network for Synaptic Genes

Synaptic and ARC-complex genes intersected by a rare loss in at least 4 case or control subjects and with genic burden Benjamini-Hochberg FDR <= 25% (red discs) were used to query GeneMANIA[32] and retrieve additional protein interaction neighbors, resulting in a network of 136 synaptic genes. Genes are depicted as disks; disk centers are colored based on rare loss frequency (Freq.SZ and Freq.CT) being prevalent in cases or controls; disk borders are colored

to mark (i) gene implication in human dominant or X-linked neurological or neurodevelopmental phenotype, (ii) de-novo mutation (DeN) reported by Fromer et al. [28], split between LOF (frameshift, stopgain, core splice site) and missense or amino acid insertion / deletion, (iii) implication in mouse neurobehavioral abnormality. Pre-synaptic adhesion molecules (NRXN1, NRXN3), post-synaptic scaffolds (DLG1, DLG2, DLGAP1, SHANK1, SHANK2) and glutamatergic ionotropic receptors (GRID1, GRID2, GRIN1, GRIA4) constitute a highly connected subnetwork with more losses in cases than controls.

Figure 4: Gene Based Manhattan.
A Manhattan plot displaying the –log10 nominal deviance p-value for the gene test. P-value cutoffs corresponding to FWER 0.05 and BH-FDR 5% are highlighted in red and blue, respectively. Loci significant after multiple test correction are labeled.

Figure 5: Manhattan plot of probe-level associations across the Neurexin-1 locus. Empirical P-values at each deletion breakpoint reveal a sawtooth pattern of association. Predominant peaks correspond to exons and transcriptional start sites of NRXN1 isoforms.

Methods

Overview

We assembled a CNV analysis group with members from Broad Institute, Children's Hospital of Philadelphia, University of Chicago, University of California San Diego, University of Michigan, University of North Carolina, Colorado University Boulder, and University of Toronto/SickKids Hospital. Our aim was to leverage the extensive expertise of the group to develop a fully automated centralized pipeline for consistent and systematic calling of CNVs for both Affymetrix and Illumina platforms. An overview of the analysis pipeline is shown in **Extended Data Figure 1**. After an initial data formatting step we constructed batches of samples for processing using four different methods, PennCNV, iPattern, C-score (GADA and HMMSeg) and Birdsuite for Affymetrix 6.0. For Affymetrix 5.0 data we used Birdsuite and PennCNV, for Affymetrix 500 we used PennCNV and C-score, and for all Illumina arrays we used PennCNV and iPattern. We then constructed a consensus CNV call dataset by merging data at the sample level and further filtered calls to make a final dataset **Extended data table 2**. Prior to any filtering, we processed raw genotype calls for a total of **57,577** individuals, including **28,684** SCZ cases and **28,893** controls.

Study Sample

A complete list of datasets that were included in the current study can be found in **Extended Data Table 2**. A more detailed description of the original studies can be found in a previous publication[1]

Copy Number Variant Analysis Pipeline Architecture and Sample Processing

All aspects of the CNV analysis pipeline were built on the Genetic Cluster Computer (GCC) in the Netherlands. PGC members sent external drives of raw data to the Netherlands for upload to the server as well as the corresponding sample metadata files.

Input Acceptance and Preprocessing: For Affymetrix we used the *.CEL files (all converted to the same format) as input, whereas for Illumina we required Genome or Beadstudio exported *.txt files with the following values: Sample ID, SNP Name, Chr, Position, Allele1 – Forward, Allele2 – Forward, X, Y, B Allele Freq and Log R Ratio. Samples were then partitioned into 'batches' to be run through each pipeline. For Affymetrix samples we created analysis batches based on the plate ID (if available) or genotyping date. Each batch had approximately 200 samples with an equal mix of male and female samples. Affymetrix Power Tools (APT - apt-copynumber-workflow) was

then used to calculate summary statistics about chips analyzed. Gender mismatches identified and excluded as were experiments with MAPD > 0.4. For Illumina data, we first determined the genome build and converted to hg18 if necessary and created analysis batches based on the plate ID or genotyping date. Each batch had approximately 200 samples, and equal mix of male and female samples.

Composite Pipeline: The composite pipeline comprises CNV callers PennCNV [2], iPattern [3], Birdsuite [4] and C-Score [5] organized into component pipelines. We used all four callers for Affymetrix 6.0 data, PennCNV and C-Score for Affymetrix 500, Probe annotation files were preprocessed for each platform. Once the array design files and probe annotation files were pre-processed, each individual pipeline component pipeline was run in two steps: 1) processing the intensity data by the core pipeline process to produce CNV calls, 2) parsing the specific output format of the core pipeline and converting the calls to a standard form designed to capture confidence scores, copy number states and other information computed by each pipeline

Merging of CNV data and Quality control filtering

Merging of CNV data: After standardization of outputs from each algorithm, CNV calls from each algorithm were merged at the sample level to increase specificity [3]. For CNVs generated from Affymetrix 6.0 array, we took the intersection of the four outputs (Birdsuite, iPattern, C-Score, PennCNV) at the sample level to create a consensus CNV. For the Affymetrix 500, Affymetrix 5.0, and Illumina platforms, CNV merging was performed by taking the intersection of the calls made by the two algorithms (PennCNV and C-Score for Affymetrix 500, Birdsuite and PennCNV for Affymetrix 5.0, and iPattern and PennCNV for Illumina) at the sample level. CNV calls that were made by only one of the algorithm were excluded. Calls discordant for type of CNV (gain or loss) were also excluded.

Quality control filtering: Following merging we applied filtering criteria for removal of arrays with excessive probe variance or GC bias and removal of samples with mismatches in gender or ethnicity or chromosomal aneuploidies. For Affymetrix we extracted the MAPD and waviness-sd from the APT summary file. We also calculated the proportion of each chromosome (excluding chrY) tagged as copy number variable and computed the number of CNV calls made for each sample. We then retained experiments if each of these measures was within 3 SD of the median. For Illumina data we extracted LRRSD, BAFSD, GCWF (waviness) from PennCNV log files. As with the Affymetrix data, we calculated the proportion of each chromosome (excluding chrY) tagged as copy number variable and computed the number of CNV calls made for each

sample. We retained samples if each of the above measures was within 3 SD of the median. For both Illumina and Affymetrix datasets, large CNVs that appeared artificially split were combined together if one of the methods detected a CNV spanning the gap. However, samples where > 10% of the chromosome was copy number variable were excluded as possible aneuploidies. Further, we excluded CNVs that: 1) spanned the centromere or overlapped the telomere (100 kb from the ends of the chromosome); 2) had > 50% of its length overlapping a segmental duplication; 3) had >50% overlap with immunoglobulin or T cell receptor. The final filtered CNV dataset was annotated with Refseq genes (transcriptions and exons). After this stage of quality control (QC), we had a total of **52,511** individuals, with **27,034** SCZ cases and **25,448** controls.

Filtering for rare CNVs: To make our final dataset of rare CNVs for all subsequent analysis we universally filtered out variants that present at >= 1% (50% reciprocal overlap) frequency in cases and controls combined. CNVs that overlapped > 50% with regions tagged as copy number polymorphic on any other platform were also excluded. CNVs < 20kb or having fewer than 10 probes were also excluded.

Post-CNV Calling QC

Overview: A number of steps were undertaken after CNV calling and initial filtering QC to minimize the impact of technical artifacts and potential confounds. In summary, we removed individuals not present in the PGC2 GWAS analysis [1], removed datasets with non-matching case or control samples that could not be reconciled using consensus platform probes, and removed any additional outliers with respect to overall CNV burden, CNV calling metrics, or SCZ phenotype residuals. All steps are described in more detail below.

Merging with GWAS cohort: By matching the unique sample identifiers, we retained only individuals that also passed QC filtering from the companion PGC GWAS study in Schizophrenia [1]. This step filtered out samples with low-quality SNP genotyping, related individuals, and repeated samples across cohorts. An additional benefit of the PGC analytical framework is the ability to account for population stratification across cohorts using principal components derived from probe level analysis. After the post-CNV calling quality control steps described below, we re-calculated principal components using the Eigenstrat software package [6]. Sample information and subsequent CNV and GWAS filtered sample sets are presented in **Extended data table 2**. In the process of matching to the GWAS-specific cohort, all individuals of non-European ancestry were removed from analysis (~5.8% of the post-QC sample comprising three separate datasets). We

also removed 42 samples that had discordant phenotype designations between the GWAS analysis and CNV genotype submission.

Individual dataset removal: Some datasets submitted to the PGC consisted of only case or control samples, affected trios, or recruited external samples as controls. This asymmetry in case-control ascertainment and genotyping can present serious biases for CNV analysis, as the sensitivity to detect CNV will vary considerably across genotyping platforms, as well as within dataset and genotyping batch. Unlike imputation protocols commonly used for SNP genotyping, there is no equivalent process to infer unmeasured probe intensity from nearby markers. We took a number of steps to identify and remove datasets that showed strong signs of case-control ascertainment or genotyping asymmetry:

1) Identify genotyping platforms where case-control ratio was not between 40-60%
2) Where possible, merge similar genotyping platforms using consensus probes prior to CNV-calling pipeline in order to improve case-control ratio.
3) Examine overall CNV burden and association peaks for spurious results
4) Remove datasets that remain problematic due to unusual CNV burden or multiple spurious CNV associations.

The genotyping platforms identified and processed are listed in **Extended data table 2**. We were able to combine the Illumina OmniExpress and Illumina OmniExpress plus Exome Chip platforms with success by removing probe content specific to the Exome chip platform. We removed the *caws* Affymetrix 500 datasets due to a number of strong CNV association peaks not seen in any other dataset. We also remove the *fii6* dataset due to a 2-fold CNV burden in cases relative to controls. In order to improve case-control balance, we had to remove the affected proband trio datasets (*boco*, *lacw*, and *lemu*) in the Illumina 610 platform, and the control-only *uclo* dataset in the Affymetrix 500 platform.

Individual sample removal: We re-analyzed CNV burden estimates in the reduced sample to flag any lingering outliers missed in the initial QC. We identified outliers for CNV count and Kb burden in the autosome (> 30 CNVs or 8 Mb, respectively) and in the X chromosome (> 10 CNVs or 5 Mb, respectively), removing an additional 15 individuals.

Genome-wide CNV intensity and quality measurements produced by CNV calling algorithms (i.e. “CNV metrics”) were examined for additional outliers and potential relationships with case-control status. Each CNV metric was re-examined across studies

to assess if any additional outliers were present. Only three outliers were removed as their mean B allele (or minor allele) frequency deviated significantly from 0.5. Many CNV metrics are auto-correlated, as they measure similar patterns of variation in the probe intensity. Thus, we focused on the main intensity metrics - median absolute pairwise difference (MAPD) for projects genotyped on the Affymetrix 6.0 platform, and Log R Ratio standard deviation (LRRSD) in all other genotyping platforms. Among Affymetrix 6.0 datasets, MAPD did not differ between in cases and controls (t=1.14, p = 0.25). However, among non-Affymetrix 6.0 datasets, LRRSD showed significant differences between cases and controls (t=-35.3, p < 2e-16), with controls having a higher standardized mean LRRSD (0.227) than cases (-0.199). To control for any spurious associations driven by CNV calling quality, we included LRRSD (MAPD for Affymetrix 6.0 platforms) as a covariate in downstream analysis. CNV metrics were normalized with their genotyping platform prior to inclusion in the combined dataset.

Regression of potential confounds on case-control ascertainment

The PGC cohorts are a combination of many datasets drawn from the US and Europe, and it is important to ensure that any bias in sample ascertainment does not drive spurious association to SCZ. In order to ensure the robustness of the analysis, we controlled for a number of covariates that could potential confound results. Burden and gene-set analyses included covariates in a logistic regression framework. Due to the number of tests run at probe level association, we employed a step-wise logistic regression approach to allow for the inclusion of covariates in our case-control association, which we term the *SCZ residual* phenotype.

Covariates include sex, genotyping platform, CNV metrics, and ancestry principal components derived from SNP genotypes on the same samples in a previous study[1]. We were unable to control for dataset or genotyping batch, as a subset of the contributing datasets are fully confounded with case/control status. CNV metric is normalized within genotyping platform prior to inclusion in the logistic model. Only principal components that showed a significant association to small CNV burden were used (small CNV being defined as autosomal CNV burden with CNV < 100 kb in size). Among the top 20 principal components, only the 1st, 2nd, 3rd, 4th, and 8th principal component showed association with small CNV burden (with p < 0.01 used as the significance cutoff). To calculate the SCZ residual phenotype, we first fit a logistic regression model of covariates to affection status, and then extracted the Pearson residual values for use in a quantitative association design for downstream analyses. Residual phenotype values in cases are all above zero, and controls below zero, and are graphed against overall kb burden in **Extended data figure 7**. We removed three individuals with an SCZ residual

phenotype greater than three (or negative three in controls). After the post-processing round of QC, we retained a dataset with a total of **41,321** individuals comprising **21,094** SCZ cases and **20,227** controls.

Identifying previously implicated CNV loci in the literature

To delineate CNV burden effects coming from CNV loci that have previously been reported as putative SCZ risk factors from CNV in remainder of the genome, we flagged CNV loci with $p < 0.01$ that have either been reviewed [7,8] or otherwise reported [8-10] as potential SCZ risk factors in the literature. Previously reported loci meeting inclusion are listed in **Extended data table 1**. While a number of CNV loci have been reported in multiple studies, we sought the most recent reports that incorporated the largest sample sizes. To identify putatively associated CNV loci with SCZ from the full list, we applied the genome-wide *p*-value cutoff of **8e-5**, derived from the Cochran-Mantel-Haenzel (CMH) test in the current probe-level analysis as the *p*-value cutoff for inclusion as SCZ implicated CNV loci. While the CMH test is not the primary probe-level test in the current PGC analysis, it corresponds more closely to the tests used in published reports. In all, nine independent CNV loci from published reports surpass genome-wide correction. All published CNV loci, even those excluded as an SCZ implicated regions, are examined in the probe-level association analysis.

CNV burden analysis

We analyzed the overall CNV burden in a variety of ways to discern which general properties of CNV are contributing to SCZ risk. Overall individual CNV burden was measured in 3 distinct ways – 1) Kb burden of CNVs, 2) Number of genes affected by CNVs, and 3) Number of CNVs. In particular, we only counted gene as affected when the CNV overlapped a coding exon. We also partitioned our analyses by CNV type, size, and frequency. CNV type is defined as copy number losses (or deletions), copy number gains (or duplications), and both copy number losses and gains. To assign a specific allele frequency to a CNV, we used the --cnv-freq-method2 command in PLINK, whereby the frequency is determined as the total number of CNV overlapping the target CNV segment by at least 50%. This method differs from other methods that assign CNV frequencies by genomic region, whereby a single CNV spanning multiple regions may be included in multiple frequency categories.

For **Figure 1**, and **Extended data figures 2 and 3**, we partitioned CNV burden by genotyping platform, and the abbreviations for each platform are expanded below:

A500: Affymetrix 500

I300: Illumina 300K
I600: Illumina 610K and Illumina 660W
A5.0: Affymetrix 5.0
A6.0: Affymetrix 6.0
omni: OmniExpress and OmniExpress plus Exome

Due to the small size of the Omni 2.5 array (28 cases and 10 controls), they were excluded from presentation in the figure, but are included in all burden analyses with the total PGC sample. Burden tests use a logistic regression framework with the inclusion of covariates detailed above. Using a logistic regression framework, we predicted SCZ status using CNV burden as an independent predictor variable, thus allowing us to get an accurate estimate of the unique contribution of CNV burden in a multiple regression framework. To gain insight into the proportion of CNV burden risk coming from loci outside of the previously implicated SCZ regions, we ran all burden analyses after removing CNV that overlapped previously implicated CNV boundaries by more than 10%.

CNV probe level association

Genome-wide interrogation of CNV signals was tested at each respective CNV. Probe level tests were examined at the start, end, and single base position after the end of the called CNV. Three categories of CNV were tested: CNV deletions, CNV duplications, and deletions and duplications together. All analyses were run using PLINK software [11].

We ran probe level association using the SCZ residual phenotype as a quantitative variable, with significance determined through permutation of phenotype residual labels. An additional z-scoring correction, explained below, is used to control for any extreme values in the SCZ residual phenotype and efficiently estimate two-sided empirical *p*-values for highly significant loci. To ensure against the potential loss of power from the inclusion of covariates, we also ran a single degree of freedom Cochran-Mantel-Haenzel (CMH) test stratified by genotyping platform, with a 2 (CNV carrier status) x 2 (phenotype status) x N (genotyping platform) contingency matrix. While the CMH test does not account for more subtle biases that could drive false positive signals, it is robust to signals driven by a single platform and allows for each CNV carrier to be treated equally. Loci the surpassed genome-wide correction in either test was followed up for further evaluation.

Z-score recalibration of empirical testing: Probe level association *p*-values from the SCZ residual phenotype were initially obtained by performing one million permutations at

each CNV position, whereby each permutation shuffles the SCZ residual phenotype among all samples, and retains the SCZ residual mean for CNV carriers and non-carriers. For extremely rare CNV, however, CNV carriers at the extreme ends of the SCZ residual phenotype can produce highly significant *p*-values. While we understand that such rare events are unable to surpass strict genome-wide correction, we wanted to retain all tests to help delineate the potential fine-scale architecture within a single region of association. To properly account for the increased variance when only a few individuals are tested, we applied an empirical Z-score correction to the CNV carrier mean. In order to get an empirical estimate of the variance for each test, we calculated the standard deviation of residual phenotype mean differences in CNV carriers and non-carriers from 5,000 permutations. Z-scores are calculated as the observed case-control mean difference divided by the empirical standard deviation, with corresponding *p*-values calculated from the standard normal distribution. Concordance of the initial empirical and z-score *p*-values are close to unity for association tests with six or more CNV, whereas Z-score *p*-values are more conservative among tests with less than six CNV. Furthermore, the Z-score method naturally provides an efficient manner to estimate highly significant empirical *p*-values that would involve hundreds of millions of permutations to achieve.

Genome-wide correction for multiple tests

Beyond identifying significant CNV at the probe level, we also estimated the genome-wide testing space for rare CNV analysis. With the large PGC cohort being called through a consistent pipeline, we saw an opportunity to characterize the null expectation of segregating and recurrent *de novo* rare CNV in populations of European ancestry. Accepted thresholds for significance among published risk CNV have been limited in scope, as accurate population estimates of rare CNV frequency and distribution across the genome require large representative samples.

Genome-wide significance thresholds were calculated using the 5% family-wise error rate from 5,000 permutations in both the SCZ residual phenotype and CMH test. Specifically, we selected the 95^{th} percentile of the minimum *p*-values obtained across permutations. Below are the genome-wide correction *p*-value thresholds determined in this manner:

SCZ residual phenotype FWER correction:
CNV losses and gains: 6.73e-6
CNV losses: 1.5e-5
CNV gains: 1.35e-5

CMH test FWER correction:
CNV losses and gains: 3.65e-5
CNV losses: 8.25e-5
CNV gains: 7.8e-5

This method differs slightly from those used in **Levinson et al.** [9] to estimate the multiple test correction for rare CNV, however their genome-wide correction of p = 1e-5 corresponds quite closely to the estimates observed using the SCZ residual phenotype. The observed family-wise correction serves as good approximation of the independent rare CNV signals found among European ancestry populations for array-based CNV capture, but as sample sizes increase, so too will the effective number of tests, necessitating further evaluation of the multiple testing burden.

Gene-set burden enrichment analysis: gene-sets
Gene-sets with an a priori expectation of association to neuropsychiatric disorders were compiled based on gene annotations (Gene Ontology and curated pathway databases, downloaded June 2013) and published article materials (for details, see **Extended Data Table 3**). Gene-sets based on brain expression were compiled by processing the BrainSpan RNA-seq gene expression data-set (http://www.brainspan.org/static/download.html, downloaded Sept 2012). Four roughly equally sized gene-sets (about 4600 genes each) were derived to represent four expression tiers (very high, medium-to-high, medium-to-low, very low or absent); genes were selected if they passed a fixed expression threshold in at least 5/508 experimental data points (corresponding to different regions of donor brains, different donor ages corresponding to different developmental brain stages, and different donor sexes). Gene-sets based on mouse phenotypes were assembled by downloading MPO (Mammalian Phenotype Ontology) annotations from MGI (www.informatics.jax.org, downloaded August 2013), up-propagating annotations following ontology relations, and mapping to human orthologs using NCBI Homologene (www.ncbi.nlm.nih.gov/homologene); finally, top-level organ systems with fewer genes were aggregated while striving to preserve biological homogeneity, so to have roughly equal-sized sets (2,600-1,300 genes). For all gene-sets, gene identifiers in the primary source were mapped to Entrez-gene identifiers using the R/Bioconductor package *org.Hs.eg.db*.

Gene-set burden enrichment analysis: pre-processing

Subjects were restricted to the ones with at least one rare CNV. For copy number gains and losses, we separately calculated the following subject-level totals: variant number, variant length and number of genes impacted; these covariates are then used to model global burden and correct gene-set burden to ensure it is specific (i.e. not a mere reflection of genome-wide burden with some stochastic deviation due to sampling). The subject-level total number of genes impacted was also calculated for each gene-set, again separately for gains and losses. Subjects were flagged if they carried at least one CNV matching a locus previously implicated in schizophrenia (see section "Identifying previously implicated CNV loci in the literature"); this was then used to analyzed gene-set burden for all subjects, or excluding subjects with an already implicated CNV.

Gene-set burden enrichment analysis: statistical test

For each gene-set, we fit the following logistic regression model (as implemented by the R function *glm* of the *stats* package), where subjects are statistical sampling units:

y ~ covariates + global + gene-set

Where:

- *y* is the dicotomic outcome variable (schizophrenia = 1, control = 0)
- *covariates* is the set of variables used as covariates also in the genome-wide burden and probe association analysis (sex, genotyping platform, CNV metric, and CNV associated principal components)
- global is the measure of global burden; for the results in the main text, we used the total gene number (abbreviated as *U* from universe gene-set count); we also calculated results for total length (abbreviated as *TL*) and variant number plus variant mean length (abbreviated as *CNML*)
- *gene-set* is the gene-set gene count

The gene-set burden enrichment was assessed by performing a chi-square deviance test (as implemented by the R function *anova.glm* of the *stats* package) comparing these two regression models:

y ~ covariates + global

y ~ covariates + global + gene-set

We reported the following statistics:

- coefficient beta estimate (abbreviated as *Coeff*)
- t-student distribution-based coefficient significance p-value (as implemented by the R function *summary.glm* of the *stats* package, abbreviated as *Pvalue_glm*)
- deviance test p-value (abbreviated as *Pvalue_dev*)
- gene-set size (i.e. number of genes is the gene-set, regardless of CNV data)
- BH-FDR (Benjamini-Hochberg False Discovery rate)

- percentage of schizophrenia and control subjects with at least 1 gene, 2 genes, etc… impacted by a CNV of the desired type (loss or gain) in the gene-set (abbreviated as *SZ_g1n, SZ_g2n, … CT_g1n, …*)

Please note that, by performing simple simulation analyses, we realized that *Pvalue_glm* can be extremely over-conservative in presence of very few gene-set counts different than 0, while *Pvalue_dev* tends to be slightly under-conservative. While the two p-values tend to agree well for gene-set analysis, *Pvalue_glm* is systematically over-conservative for gene analysis since smaller counts are typically available for single genes.

Gene burden analysis: pre-processing

Subjects were restricted to the ones with at least one rare CNV. Only genes with at least a minimum number of subjects impacted by CNV were tested; this threshold was picked by comparing the BH-FDR to the permutation-based FDR and ensuring limited FDR inflation (permuted FDR < 1.65 * BH-FDR at BH-FDR threshold = 5%) while maximizing power. For gains the threshold was set to 12 counts, while for losses it was set to 8 counts.

Gene burden analysis: statistical test

For each gene, we fit the following logistic regression model (as implemented by the R function *glm* of the *stats* package), where subjects are statistical sampling units:

y ~ covariates + gene

Where:

- *y* is the dichotomous outcome variable (schizophrenia = 1, control = 0)
- *covariates* is the set of variables used as covariates also in the genome-wide burden and probe association analysis (sex, genotyping platform, CNV metric, and CNV associated principal components)
- *gene* is the binary indicator for the subject having or not having a CNV of the desired type (loss or gain) mapped to the gene

The gene burden was assessed by performing a chi-square deviance test (as implemented by the R function *anova.glm* of the *stats* package) comparing these two regression models:

- *y ~ covariates*
- *y ~ covariates + gene*

Gene burden analysis: multiple test correction

Multiple test correction was performed for loci rather than for genes, to avoid the strong correlation between test introduced by multi-genic CNVs; for the same reason, it

is more useful to count false positives as loci rather than genes. We followed a greedy step-down procedure:

- start from gene with most significant deviance p-value G1, create locus L1
- remove from the gene list all genes that share at least 50% of their carrier subjects with G1, and add them to locus L1
- do the same for the next gene most significant gene in the list (thus creating a new locus L2), and proceed recursively until there is no gene left
- define locus p-value as the smallest deviance p-value of its genes

We computed permutation-based FDR by permuting subjects' condition labels (schizophrenia, control), but not covariates (as those are expected to correlate to CNV distribution), 1,000 times. The FDR was then defined as the ratio between the average number of tests passing a given p-value threshold across the 1,000 permutations and the number of tests passing the same p-value threshold for real data. FDRs were also generated counting only the subset of genes with positive and negative regression coefficients (i.e. risk and presumed protective). The p-value threshold for permutation-based FDR calculation was picked by choosing the maximum nominal p-value corresponding to a given BH-FDR threshold (e.g. 5%). BH-FDR is supposed to be slightly inflated because (i) the deviance test p-value is slightly under-conservative in presence of very few gene indicators different than 0, (ii) we use the smallest gene p-value to define the locus p-value.

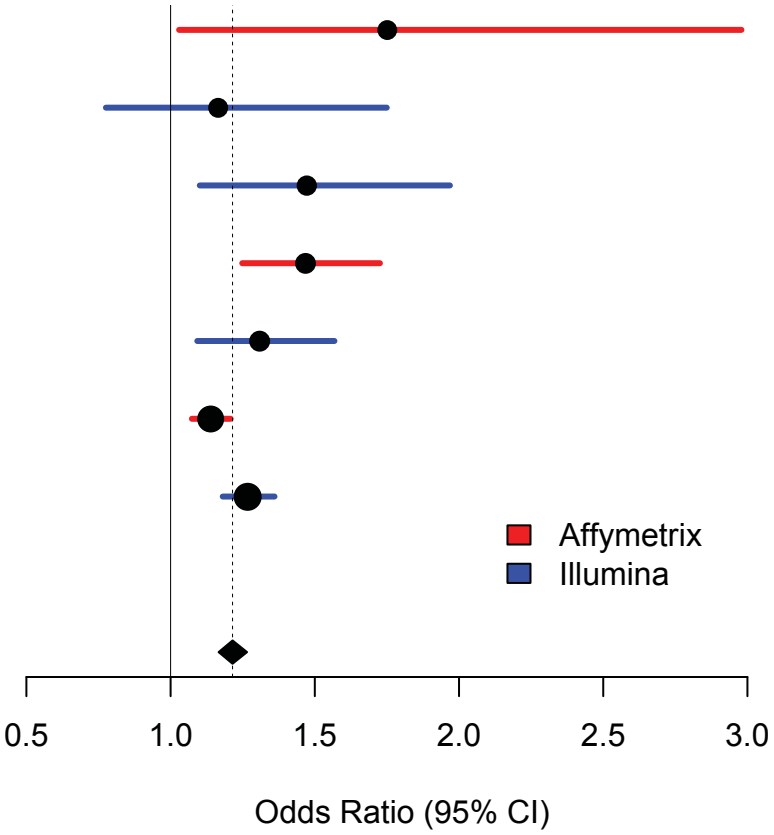
References

1. Schizophrenia Working Group of the Psychiatric Genomics, C. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421-7 (2014).
2. Wang, K. *et al.* PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res* **17**, 1665-74 (2007).
3. Pinto, D. *et al.* Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* **466**, 368-72 (2010).
4. Korn, J.M. *et al.* Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat.Genet.* **40**, 1253-1260 (2008).
5. McCarthy, S.E. *et al.* Microduplications of 16p11.2 are associated with schizophrenia. *Nat Genet* **41**, 1223-7 (2009).
6. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904-909 (2006).
7. Malhotra, D. & Sebat, J. CNVs: harbingers of a rare variant revolution in psychiatric genetics. *Cell* **148**, 1223-41 (2012).

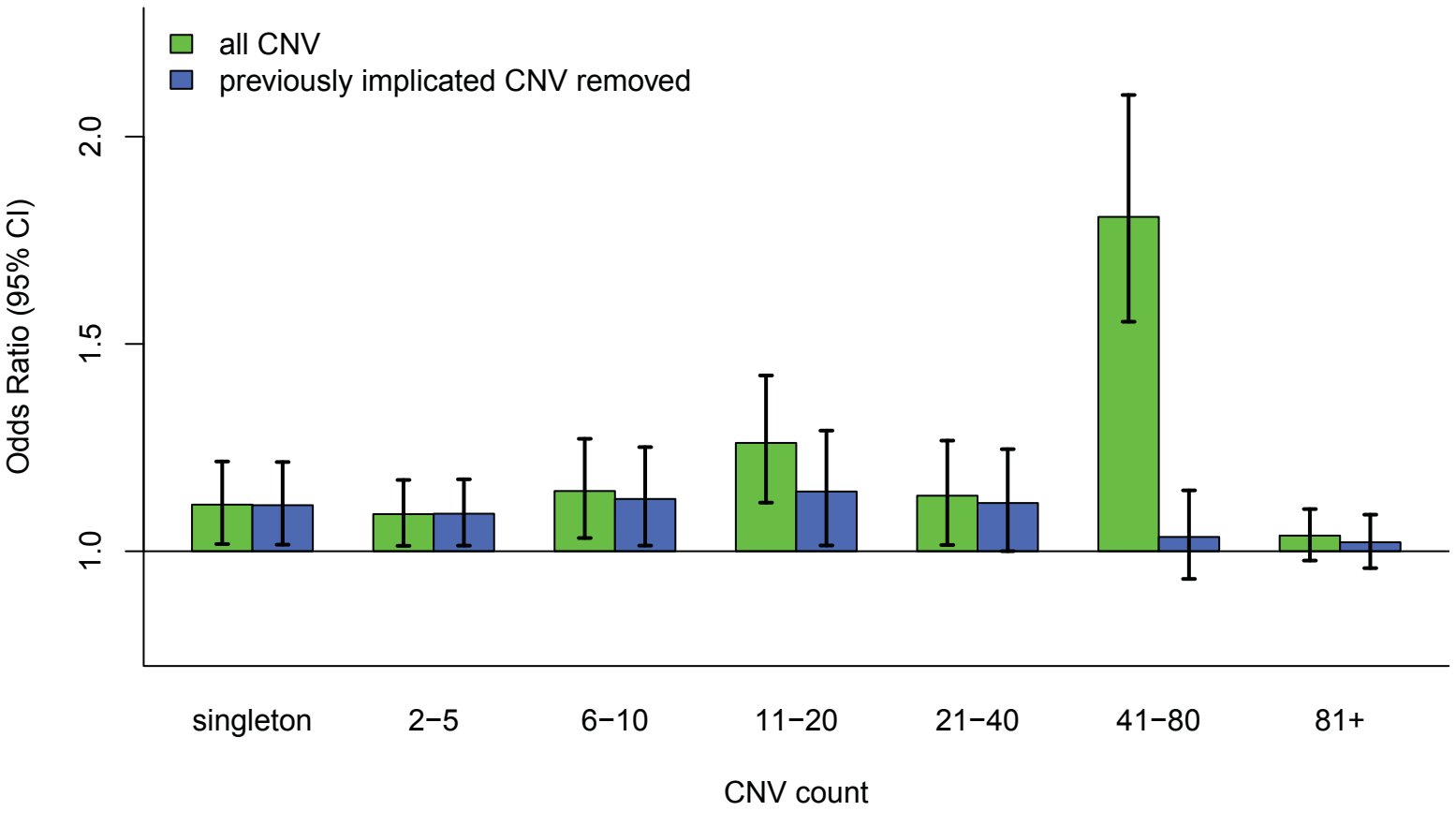
8. Rees, E. *et al.* Analysis of copy number variations at 15 schizophrenia-associated loci. *Br J Psychiatry* **204**, 108-14 (2014).
9. Levinson, D.F. *et al.* Copy number variants in schizophrenia: confirmation of five previous findings and new evidence for 3q29 microdeletions and VIPR2 duplications. *Am J Psychiatry* **168**, 302-16 (2011).
10. Bergen, S.E. *et al.* Genome-wide association study in a Swedish population yields support for greater CNV and MHC involvement in schizophrenia compared to bipolar disorder. *Molecular Psychiatry* **17**, 880-6 (2012).
11. Purcell, S. *et al.* PLINK: a toolset for whole-genome association and population-based linkage analysis. *American Journal of Human Genetics* **81**, 559-75 (2007).

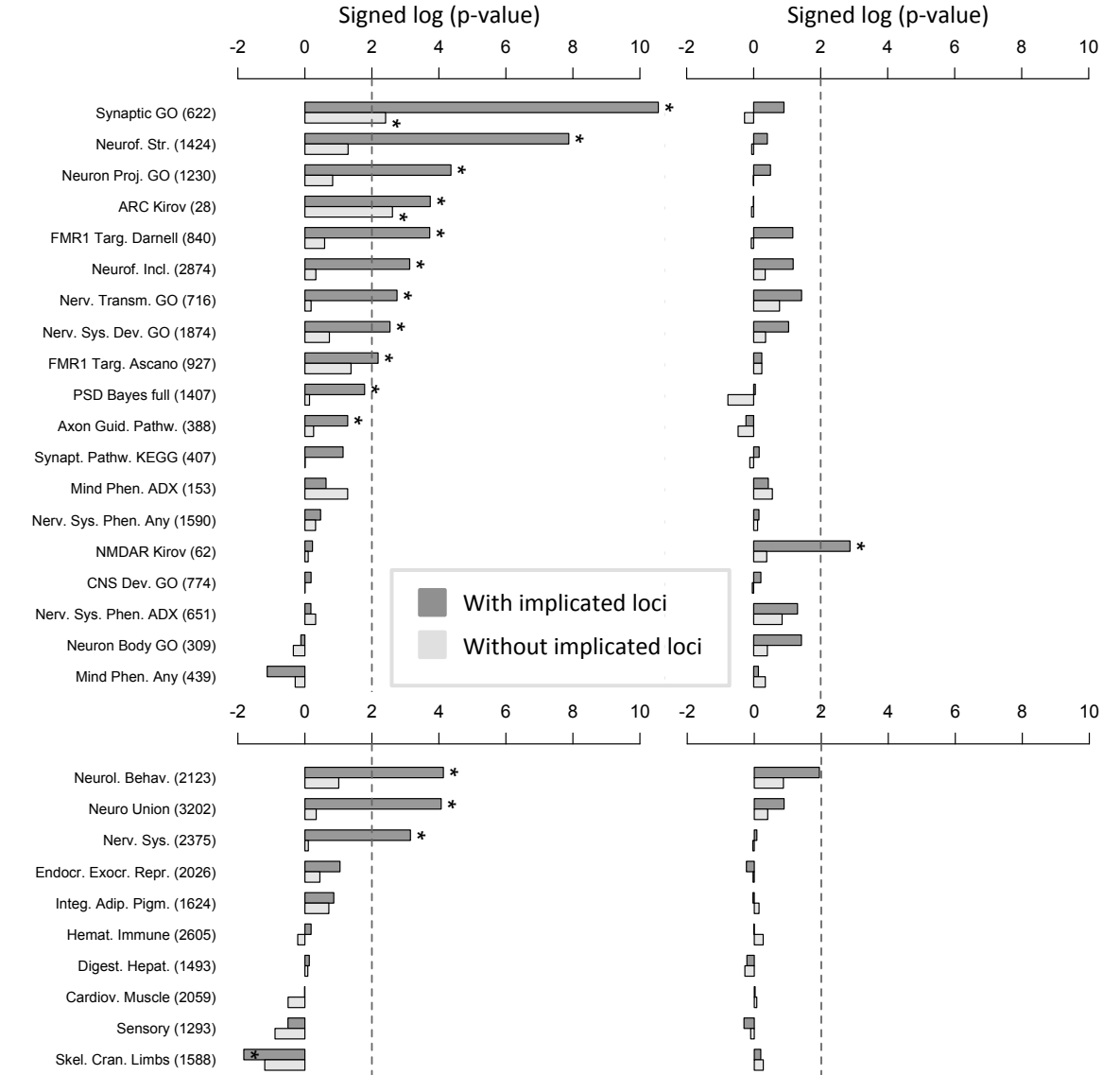
A

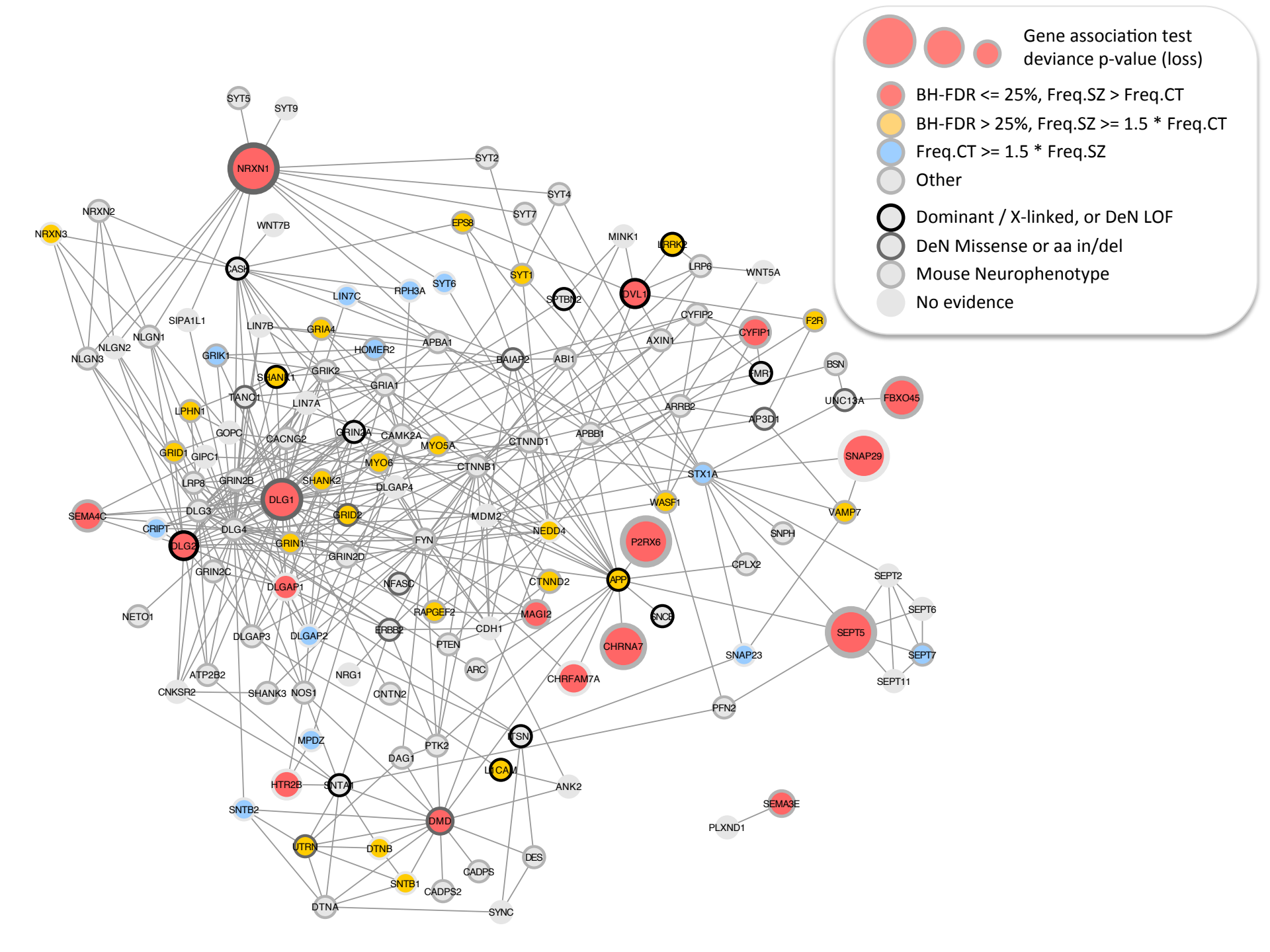
Platform	Cases	Controls	Genes	pval	OR
A500	314	201	1.7	0.039	1.75
I300	410	285	2	0.461	1.16
I600	818	623	1.3	0.009	1.47
A5.0	1130	1001	1.7	3.6e−06	1.47
I550	1572	1705	1.9	0.004	1.31
A6.0	7027	7980	2.4	1.6e−05	1.14
omni	9795	8422	2.1	5.8e−11	1.27
PGC	**21094**	**20227**	**2.2**	**6.6e−21**	**1.21**

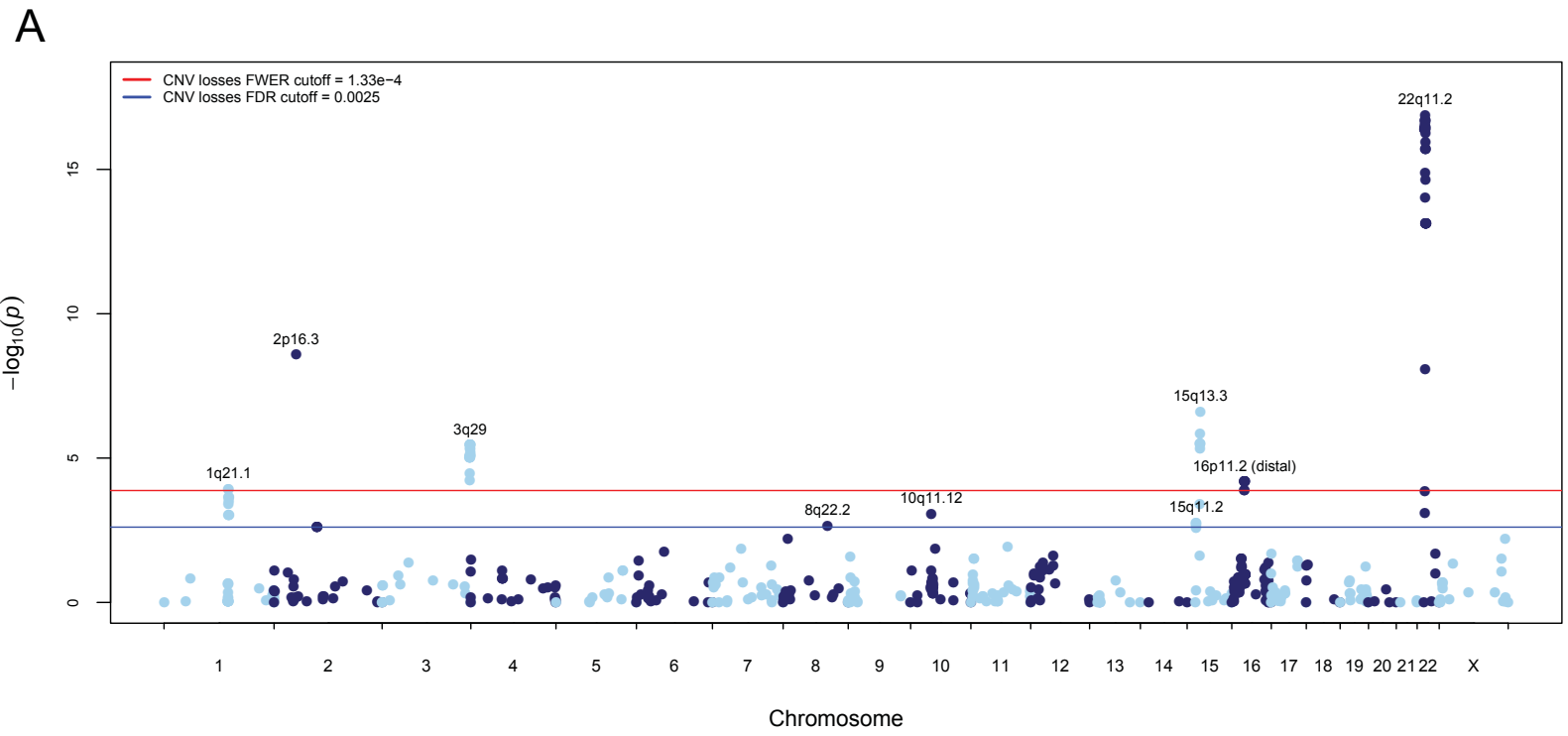


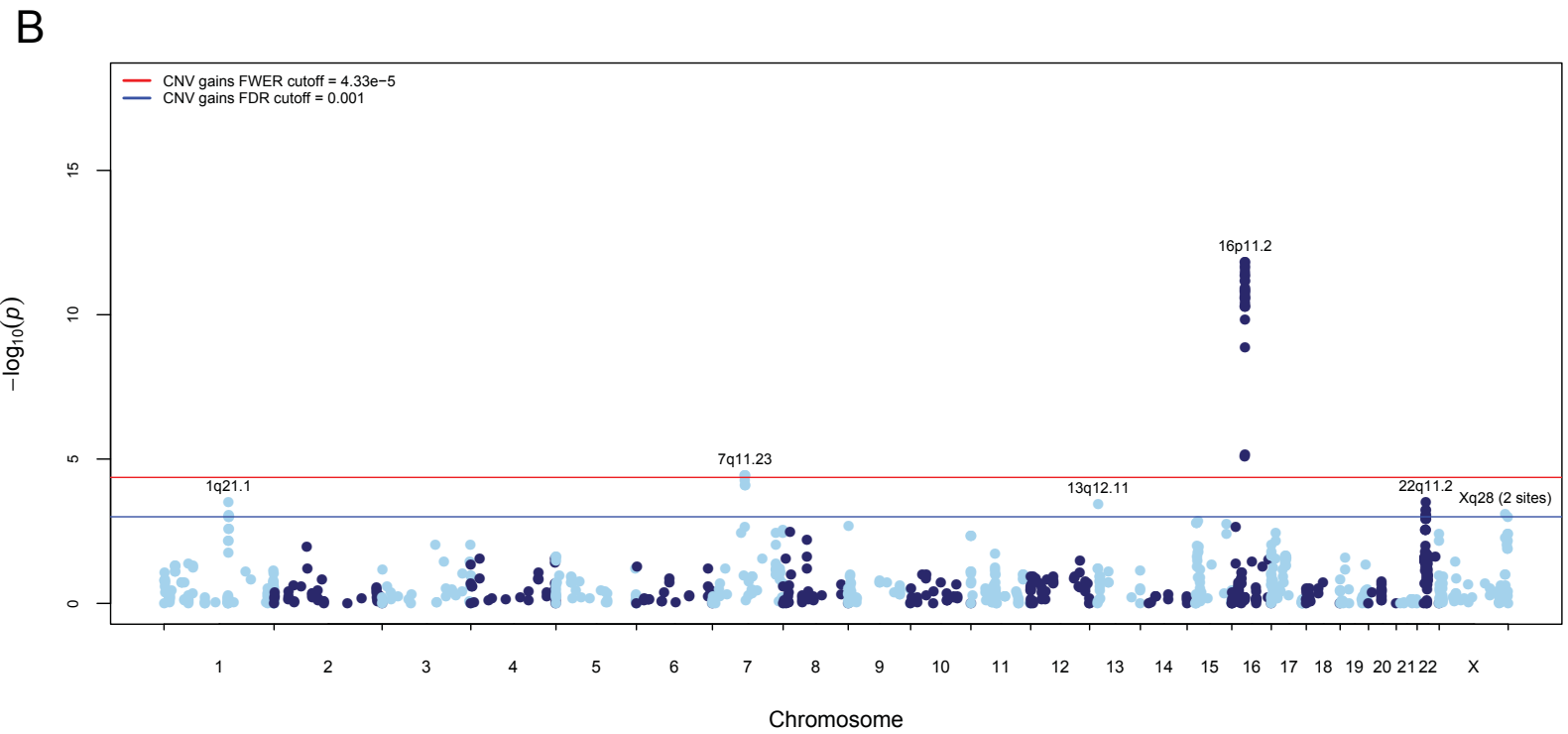
B

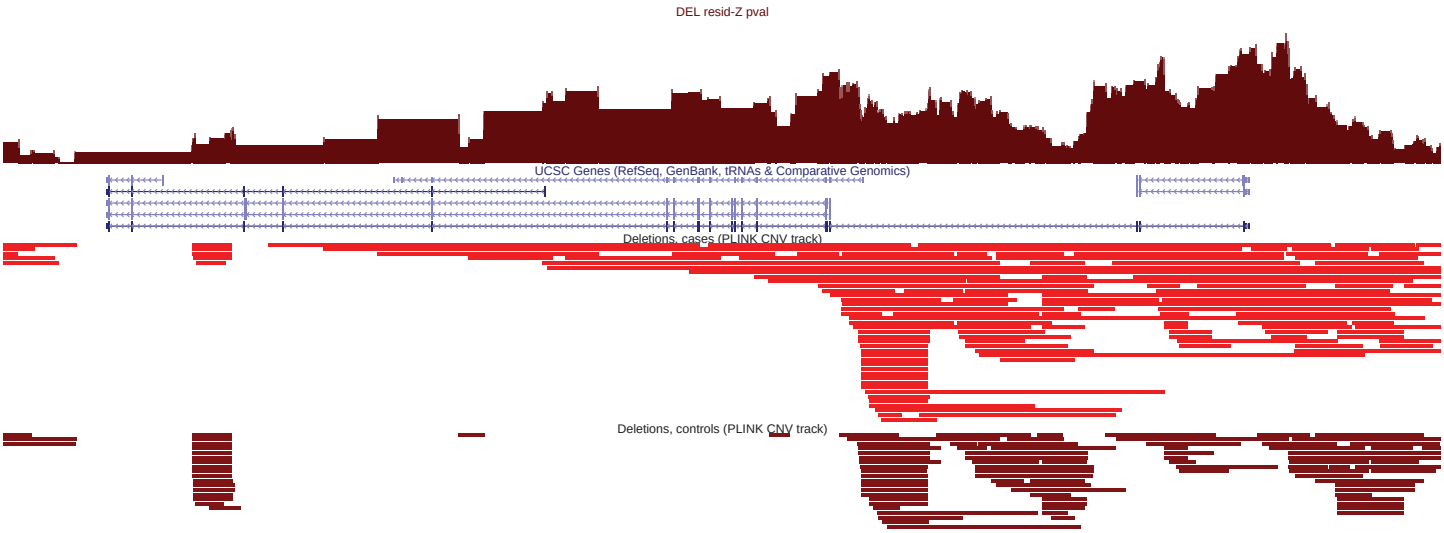



Signed log (p-value)
Signed log (p-value)
-2
0
2
4
6
8
10
Synaptic GO (622)
Neurof. Str. (1424)
Neuron Proj. GO (1230)
ARC Kirov (28)
FMR1 Targ. Darnell (840)
Neurof. Incl. (2874)
Nerv. Transm. GO (716)
Nerv. Sys. Dev. GO (1874)
FMR1 Targ. Ascano (927)
PSD Bayes full (1407)
Axon Guid. Pathw. (388)
Synapt. Pathw. KEGG (407)
Mind Phen. ADX (153)
Nerv. Sys. Phen. Any (1590)
NMDAR Kirov (62)
CNS Dev. GO (774)
Nerv. Sys. Phen. ADX (651)
Neuron Body GO (309)
Mind Phen. Any (439)
With implicated loci
Without implicated loci
Neurol. Behav. (2123)
Neuro Union (3202)
Nerv. Sys. (2375)
Endocr. Exocr. Repr. (2026)
Integ. Adip. Pigm. (1624)
Hemat. Immune (2605)
Digest. Hepat. (1493)
Cardiov. Muscle (2059)
Sensory (1293)
Skel. Cran. Limbs (1588)


Gene association test deviance p-value (loss)
BH-FDR <= 25%, Freq.SZ > Freq.CT
BH-FDR > 25%, Freq.SZ >= 1.5 * Freq.CT
Freq.CT >= 1.5 * Freq.SZ
Other
Dominant / X-linked, or DeN LOF
DeN Missense or aa in/del
Mouse Neurophenotype
No evidence
SYT5
SYT9
NRXN1
SYT2
SYT4
NRXN2
SYT7
NRXN3
WNT7B
EPS8
MINK1
LRRK2
CASK
LRP6
WNT5A
SYT1
RPH3A
SYT6
LIN7C
SPTBN2
DVL1
SIPA1L1
CYFIP2
F2R
LIN7B
GRIA4
NLGN1
APBA1
AXIN1
CYFIP1
NLGN2
HOMER2
NLGN3
GRIK1
BAIAP2
ABI1
FMR1
BSN
SHANK1
GRIK2
GRIA1
TANC1
LIN7A
UNC13A
FBXO45
LPHN1
ARRB2
AP3D1
GOPC
CACNG2
GRIN2A
CAMK2A
MYO5A
CTNND1
APBB1
SNAP29
GRID1
GIPC1
MYO6
CTNNB1
STX1A
LRP8
GRIN2B
SHANK2
DLGAP4
DLG1
WASF1
VAMP7
SEMA4C
DLG3
GRID2
MDM2
CRIPT
DLG4
NEDD4
SNPH
P2RX6
DLG2
GRIN1
FYN
GRIN2D
CPLX2
GRIN2C
SEPT2
APP
NFASC
CTNND2
DLGAP1
SEPT6
RAPGEF2
MAGI2
SNCB
NETO1
SEPT5
ERBB2
CDH1
DLGAP3
DLGAP2
PTEN
CHRNA7
SNAP23
SEPT7
ATP2B2
ARC
NRG1
CHRFAM7A
SEPT11
CNKSR2
SHANK3
NOS1
CNTN2
PFN2
ITSN1
MPDZ
PTK2
DAG1
L1CAM
HTR2B
SNTA1
ANK2
SEMA3E
SNTB2
DMD
PLXND1
UTRN
DTNB
DES
CADPS
SNTB1
CADPS2
DTNA
SYNC


A
CNV losses FWER cutoff = 1.33e−4
CNV losses FDR cutoff = 0.0025
22q11.2
2p16.3
15q13.3
3q29
16p11.2 (distal)
1q21.1
8q22.2
10q11.12
15q11.2
−log10(p)
0
5
10
15
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X
Chromosome


B
CNV gains FWER cutoff = 4.33e−5
CNV gains FDR cutoff = 0.001
16p11.2
7q11.23
1q21.1
13q12.11
22q11.2
Xq28 (2 sites)
−log10(p)
0
5
10
15
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X
Chromosome


DEL resid-Z pval
UCSC Genes (RefSeq, GenBank, tRNAs & Comparative Genomics)
Deletions, cases (PLINK CNV track)
Deletions, controls (PLINK CNV track)

CHR	BP1	BP2	locus GENE	Putative CNV Mechanism	CNV test	Direction	FWER	BH-FDR	Cases	Controls	Regional P-value	Odds Ratio [95% CI]
22	17,400,000	19,750,000	22q11.21	NAHR	loss	risk	yes	3.54E-15	64	1	5.70E-18	67.7 [9.3-492.8]
16	29,560,000	30,110,000	16p11.2 (proximal)	NAHR	gain	risk	yes	5.82E-10	70	7	2.52E-12	9.4 [4.2-20.9]
2	50,000,992	51,113,178	2p16.3 **NRXN1**	NHEJ	loss	risk	yes	3.52E-07	35	3	4.92E-09	14.4 [4.2-46.9]
15	28,920,000	30,270,000	15q13.3	NAHR	loss	risk	yes	2.22E-05	28	2	2.13E-07	15.6 [3.7-66.5]
1	144,646,000	146,176,000	1q21.1	NAHR	loss+gain	risk	yes	0.00011	60	14	1.50E-06	3.8 [2.1-6.9]
3	197,230,000	198,840,000	3q29	NAHR	loss	risk	yes	0.00024	16	0	1.86E-06	INF
16	28,730,000	28,960,000	16p11.2 (distal)	NAHR	loss	risk	yes	0.0029	11	1	5.52E-05	20.6 [2.6-162.2]
7	72,380,000	73,780,000	7q11.23	NAHR	gain	risk	yes	0.0048	16	1	1.68E-04	16.1 [3.1-125.7]
X	153,800,000	154,225,000	Xq28 (distal)	NAHR	gain	risk	no	0.049	18	2	3.61E-04	8.9 [2.0-39.9]
22	17,400,000	19,750,000	22q11.21	NAHR	gain	protective	no	0.024	3	16	4.54E-04	0.15 [0.04-0.52]
7	64,476,203	64,503,433	7q11.21 **ZNF92**	NAHR	loss+gain	protective	no	0.033	131	180	6.71E-04	0.66 [0.52-0.84]
13	19,309,593	19,335,773	13q12.11 **ZMYM5**	NHAR	gain	protective	no	0.024	15	38	7.91E-04	0.36 [0.19-0.67]
X	148,575,477	148,580,720	Xq28 **MAGEA11**	NAHR	gain	protective	no	0.044	12	36	1.06E-03	0.35 [0.18-0.68]
15	20,350,000	20,640,000	15q11.2	NAHR	loss	risk	no	0.044	98	50	1.34E-03	1.8 [1.2-2.6]
9	831,690	959,090	9p24.3 **DMRT1**	NHEJ	loss+gain	risk	no	0.049	13	1	1.35E-03	12.4 [1.6-98.1]
8	100,094,670	100,958,984	8q22.2 **VPS13B**	NHEJ	loss	risk	no	0.048	7	1	1.74E-03	14.5 [1.7-122.2]
7	158,145,959	158,664,998	7p36.3 **VIPR2 WDR60**	NAHR	loss+gain	risk	no	0.046	20	6	5.79E-03	3.5 [1.3-9.0]