nature medicine

BRIEF COMMUNICATION

https://doi.org/10.1038/s41591-020-1007-0

A multidimensional precision medicine approach identifies an autism subtype characterized by dyslipidemia

Yuan Luo[1,2,3,4,10], Alal Eran[5,6,7,10], Nathan Palmer[6], Paul Avillach[6], Ami Levy-Moonshine[8], Peter Szolovits[9] and Isaac S. Kohane[6]✉

The promise of precision medicine lies in data diversity. More than the sheer size of biomedical data, it is the layering of multiple data modalities, offering complementary perspectives, that is thought to enable the identification of patient subgroups with shared pathophysiology. In the present study, we use autism to test this notion. By combining healthcare claims, electronic health records, familial whole-exome sequences and neurodevelopmental gene expression patterns, we identified a subgroup of patients with dyslipidemia-associated autism.

The National Academy of Sciences' Precision Medicine Report proposes that 'when multiple molecular indicators are used in combination with conventional clinical, histological, and laboratory findings, they offer the opportunity for a more accurate and precise description and classification of disease'[1]. We tested whether this proposed multimodal precision medicine approach could identify distinct disease subtypes in large clinical, genomic and transcriptomic data. We used autism spectrum disorder (ASD) as a test case, in light of its extreme complexity and societal impact. ASD is now estimated to affect 1 in 54 children in the USA, 80% of whom are boys[2]. This behaviorally defined set of neurodevelopmental disorders currently lacks effective treatment and is debilitating for many families[3]. Recent genomic studies show that the clinical heterogeneity of ASD is matched by extreme genetic heterogeneity[4]. Dissecting the patient subgroups and matched molecular networks that underlie such complexity is essential to enable accurate early diagnosis, improve outcomes and ultimately facilitate precision medicine approaches to ASD[5,6], as exemplified in oncology[7–9].

Toward this goal, we integrated large datasets of familial whole-exome sequences (WESs), neurodevelopmental expression patterns, electronic health records (EHRs) and healthcare claims (Extended Data Fig. 1 and Methods). We identified variants of interest by examining clusters of neurodevelopmentally co-regulated, sex-differentially expressed, ASD-segregating deleterious variations, consistent with our current understanding of the complexity, origin and epidemiology of ASD[3]. Specifically, we first obtained spatiotemporal expression data from typically developing human brains using 524 samples from 26 brain regions of 42 subjects (23 males, 19 females) of the BrainSpan Atlas of the Developing Human Brain. As ASD is thought to arise during prenatal brain development and be driven by multiple genomic variants within each individual, we identified clusters of exons that are co-expressed during early human brain development (Fig. 1a). Moreover, in light of the 4:1 male:female ratio in ASD, the enrichment of sexually different prenatal gene expression for ASD risk genes[10] and ASD-dysregulated co-expression modules[11], as well as findings of sexually different gene expression programs in ASD mouse models[12,13], we focused on clusters that are differentially expressed between males and females during prenatal neurodevelopment, based on data from 20 individuals (10 males and 10 females; Fig. 1b). Second, we compiled WES datasets of 3,531 individuals from 1,704 families who have 1 child with ASD and 1 unaffected sibling (simplex families) and 50 families with 2–5 affected siblings (multiplex families) via the National Institute of Mental Health (NIMH) Data Archive's National Database for Autism Research (NDAR). After joint variant calling within the BrainSpan intervals, we identified variants that are discordant between siblings of simplex families, and those that are shared among all affected siblings of multiplex families (Fig. 1c). Of all these ASD-segregating variants, we focused on inherited, likely gene-disrupting (LGD) ones, namely nonsense, frameshift and splice-site mutations (Fig. 1d).

We mapped variants back to exon clusters to identify neurodevelopmentally co-regulated, ASD-segregating deleterious variants that might have sex-specific effects during early human neurodevelopment (Fig. 1e). We used affected sibling-pair (ASP) analysis to assess the significance of multiplex family variant sharing, and permutation tests to assess the increased burden of deleterious, neurodevelopmentally co-regulated, sex-differentially expressed variation in probands compared with their unaffected siblings, while stringently controlling for multiple hypothesis testing (Methods). Starting with an average of 32,000 SNPs and 3,500 high-confidence insertions–deletions (indels) per individual exome, this approach highlighted, on average, 50 ASD-relevant SNPs and 130 indels per individual (Supplementary Fig. 1). Overall, this analysis identified 33 neurodevelopmentally co-regulated, sex-differentially expressed clusters with ASD-segregating deleterious variation (Supplementary Table 1).

[1]Division of Health and Biomedical Informatics, Department of Preventive Medicine, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA. [2]Northwestern University Clinical and Translational Sciences Institute, Chicago, IL, USA. [3]Institute for Augmented Intelligence in Medicine, Northwestern University, Chicago, IL, USA. [4]Center for Health Information Partnerships, Northwestern University, Chicago, IL, USA. [5]Computational Health Informatics Program, Boston Children's Hospital, Boston, MA, USA. [6]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. [7]Department of Life Sciences and Zlotowski Center for Neuroscience, Ben Gurion University of the Negev, Beer Sheva, Israel. [8]Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, MA, USA. [9]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA. [10]These authors contributed equally: Yuan Luo, Alal Eran. ✉e-mail: isaac_kohane@harvard.edu

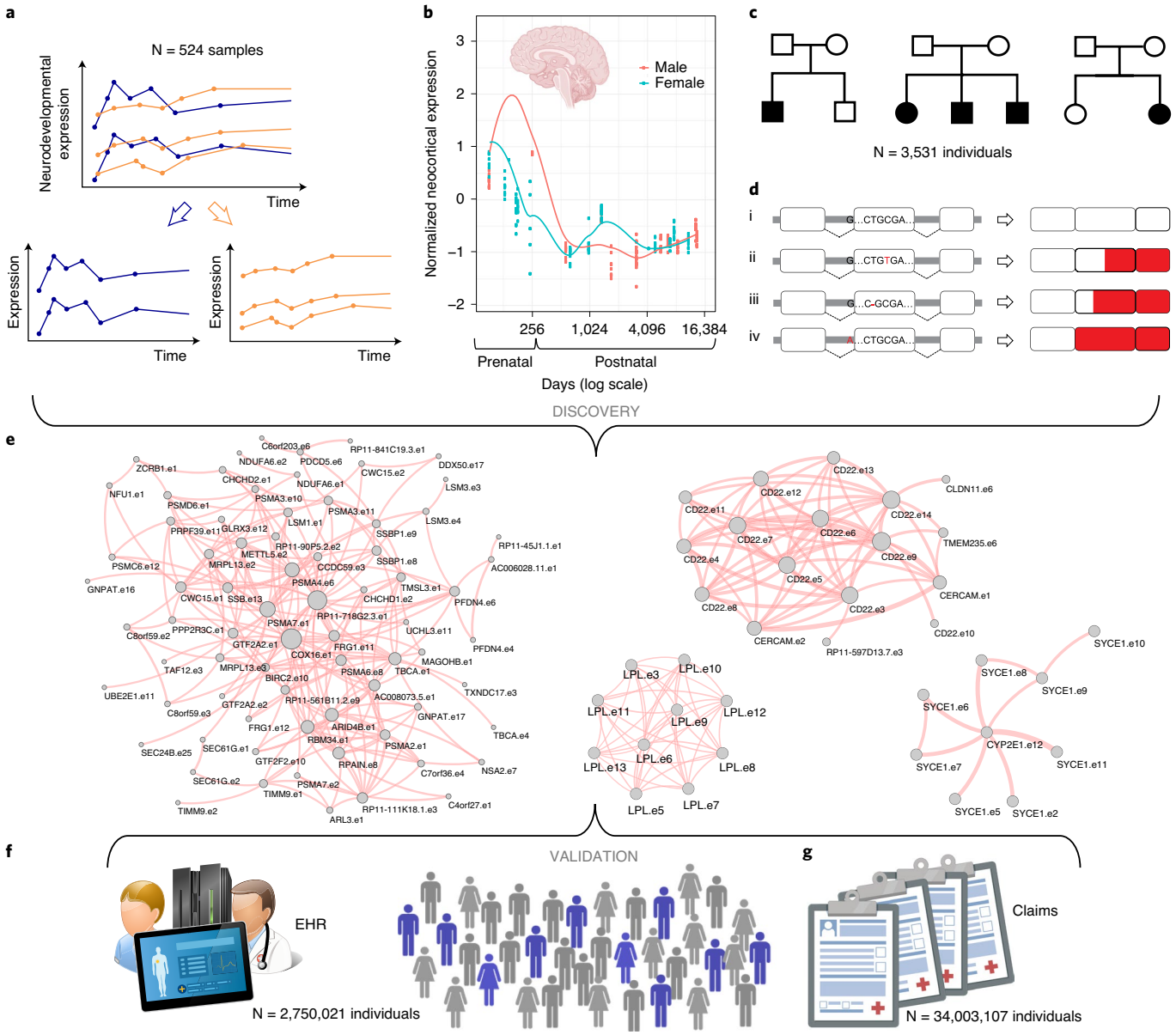


Fig. 1 | Independent sources of information used to identify molecular networks contributing to ASD. a, Neurodevelopmentally co-regulated exons identified by clustering correlated spatiotemporal RNA-seq data of the developing human brain. **b**, Differential expression analysis between males and females identified sex-different exon clusters. **c**, ASD-segregating variants detected in WES data from 3,531 individuals belonging to 1,704 simplex and 50 multiplex families. **d**, Various gene models used to identify LGD variants, including nonsense (**d**, ii), frameshift (**d**, iii) and splice-site (**d**, iv) mutations, the impact of which on the wild-type transcript (**d**, i) is depicted in red. **e**, Information streams **a**–**d** were integrated to identify clusters of sex-differentially expressed, neurodevelopmentally co-regulated, ASD-segregating deleterious variants. **f**,**g**, Lipid dysregulation, a previously unreported molecular theme in nonsyndromic ASD, was validated using large EHR data (**f**; $n = 2{,}750{,}021$ individuals) and massive health claims data (**g**; $n = 34{,}003{,}107$ individuals). **b**,**g**, Images were prepared using BioRender (BioRender.com).

Functional enrichment analysis of the identified exon clusters revealed several molecular themes (Supplementary Table 1), most of which have been previously described in ASD. These include chromatin and transcriptional regulation, immune function and synaptic function[4]. However, it also revealed a previously unrecognized molecular convergence, lipid regulation. For example, a small cluster containing five exons of the low-density-lipoprotein receptor (*LDLR*) gene (19p13.2) was found to carry ASD-segregating deleterious variation ($P = 1.93 \times 10^{-7}$, Fig. 2a,b). Another example of a large cluster, containing 1,926 exons, includes 17 protein-coding gene members of the 'Reactome metabolism of lipids and lipoproteins' pathway, a pathway of 478 human lipid and lipoprotein metabolism genes (Gene Set Enrichment Analysis (GSEA) $P = 4.53 \times 10^{-7}$, cluster $P = 1.18 \times 10^{-12}$). ASD-segregating variants predicted to disrupt the function of genes in this cluster are collectively expected to alter LDL, cholesterol and triglyceride levels (Fig. 2c). We therefore directly tested this hypothesis in two large clinical cohorts. Of note, we found no relationships between LGD variants in lipid metabolism genes and intelligence quotient (IQ; Supplementary Figs. 2 and 3) or sex (Supplementary Figs. 4 and 5) among individuals with ASD.

To test the hypothesis that dyslipidemia might be a convergent etiology in ASD, we compared blood lipid profiles and dyslipidemia diagnoses among individuals with ASD, their unaffected family members and unrelated matched controls. Using the medical

records of 2,750,021 individuals seen at Boston Children's Hospital (BCH), including 25,514 children with ASD, we compared the results of common lipid lab tests between children with ASD and matched individuals with no ASD-related diagnoses. Tests included fasting LDL, total cholesterol and triglycerides, which, of note, are interdependent[14]. We found that children with ASD have blood lipid profiles that are significantly outside the physiological range (LDL: OR = 1.48, 95% CI = (1.36, 1.61), $P = 1.06 \times 10^{-20}$; total cholesterol: OR = 1.69, 95% CI = (1.49, 1.92), $P = 7.14 \times 10^{-15}$; triglycerides: OR = 1.33, 95% CI = (1.20, 1.46), $P = 1.73 \times 10^{-8}$; Fig. 2d and Extended Data Fig. 2). Further stratification based on factors known to affect lipid levels, including age, sex and metabolic state, consistently demonstrated population-level differences between individuals with ASD and matched controls ($P = 1.79 \times 10^{-4}$; Fig. 2e and Supplementary Figs. 6 and 7). For this purpose, an altered metabolic state was defined by the presence of one or more obesity, diabetes or metabolic syndrome X diagnoses.

We also assessed the comorbidity of ASD and dyslipidemia using healthcare claims data from 34,003,107 individuals. We examined the prevalence of dyslipidemia diagnoses in 80,714 individuals diagnosed with ASD, and repeatedly sampled unrelated controls matched by age, sex and zip code, as a marker of socioeconomic status. We found significant enrichment of dyslipidemia diagnoses in individuals with ASD (OR = 1.93, 95% CI = (1.88, 1.99), $P < 1 \times 10^{-323}$; Fig. 2f). Moreover, consistent with our genetic findings in inherited variants, both maternal and paternal dyslipidemia were associated with ASD in the offspring (OR = 1.16, 95% CI = (1.12, 1.20), $P = 5.28 \times 10^{-18}$ for mothers, and OR = 1.13, 95% CI = (1.09, 1.16), $P = 1.92 \times 10^{-14}$ for fathers; Extended Data Fig. 3). To control for genetic background and familial eating habits, we further compared dyslipidemia diagnoses in individuals with ASD and their unaffected siblings, revealing an association of dyslipidemia and ASD within families (OR = 1.76, 95% CI = (1.61, 1.92), Fisher's $P = 2.25 \times 10^{-36}$). All in all, in this US-wide dataset, scaled across multiple healthcare institutions, comorbid dyslipidemia was found in 6.55% (95% CI = (6.38%, 6.72%)) of individuals with ASD.

We next compared core ASD-related features between individuals with dyslipidemia-associated ASD and individuals with ASD and no dyslipidemia (Extended Data Fig. 4). Several clinical characteristics were more common in dyslipidemia-associated ASD, including epilepsy (OR = 1.33, 95% CI = (1.18, 1.51), $P = 5.73 \times 10^{-6}$), sleep disorders (OR = 1.51, 95% CI = (1.36, 1.69), $P = 1.35 \times 10^{-13}$) and attention deficit hyperactivity disorder (OR = 1.30, 95% CI = (1.23, 1.39), $P = 4.61 \times 10^{-18}$), suggesting that dyslipidemia might contribute to altered neurodevelopment in general.

We further characterized the diagnostic spectrum of individuals with dyslipidemia-associated ASD compared with individuals with ASD and no dyslipidemia (Fig. 2g). We excluded individuals with obesity, diabetes and metabolic syndrome X, which are known to affect lipid levels. Several endocrine and metabolic diagnoses were associated with dyslipidemia in ASD, including anemia (OR = 6.00, 95% CI = (5.30, 6.80), $P = 5.84 \times 10^{-174}$), hypothyroidism (OR = 6.19, 95% CI = (5.42, 7.08), $P = 3.93 \times 10^{-157}$) and vitamin D deficiency (OR = 5.02, 95% CI = (4.40, 5.73), $P = 3.93 \times 10^{-157}$). Although each of these conditions has been previously linked to ASD, our findings of a specific association with dyslipidemia in ASD further define an emerging dyslipidemia-associated ASD subgroup.

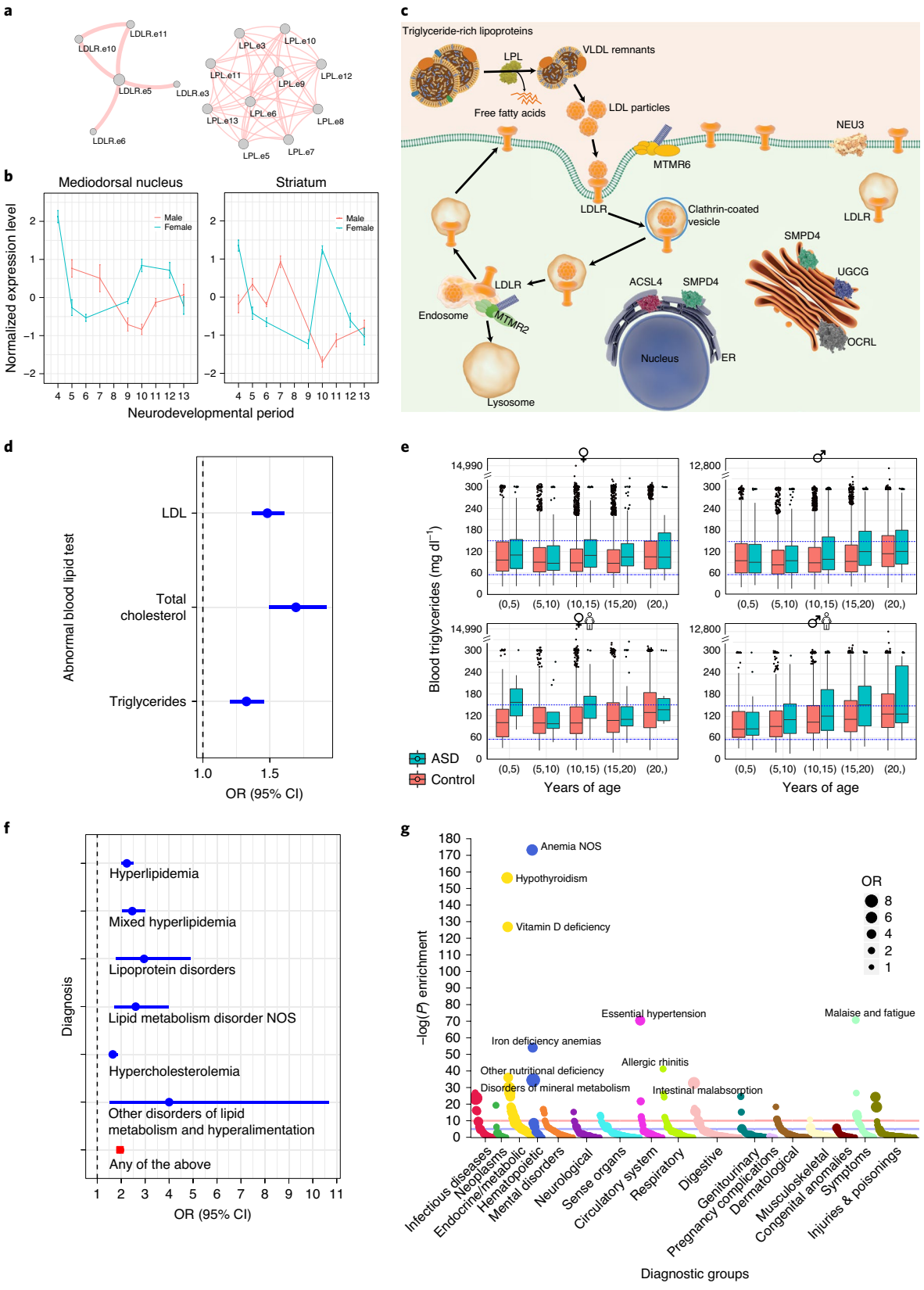
To eliminate potential confounding by drugs commonly prescribed in ASD that are known to alter lipid levels[15], we next restricted our analyses to individuals with no prescription records for atypical antipsychotic, anticonvulsant or antidiabetic drugs. Dyslipidemia remained associated with ASD in these individuals. For example, dyslipidemia diagnoses were more common in individuals with ASD, and no atypical antipsychotic, anticonvulsant or antidiabetic drug prescriptions compared with individuals without ASD and no such prescriptions (OR = 1.73, 95% CI = (1.67, 1.79), $P = 1.11 \times 10^{-201}$; Extended Data Fig. 2a,b). Consistently, in an independent cohort, abnormal blood lipid profiles were more common in individuals with ASD not taking atypical antipsychotics, anticonvulsants or antidiabetics, compared with individuals with no ASD diagnosis and no such drug prescriptions (LDL: OR = 1.48, 95% CI = (1.27, 1.73), $P = 6.16 \times 10^{-7}$; total cholesterol: OR = 1.77, 95% CI = (1.36, 2.27), $P = 2.00 \times 10^{-5}$; triglycerides: OR = 1.33, 95% CI = (1.10, 1.60), $P = 2.99 \times 10^{-3}$; Extended Data Fig. 2c–h).

Finally, we sought to systematically compare the phenotypes of specifically engineered models of dyslipidemia and ASD gene dysfunction. We mined the Mouse Genome Informatics database and clustered all reported phenotypes of mice with targeted mutations in genes implicated in ASD and dyslipidemia, according to the Online Mendelian Inheritance in Man (OMIM) compendium. Both hierarchical and *k*-means clustering of the 1,315 reported

Fig. 2 | Convergence of ASD-segregating deleterious variation on lipid regulation functions, clinically reflected by an association between altered lipid profiles and ASD, and enrichment of dyslipidemia diagnoses in individuals with ASD. a,**b**, Example of single-gene clusters: *LDLR* (Fisher's method, two-sided $P = 1.68 \times 10^{-5}$, $n = 3{,}531$ individuals) and *LPL* (Fisher's method, two-sided $P = 2.75 \times 10^{-3}$, $n = 3{,}531$ individuals). **a**, Pairwise correlation structure of neurodevelopmentally co-expressed *LDLR* and *LPL* exons. Edges connect exon nodes with an R^2 correlation coefficient ≥0.7. **b**, Sex-differential neurodevelopmental expression patterns of the *LDLR* cluster in the mediodorsal nucleus of the thalamus and striatum. The mean normalized expression pattern is shown across neurodevelopmental periods defined in Supplementary Table 3 among 18 males (cyan) and 15 females (red). **c**, Example of a large sex-different co-expression cluster harboring ASD-segregating deleterious variants (Fisher's method, two-sided $P = 1.18 \times 10^{-12}$, $n = 3{,}531$ individuals), enriched with genes of the 'Reactome metabolism of lipids and lipoproteins' pathway. Relationships between pathway genes found to carry ASD-segregating deleterious variants and blood LDL, cholesterol and triglycerides are shown. ER, endoplasmic reticulum; VLDL, very-low-density lipoprotein; MTMR6, myotubularin-related protein 6; NEU3, neuraminidase 3; SMPD4, sphingomyelin phosphodiesterase 4; UGCG, UDP-glucose ceramide glucosyltransferase; OCRL, oculocerebrorenal syndrome of Lowe (OCRL) inositol polyphosphate-5-phosphatase; ACSL4, acyl-CoA synthetase long-chain family member 4; MTMR2, myotubularin-related protein 2. Image was prepared using somersault18:24 (www.somersault1824.com) and BioRender (BioRender.com). **d**, Association of altered lipid profiles and ASD among 2,750,021 BCH patients. ORs and 95% CIs are shown by circles and horizontal lines, respectively. **e**, Fasting triglyceride levels stratified on age, sex (horizontally) and metabolic state (vertically), which are all known to affect lipid levels. Lab test results from individuals with ASD are shown in cyan and those from stratum-matched controls in red; blue dotted lines represent the reference range. Results from individuals with metabolic dysregulation, including obesity, diabetes and metabolic syndrome X, are shown at the bottom. Box plots show the median as a thick line and the 25th and 75th percentiles as upper and lower bounds of the box. The whiskers extend to 1.5× IQR from the median (Kruskal–Wallis, two-sided $P = 1.79 \times 10^{-4}$, $n = 33{,}056$ individuals). **f**, Enrichment of dyslipidemia diagnoses in individuals with ASD, as detected in healthcare claims data from 34,003,107 Americans. A forest plot depicts diagnosis-specific association estimates and their 95% CIs for ASD and dyslipidemia as circles and horizontal lines, respectively (overall OR = 1.93, 95% CI = (1.88, 1.99), Fisher's exact test, two-sided $P < 1.00 \times 10^{-323}$, $n = 80{,}714$ individuals with ASD versus repeatedly resampled matched controls). NOS, not otherwise specified. **g**, Diagnostic characterization of individuals with dyslipidemia-associated ASD versus ASD with no dyslipidemia, based on healthcare claims data ($n = 80{,}714$ individuals). The diagnoses enriched in individuals with dyslipidemia-associated ASD, their ORs and −log(enrichment *P* value) are shown.

phenotypes demonstrated that some ASD mouse models are more similar to dyslipidemia mice than to other ASD models (Extended Data Fig. 5 and Supplementary Fig. 8). Murine models of dyslipidemia have social and nervous system abnormalities, as observed in the ASD models, and several ASD mice show lipid homeostasis alterations and growth abnormalities characteristic of dyslipidemia models (Supplementary Table 2).

In the present study, using massive multimodal data triangulation, we identified a robust nonsyndromic ASD subtype characterized by dyslipidemia. This emerging subtype is consistent not only



with a rare syndromic form of ASD[16,17] but also with the reported neuronal functions of bona fide dyslipidemia genes during mammalian brain development. For example, lipoprotein lipase (LPL) plays a key role during neuronal differentiation[18], LDLR is an important neuronal signaling mediator[19] and proprotein convertase subtilisin/kexin type 9 (PCSK9) regulates neuronal apoptosis[20]. Our findings are also consistent with a previous pathway-level analysis of ASD-implicated genes that reported an enrichment of lipid metabolism pathways[21]. Moreover, several studies have shown that MeCP2, the dysfunction of which causes Rett syndrome, regulates cholesterol metabolism[22,23]; Rett syndrome, once considered part of the autism spectrum, is characterized by overlapping neurodevelopmental deficits[24]. Other studies have demonstrated how lipid dysregulation during early prenatal neurodevelopment could lead to altered Wnt-dependent migration and proliferation via prostaglandin E_2 (refs. [25,26]), thereby providing a testable mechanism for future studies. Notably, the convergence on lipid dysregulation seems to be focused on non-fatty acid-related lipids.

The link between ASD and dyslipidemia is exemplified by Smith–Lemli–Opitz syndrome (SLOS), a rare syndrome of congenital malformations and intellectual disability caused by a cholesterol biosynthesis defect, with 50–88% of cases reported to be in individuals on the autism spectrum[16,17]. Together with pioneering reports of dyslipidemia in individuals with ASD[27–29], the present study supports the existence of a robust dyslipidemia-associated, nonsyndromic ASD subtype. In light of their modest effect sizes, the potential clinical utility of the association of parental dyslipidemia with ASD in the offspring, as well as findings of altered blood lipid profiles in infants later diagnosed with ASD, should be directly tested by subsequent studies. Our results enable the selection of better-defined populations for further research and offer rational targets for intervention and prevention. As in oncology, where the identification of cancer subtypes enabled the development of effective targeted treatments[7–9], the identification of ASD subtypes is expected to result in similar opportunities for therapeutic development. Overall, the work presented in the present study represents a proof of concept for the value of using massive amounts of existing multimodal data to push the boundaries of existing knowledge, thereby moving us closer to precision medicine for ASD.

Online content

Any methods, additional references, Nature Research reporting summaries, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41591-020-1007-0.

Received: 25 August 2017; Accepted: 2 July 2020;
Published online: 10 August 2020

References

1. National Research Council (US) Committee ona Framework for Developing a New Taxonomy of Disease. *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease* (National Academies Press, 2011).
2. Maenner, M. J. et al. Prevalence of autism spectrum disorder among children aged 8 years—autism and developmental disabilities monitoring network, 11 sites, United States, 2016. *MMWR Surveill. Summ.* **69**, 1–12 (2020).
3. de la Torre-Ubieta, L., Won, H., Stein, J. L. & Geschwind, D. H. Advancing the understanding of autism disease mechanisms through genetics. *Nat. Med.* **22**, 345–361 (2016).
4. Lord, C. et al. Autism spectrum disorder. *Nat. Rev. Dis. Prim.* **6**, 5 (2020).
5. Li, J. et al. Integrated systems analysis reveals a molecular network underlying autism spectrum disorders. *Mol. Syst. Biol.* **10**, 774 (2014).
6. Parikshak, N. N. et al. Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell* **155**, 1008–1021 (2013).
7. Herbst, R. S. et al. Predictive correlates of response to the anti-PD-L1 antibody MPDL3280A in cancer patients. *Nature* **515**, 563–567 (2014).
8. Chapman, P. B. et al. Improved survival with vemurafenib in melanoma with *BRAF* V600E mutation. *N. Engl. J. Med.* **364**, 2507–2516 (2011).
9. Parsons, D. W. et al. An integrated genomic analysis of human glioblastoma multiforme. *Science* **321**, 1807–1812 (2008).
10. Shi, L., Zhang, Z. & Su, B. Sex biased gene expression profiling of human brains at major developmental stages. *Sci. Rep.* **6**, 21181 (2016).
11. Werling, D. M., Parikshak, N. N. & Geschwind, D. H. Gene expression in human brain implicates sexually dimorphic pathways in autism spectrum disorders. *Nat. Commun.* **7**, 10717 (2016).
12. Jung, H. et al. Sexually dimorphic behavior, neuronal activity, and gene expression in Chd8-mutant mice. *Nat. Neurosci.* **21**, 1218–1228 (2018).
13. Grissom, N. M. et al. Male-specific deficits in natural reward learning in a mouse model of neurodevelopmental disorders. *Mol. Psychiatry* **23**, 544–555 (2018).
14. Rosenson, R. Measurement of blood lipids and lipoproteins. in UpToDate (ed. Post, T. W.) https://www.uptodate.com (accessed 22 January 2018).
15. Coleman, D. M., Adams, J. B., Anderson, A. L. & Frye, R. E. Rating of the effectiveness of 26 psychiatric and seizure medications for autism spectrum disorder: results of a national survey. *J. Child Adolesc. Psychopharmacol.* **29**, 107–123 (2019).
16. Sikora, D. M., Pettit-Kekel, K., Penfield, J., Merkens, L. S. & Steiner, R. D. The near universal presence of autism spectrum disorders in children with Smith–Lemli–Opitz syndrome. *Am. J. Med. Genet. A* **140**, 1511–1518 (2006).
17. Tierney, E. et al. Behavior phenotype in the RSH/Smith–Lemli–Opitz syndrome. *Am. J. Med. Genet.* **98**, 191–200 (2001).
18. Gong, H. et al. Lipoprotein lipase (LPL) is associated with neurite pathology and its levels are markedly reduced in the dentate gyrus of Alzheimer's disease brains. *J. Histochem. Cytochem.* **61**, 857–868 (2013).
19. Beffert, U., Stolt, P. C. & Herz, J. Functions of lipoprotein receptors in neurons. *J. Lipid Res.* **45**, 403–409 (2004).
20. Kysenius, K., Muggalla, P., Matlik, K., Arumae, U. & Huttunen, H. J. PCSK9 regulates neuronal apoptosis by adjusting ApoER2 levels and signaling. *Cell. Mol. Life Sci.* **69**, 1903–1916 (2012).
21. David, M. M. et al. Comorbid analysis of genes associated with autism spectrum disorders reveals differential evolutionary constraints. *PLoS ONE* **11**, e0157937 (2016).
22. Buchovecky, C. M. et al. A suppressor screen in Mecp2 mutant mice implicates cholesterol metabolism in Rett syndrome. *Nat. Genet.* **45**, 1013–1020 (2013).
23. Kyle, S. M., Saha, P. K., Brown, H. M., Chan, L. C. & Justice, M. J. MeCP2 co-ordinates liver lipid metabolism with the NCoR1/HDAC3 corepressor complex. *Hum. Mol. Genet.* **25**, 3029–3041 (2016).
24. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*, 5th edn (American Psychiatric Association, 2013).
25. Wong, C. T., Wais, J. & Crawford, D. A. Prenatal exposure to common environmental factors affects brain lipids and increases risk of developing autism spectrum disorders. *Eur. J. Neurosci.* **42**, 2742–2760 (2015).
26. Wong, C. T. et al. Prostaglandin E_2 promotes neural proliferation and differentiation and regulates Wnt target gene expression. *J. Neurosci. Res.* **94**, 759–775 (2016).
27. El-Ansary, A. & Al-Ayadhi, L. Lipid mediators in plasma of autism spectrum disorders. *Lipids Health Dis.* **11**, 160 (2012).
28. Kim, E. K., Neggers, Y. H., Shin, C. S., Kim, E. & Kim, E. M. Alterations in lipid profile of autistic boys: a case control study. *Nutr. Res.* **30**, 255–260 (2010).
29. Tierney, E. et al. Abnormalities of cholesterol metabolism in autism spectrum disorders. *Am. J. Med Genet. B Neuropsychiatr. Genet.* **141B**, 666–668 (2006).

Methods

Neurodevelopmental co-expression analysis. To understand which variants might function together during human brain development, we examined exonic spatiotemporal co-expression patterns in the BrainSpan RNA-sequencing (RNA-seq) data (http://www.brainspan.org/api/v2/well_known_file_download/267666524). This dataset contains normalized read counts (in RPKM: reads per kilobase transcript per million mapped reads) for 309,223 coding and noncoding exons measured across 524 samples from 26 brain regions, dissected from 23 males and 19 females throughout typical human neurodevelopment (Supplementary Tables 1 and 4). These samples create a spatiotemporal profile for each exon during typical human neurodevelopment. Co-expression analysis based on these profiles can identify exons that might be functionally co-regulated. The motivation for focusing on exon-level co-regulation stems from the notion that alternative splicing plays a central role during brain development and different isoforms of the same gene may exert different functions during human neurodevelopment in health and disease[30–32].

We first processed the RNA-seq data by treating all 0 values as not available (NA) and $\log_2$-transforming the RPKM values by using the formula $\log_2(x+1)$ to reduce the effects of wild points and measurement noise. We then filtered exons based on the following exclusion criteria:

1. Variability filter: if there was no change in the expression profile (that is, expression levels were the same for different brain areas at different developmental stages from different donors), then the exon was excluded.
2. Multi-sample filter: an exon was excluded if it was present in samples from a single donor only.
3. Duplicate filter: some exonic intervals were duplicated in the BrainSpan Gencode v.10 RNA-seq data, where duplicates may have been labeled with different (and some temporary) names. These were consolidated based on the criterion of sharing a genomic interval.
4. Detection filter: we kept those exons with at least 5% non-NA expression values.
5. Sample overlap: we also required that candidate pairs of exons be detected in 75% of samples or more; that is, $|\text{intersect}(\text{nna}(e_1), \text{nna}(e_2))| \geq 75\% \times \max(|\text{nna}(e_1)|, |\text{nna}(e_2)|)$, where nna($e$) refers to the number of samples for which exon e has non-NA expression values, intersect() denotes the operation of set intersection and $|\ |$ returns the length of a vector.

After exon filtration, we identified co-regulated exons by calculating their expression correlation across the BrainSpan dataset. We measured the coefficient of determination $R^2 = \text{cor}(e_1,e_2)^2$, where $\text{cor}(e_1,e_2)$ represents Pearson's correlation between expression profiles of two exons: e_1 and e_2 (ref. [33]). The R^2 coefficient measures how well e_1 might be constructed from e_2 (by creating a revised predictor of the form $\alpha + \beta e_2$), and vice versa.

Pairwise correlation calculation across 248,898 exons after filtering amounts to over 47 billion pairs and is a daunting task that is intensive in both computation and storage. Thus, we adopted a distributed block-wise approach to calculate pairwise exon correlations, as shown in Supplementary Fig. 9. By dividing all exons into blocks of size 10,000, the correlation calculation was paralleled in a block-wise fashion. Let the blocks be $b_1, \ldots, b_n$; we then need to compute correlations between the exons within b_1, correlations between exons in b_1 and b_2, …, correlations between exons in b_1 and b_n, correlation between exons in b_2 and b_3 (b_2–b_1 block correlations can be omitted due to symmetry), and so on. Each block-wise correlation was dispatched to its own compute node in a 2,000-core computing cluster, thereby achieving a 1,000-fold speed up. We used Python 2.7 to generate the parallel processing scripts.

Identification of co-regulated exon clusters. The distribution of the coefficients of determination R^2 across all exon pairs is shown in Supplementary Fig. 10. As depicted in Supplementary Fig. 10a, cluster frequency falls at a speed faster than exponential as R^2 increases. This distribution applies to all exons passing the five filters described under Neurodevelopmental co-expression analysis. As the focus of our analysis is highly co-expressed exons, we set an empirical criterion that two exons must have $R^2 \geq 0.7$ to be considered co-expressed, thereby focusing on the 0.02% most correlated exon pairs. We kept all exons that are co-expressed with at least one other exon, and represented them as a graph. This graph has exons as nodes and draws edges connecting nodes e_1 and e_2 if $R^2(e_1,e_2) \geq 0.7$. Thus, a large sparse exon co-expression graph with 92,240 nodes and 6,205,327 edges was produced. A small part of this exon graph is shown in Supplementary Fig. 11, demonstrating that the whole exon graph consists of smaller exon clusters.

Within the entire exon graph, we identified co-expression clusters by finding maximally connected components and their community structures, using the R igraph package v.1.2.0 (ref. [34]). First, we identified 6,242 distinct co-expressed exon clusters with a mean R^2 of 0.82 and an averaged exon count of 15. This collection of exon clusters is remarkably heterogeneous in size, that is, clusters contain different number of exons (Supplementary Table 5) and genes (Supplementary Table 6). Although these distributions were skewed toward smaller exon clusters, there were numerous exon clusters representing tight multi-gene co-expression. On the other hand, one 69,587-exon megacluster arose, calling for finer resolution clustering for further module detection. We therefore adopted a second graph-clustering algorithm that maximizes the modularity of information flow across clustered graph components[35]. The distribution of cluster sizes and number of genes of the additional 3,382 connected exon clusters are shown in Supplementary Tables 7 and 8, respectively.

Quantifying cluster-level expression throughout neurodevelopment. We next tracked the temporal expression profiles of co-regulated exon clusters identified in the previous step, across the BrainSpan regions. As defined in Kang et al.[30] and shown in Supplementary Table 4, the measured brain regions and areas can be summarized into six brain structures: amygdaloid complex (AMY), cerebellar cortex (CBC), neocortex (NCX), hippocampus (HIP), mediodorsal thalamus (MD) and striatum (STR). Based on the mapping in Supplementary Table 4, we derived the expression profile for the areas with equations (1)–(8):

$$\text{FC} = \text{mean}(\text{OFC}, \text{DFC}, \text{VFC}, \text{MFC}, (\text{M1C}|\text{M1C} - \text{S1C})) \quad (1)$$

$$\text{PC} = \text{mean}(\text{PCx}, \text{IPC}, \text{S1C}) \quad (2)$$

$$\text{TC} = \text{mean}(\text{TCx}, \text{ITC}, \text{A1C}, \text{STC}) \quad (3)$$

$$\text{OC} = \text{mean}(\text{OCx}, \text{V1C}) \quad (4)$$

$$\text{NCX} = \text{mean}(\text{FC}, \text{PC}, \text{TC}, \text{OC}) \quad (5)$$

$$\text{STR} = \text{mean}(\text{STR}, \text{MGE}, \text{LGE}, \text{CGE}) \quad (6)$$

$$\text{MD} = \text{mean}(\text{MD}, \text{DTH}) \quad (7)$$

$$\text{CBC} = \text{mean}(\text{CBC}, \text{CB}) \quad (8)$$

FC, frontal cortex; OFC, orbital prefrontal cortex; DFC, dorsolateral prefrontal cortex; VFC, ventrolateral prefrontal cortex; MFC, medial prefrontal cortex; M1C, primary motor cortex; PC, parietal cortex; S1C, primary somatosensory cortex; IPC, posterior inferior parietal cortex; TC, temporal cortex; A1C, primary auditory cortex; STC, posterior superior temporal cortex; ITC, inferior temporal cortex; OC, occipital cortex; V1C, primary visual cortex; NCX, neocortex; MGE, medial ganglionic eminence; LGE, lateral ganglionic eminence; CGE, caudal ganglionic eminence; STR, striatum; DTH, dorsal thalamus; MD, mediodorsal nucleus; CB, cerebellum; CBC, cerebellar cortex. Using these equations, we calculated the aggregated expressions of all exons, in each brain area across the entire cohort. To compare across individuals and brain regions, we first normalized the expression levels using a Z transformation (that is, centering the expression vector on the mean and dividing by its s.d.). We next tracked the temporal expression patterns in each sex using three approaches:

1. LOWESS smoothed line and scatter plot: in this approach, for each exon cluster, each brain structure and each sex, we plotted the expression of every exon in the cluster from all matching sample donors, with the axes being expression and time. We then fit a temporal expression curve across all points using locally weighted scatter-plot smoothing (LOWESS)[36]. We used the implementation from R stats package v.4.1.0 with the following parameters: span = 0.5, degree = 1.
2. Mean and s.e.m. plot: in this approach, for each exon cluster, each brain structure, each sex and each time point, we computed the mean expression and its s.e.m. by using the aggregated expressions of all exons in that exon cluster, from all matching sample donors. We plot the temporal expression profile for each combination of exon cluster, brain area and sex using line graph with means as values and the s.e.m. as an error bar at each time point.
3. Mean and s.e.m. period plot: the spatiotemporal dynamics of the human brain transcriptome is a staged process and can be tracked as a multi-period system, as detailed in Supplementary Table 3. For each exon cluster, each brain structure, each sex and each neurodevelopmental period, we computed the mean expression level and its s.e.m. by using the aggregated expression of all exons in that exon cluster, from all matching sample donors. The temporal expression profiles were then similarly plotted as in approach 2.

Identification of sex-different co-regulated exons. To identify sex-different co-regulated exons during human brain development, we compared the temporal expression profiles of each exon cluster in each brain structure between male and female individuals. Following Werling et al.[11], we performed differential expression analysis using Limma v.3.42.2 (ref. [37]), a robust method of linear mixed models and Bayesian t-tests for analyzing small cohorts. We used sex as the main contrast in the regression models, and included subject as a random effect to account for the non-independence of samples from the same individual. As fixed effects in the

model, we included age, brain structure, postmortem interval and RNA integrity number, as detailed in Kang et al.[30]. *P* values were adjusted for multiple testing of all clusters using the Benjamini–Hochberg approach[38]. Supplementary Table 9 details the results of this analysis.

WES data compilation. We compiled familial WES studies from the NIMH Data Archives' NDAR, as detailed in Supplementary Table 10. The BCH Institutional Review Board (IRB) has determined that this research qualifies as exempt from the requirements of human subject protection regulations. Supplementary Table 10 shows the number of included families from each dataset. Inclusion criteria were: (1) families with at least two siblings who have a similar degree of sequence coverage, as determined by the Genome Analysis Toolkit (GATK) CallableLoci analysis[39]; and (2) families with variant data enabling the performance of variant quality score recalibration, offering a tunable sensitivity–specificity tradeoff[39]. A total of 1,754 families met the inclusion criteria and were included in our analysis. Of these, 50 were multiplex families with 2–5 affected siblings, and 1,704 were simplex families with 1 affected and 1 unaffected full sibling. The total number of individuals included in our analysis amounts to 3,531. To accurately and consistently call variants across all datasets, we followed the GATK framework[39] for a standardized preprocessing of WES data into analysis-ready reads, followed by joint variant calling.

WES preprocessing. For each individual included in the present study, multiple BAM files may have been generated by multiple sequencing runs. Furthermore, different studies used different aligners and different variant calling frameworks. To standardize variant calling and data analysis across studies, our data preprocessing began with converting BAM files back to interleaved fastq files and aligning these in a standardized manner using BWA-MEM v.0.7.10 (http://bio-bwa.sourceforge.net/bwa.shtml). This step ensures that the BAM files are processed in the same standard way to improve variant calling accuracy. Before converting a BAM file to a fastq file, we first split the BAM files into multiple read groups. We then used Picard v.1.119 (http://broadinstitute.github.io/picard) to undo possible post-alignment processing for each split BAM file, using the RevertSAM utility. The actual conversion from BAM files to fastq files included the following two substeps: the first used the 'bamshuf' utility from Samtools v.1.1 (https://github.com/samtools/samtools) to shuffle the reads in the BAM file for them not to be in any biased order, so that subsequent alignment could correctly estimate the insert size using blocks of paired reads. The second substep used the 'bam2fq' utility from Samtools to convert the BAM file to an interleaved fastq file, where each pair of reads (forward and reverse reads) is in the same file. The interleaved fastq files from all individuals were then mapped to a single human reference genome (GRCh37/hg19, including decoy contigs) using BWA-MEM. The newly aligned BAM files containing different read groups were then merged using the Picard MergeSamFiles utility. For the merged BAM file, duplicates were marked and removed using the Picard MarkDuplicates utility, and read group information was added using the Picard AddOrReplaceReadGroups utility.

For efficiency, we restricted variant calling to a limited set of chromosomal regions specified by the BrainSpan exon intervals. This is because we are interested only in neurodevelopmentally co-regulated variants in the present study. We padded each BrainSpan exon with a 100-basepair (bp) buffer. We sorted the padded intervals and divided them into two collections based on whether they are on the forward or the reverse strands. We then merged intervals overlapping with other intervals in the same collection to provide a nonoverlapping collection of intervals on each strand. The union of the two collections of merged intervals then formed the BrainSpan reference interval. Supplementary Fig. 12 shows the size distribution of padded merged BrainSpan intervals. This figure also categorizes the intervals based on their strands (forward or reverse), and depicts the distributions of those intervals respectively, which are similar to each other and similar to that of all intervals.

Joint variant calling in BrainSpan intervals. After preprocessing, we performed joint variant calling using the GATK v.3.3 (ref. [39]) in the padded BrainSpan intervals. The preprocessed BAM files underwent local realignment, which transformed regions with misalignments due to indels into clean reads with a consensus indel model, using the GATK RealignerTargetCreator and IndelRealigner utilities. The base quality scores were then recalibrated to correct for artifact and offset bias, using the GATK BaseRecalibrator utility, producing analysis-ready reads.

The analysis-ready reads were then processed using the GATK Haplotype Caller. This step simultaneously calls SNPs and indels using local reassembly of haplotypes in an active region, resulting in per-position genotype likelihood. We used the human reference genome GRCh37/hg19 including decoy contigs as reference for the Haplotype Caller, using the recommended setting for single-sample all-sites calling on DNA-seq: emitRefConfidence = GVCF, variant_index_type = LINEAR, variant_index_parameter = 128000.

We then combined the resulting per-sample variants and performed joint genotyping step using the GATK's GenotypeGVCFs utility. Joint genotyping aggregated multi-sample variants and merged the records to re-estimate the genotype likelihood by combining all records spanning the target location. Based on our joint genotyping results, we applied a variant quality score recalibration, a machine-learning-based, variant-filtering step. SNPs and indels were recalibrated separately in two passes. The first pass recalibrated SNPs, with indels left untouched; the second pass recalibrated indels, with recalibrated SNPs left untouched.

We applied the WES preprocessing and joint variant calling steps to samples from the multiplex family cohort. For the discordant family cohort, we used a subset of the dataset produced by Krumm et al.[40], which is based on a similar GATK pipeline. There are two main differences between the pipelines by Krumm et al.[40] and our pipeline: (1) Krumm et al. performed joint variant calling separately for each quad (parents, proband and unaffected sibling) instead of the entire cohort; (2) Krumm et al. called variants within 20 bp of the NimbleGen EZ-SeqCap v.2.0 targets rather than within 100 bp of BrainSpan interval targets. The first difference may introduce some bias when directly comparing called samples from the two cohorts. However, we performed segregation analysis separately on the two cohorts, thus avoiding such bias. The second difference resulted in disparate numbers of variants per individual between the two cohorts. However, as detailed below and depicted in Supplementary Figs. 13 and 14, our subsequent filtering steps (mapping to BrainSpan exon clusters in particular) made the total number of variants per individual comparable between the two cohorts. Finally, we retained only those variants with the most confident genotype assignments, as quantified by the GATK's genotype quality (GQ) score, requiring GQ = 99.

Annotation-based variant filtering and deleterious variant detection. We next used ANNOVAR v.2018Apr16 (https://annovar.openbioinformatics.org) to comprehensively annotate called variants with a wide array of information. Annotations included the host gene (using several gene models such as RefSeq, UCSC Known Gene and Gencode), variant function and minor allele frequency among various populations as determined by gnomAD, and phenotype associations according to ClinVar and HGMD. To address issues of reference misannotation, we filtered out variants with minor allele frequency >90% among the 125,748 individual exomes aggregated by gnomAD v.2.1 (https://gnomad.broadinstitute.org). We further focused on likely gene disruptive (LGD) variants, which include frame-shift insertions, frame-shift deletions, nonsense variants and splice-site mutations.

Segregation pattern analysis in families discordant for ASD. Simplex families in this study refer to those that have one child diagnosed with ASD. We focused on discordant families, a special case of simplex families that have two siblings: one proband (affected with ASD) and one unaffected sibling. In each discordant family, discordant sibling pairs were formed by pairing a proband with his or her own unaffected sibling. With the collection of discordant pairs, we compared neurodevelopmentally co-regulated LGD variants found in the proband with those carried by the unaffected sibling, in each exon cluster. We used permutation tests[41] to assess the statistical significance of an exon cluster's excess deleterious variation in probands compared with their unaffected siblings. Treating each family as rows and probands and sibling as columns, we filled the entries of this matrix with the total number of variants detected in each individual in each exon cluster. This created an exon cluster mutational profile among discordant families. To obtain an empirical *P* value for excess mutational burden, we randomly shuffled paired probands and siblings. Repeating the permutation created a distribution of mutational profiles that simulates exon cluster LGD variants by chance. With this simulated distribution, we then calculated the *P* value of differential variation (that is, $\sum_i \left(m^e_{p_i} - m^e_{s_i} \right)$, where e indexes the exon clusters, i indexes discordant families, and p_i and s_i are the proband and unaffected sibling in the ith family, respectively).

Segregation pattern analysis in multiplex families. Multiplex families have two or more affected probands. In this segregation analysis we searched for neurodevelopmentally co-regulated LGD variants that are shared among all affected siblings. Although most multiplex families have two affected siblings, there were 16 families with 3–5 affected siblings. We used ASP analysis[42] to assess the significance of variant sharing among all proband siblings. We followed an extended version of the ASP test[42]. The null hypothesis of this test is that variant sharing is by chance, and therefore not related to the phenotype. This hypothesis was tested using the nonparametric linkage z-score. To deal with multiplex families of more than two sibs, we divided each family into sib pairs (that is, a family with s sibs would result in $s \times (s-1)/2$ affected sib pairs). Although the artificially created pairs are dependent, we did not scale them down by $s/2$, as if there were only $s-1$ pairs in the sibship. This was done to retain optimal power because most of the multiplex families had four or fewer sibs (only two families had five sibs)[42].

Mapping variants onto co-regulated exon clusters. To identify neurodevelopmentally co-regulated LGD variants, we next mapped the identified LGD variants to the discovered exon clusters. In doing so we first performed an interval search to map variants into exons using the GenomicRanges toolkit[43]. A variant was mapped into an exon if it overlapped the exon's genomic interval. This mapping of LGD variants to exon clusters allowed us to aggregate deleterious mutations in each co-regulated exon cluster, thereby reducing the complexity of the problem by several orders of magnitude. Supplementary Figs. 13 and 14 show the distributions of the number of variants per individual at each stage of the analysis, for the discordant family cohort and the multiplex family cohort, respectively.

From Supplementary Fig. 13, it can easily be seen that the steps of restricting to LGD variants—restricting variants to co-regulated exon clusters and filtering for differentially variable variants—all contribute to the reduction in the number of candidate variants. Similar reduction holds for multiplex families, where the last filtering step is based on shared variants among all proband siblings, as shown in Supplementary Fig. 14.

Integrated statistical significance. As detailed in "Segregation pattern analysis in families discordant for ASD" and "Segregation pattern analysis in multiplex families", in simplex families we calculated *P* values for the statistical significance of excess LGD variation in probands compared with their unaffected sibs in each neurodevelopmentally co-regulated exon cluster. In multiplex families, we calculated *P* values for increased deleterious allele sharing among all affected sibs for each cluster. In this step, we combined the two sources of association evidence using Fisher's method[44]. As the analytical focus was at the cluster level, the combined *P* values were then Bonferroni corrected for multiple testing of all clusters[45]. Supplementary Table 11 details the composition of the resulting 33 significant clusters.

Functional enrichment analysis. To assess the function of single gene exon clusters, we used the National Center for Biotechnology Information's gene2go table (ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2go.gz) to map genes to their molecular function, biological process and cellular compartment. For multi-gene exon clusters we used GSEA v.4.0.3 (ref. [46]) to identify gene membership in Kyoto Encyclopedia of Genes and Genomes pathways (http://www.genome.jp/kegg), Reactome pathways (http://www.reactome.org), BioCarta pathways (http://www.biocarta.com) and their pathway interactions, as recorded in the Pathway Interaction Database (https://wiki.nci.nih.gov/pages/viewpage.action?pageId=315491760). SFARI Gene, an integrated catalog of human genetic studies related to autism, was used to examine the significant cluster genes' known association with ASD (https://gene.sfari.org). Only genes belonging to evidence categories S, 1 and 2, were considered as having a strong prior for playing a role in ASD. Furthermore, ClinVar (https://www.ncbi.nlm.nih.gov/clinvar) and OMIM (https://www.omim.org) were mined in search of significant cluster genes' implication in schizophrenia and bipolar disorder, two related neurodevelopmental disorders with etiologies that overlap with those of ASD[4].

Affinity clustering and reanalysis. To ensure the reproducibility of our genomic analyses we repeated exon clustering with another clustering approach, affinity clustering[47]. We used affinity clustering to generate medium, more uniformly sized, connected exon clusters. To construct an affinity network using distance as edge weights, we converted the R^2 score between exons to the distance between exons using the following preprocessing steps. We first normalized the R^2 score of an edge, dividing it by the product of the square roots of degrees of its two nodes. We then used 1 minus the normalized R^2 score as distance. As affinity clustering requires only relative order of distances and is otherwise agnostic to absolute distance values, we needed only to ensure that the converted distance is ≥ 0. With the above preprocessing done, we obtained 491 moderately sized connected exon clusters, the sizes of which are more evenly distributed (number of exons per cluster: median 76, interquartile range (IQR) (38, 159); number of genes per cluster: median 19, IQR (7, 56)). Supplementary Table 12 details significant exon clusters identified with affinity clustering, demonstrating the similarity to the original results in terms of their identified molecular themes.

Analysis of blood lipid profiles among BCH patients. We used the i2b2/tranSMART platform v.1.7.09 (ref. [48]) to analyze EHR data from 2,750,021 individuals seen at BCH, including 25,514 children with ASD. The BCH IRB has determined that this research qualifies as exempt from the requirements of human subject protection regulations; i2b2/tranSMART enables the cohesive analysis of heterogeneous phenotypic data, including longitudinal diagnoses and lab results. Using this engine, we compared the results of common lipid lab tests in individuals with ASD and matched individuals with no ASD-related diagnoses. Tests included triglyceride levels, total cholesterol and LDL. For each lab, a 2×2 contingency table was constructed to compare the association of abnormal lab results with ASD, as indicated by the presence of one or more International Classification of Diseases, 9th revision (ICD-9) codes in the 299 group/ICD-10 codes in the F84 group (pervasive developmental disorders) in at least one record[49,50]. Fisher's exact tests were used to assess the statistical significance of the association of abnormal lipid lab results and ASD.

We also compared actual blood lipid levels between individuals with ASD and age, sex and metabolic state-matched controls. For this purpose, an altered metabolic state was defined by the presence of one or more obesity, diabetes or metabolic syndrome X diagnoses. We extracted all diagnoses and demographics of de-identified BCH patients with at least one lipid test result of interest and at least 2 years of coverage ($n=103{,}484$). For each blood test, comparisons were made across strata using a generalized linear model.

Analysis of healthcare claims data. We analyzed 4 calendar years' (2010–2013) worth of medical claims and enrollment demographics for 34,003,107 Americans who were covered by Aetna Inc. healthcare during that period. The Harvard Medical School IRB has determined that this research qualifies as exempt from the requirements of human subject protection regulations. We used the subscriber-to-member relationships in the insurance claims data to identify 38,846 families with at least one child diagnosed with ASD, indicated by the presence of one or more ICD-9 codes in the 299 group (pervasive developmental disorders) in at least one medical claim. Fathers, mothers and their affected children were matched to control populations by age, sex and zipcode, as a proxy to socioeconomic status. The large control populations were repeatedly subsampled ($n=10{,}000$) to compare the prevalence of comorbid diagnoses in equally sized samples, and the *P* value of the median statistic for each diagnostic category was taken as the representative association between that diagnostic group and the case population. In addition to case–control comparisons, we also identified 23,837 families with at least one child diagnosed with ASD (ICD-9 code 299.x) and at least one child lacking any 299.x diagnosis.

ICD-9 codes were rolled up to phenome-wide association study (PheWAS) groups[51], and dyslipidemia was defined as having any of the following diagnoses: hyperlipidemia (PheWAS code (PheCode) 272.1), mixed hyperlipidemia (PheCode 272.13), hypercholesterolemia (PheCode 272.11), lipoid metabolism disorder not otherwise specified (PheCode 272.9), lipoprotein disorders (PheCode 277.51), other disorders of lipoid metabolism and hyperalimentation (PheCode 277.5), and disorders of lipoid metabolism (PheCode 272). Fisher's exact tests were used to assess the strength and significance of all diagnostic associations.

We further assessed dyslipidemia diagnoses, defined as above, in parents of at least one child with ASD, compared with age- and sex-matched parents of children not diagnosed with ASD. Comparisons were performed using Fisher's exact tests.

We also compared drug prescription records and comorbid diagnoses between individuals with ASD and dyslipidemia, and individuals with ASD and no dyslipidemia, identified as described above. The National Drug Code directory was used to map National Drug Codes to established pharmacological classes and physiological effects, thereby identifying prescriptions of atypical antipsychotic, antiepileptic, antidiabetic (including gliptins, glitazones, biguanides, sulfonylureas, glinides, α-glucosidase inhibitors, glucagon-like peptide-1 agonists, sodium–glucose transport protein 2 inhibitors, insulin analogs and amylin analogs) and lipid-lowering drugs (including statins, cholesterol absorption inhibitors and peroxisome proliferator-activated receptor α agonists). Data were stored on SQL Server 2017 on Windows Server 2012 R2 Datacenter and accessed through Microsoft R Open v.3.3.3. The PheWAS R package v.0.12 was used to map ICD-9 codes to PheCodes as described in Analysis of healthcare claims data above. In all, 6,621,118 individuals with ASD and/or dyslipidemia had sufficient data, which were comparatively analyzed using age- and sex-adjusted, generalized linear models.

Confounder analyses. We examined whether IQ or sex might be related to LGD variants in lipid metabolism genes in individuals with ASD. First, we examined the relationships between IQ and the number of deleterious alleles in lipid metabolism genes (Supplementary Fig. 2). We found no differences in IQ between individuals in the bottom and top deciles of the number of LGD variants in lipid metabolism genes (the mean IQ of individuals with ASD in the bottom decile of deleterious variants in lipid metabolism genes was 79.63, the mean IQ of individuals with ASD in the top decile was 81.60, and the 95% CI of the difference of the means was (−6.64, 2.7), Student's *t*-test, $P=0.41$). Similarly, we compared the relationships between IQ and the number of lipid-metabolism genes hit by one or more LGD variants (Supplementary Fig. 3). Consistent with the previous result, we found no differences in IQ between individuals in the bottom and top deciles of the number of lipid metabolism genes with LGD alleles (the mean IQ of individuals with ASD in the bottom decile was 79.55, the mean IQ of individuals with ASD in the top decile was 80.62 and the 95% CI of the difference of the means was (−5.98, 3.85), Student's *t*-test, $P=0.67$). We then examined the sex distributions according to the number of LGD variants in lipid metabolism genes (Supplementary Fig. 4), as well as the number of lipid metabolism genes hit by one or more LGD variants (Supplementary Fig. 5). We found no sex differences between individuals in the bottom and top deciles of the number of LGD alleles in lipid metabolism genes (85.9% males in the bottom decile versus 86.2% in the top decile; OR = 0.967, 95% CI = (0.58, 1.59), Fisher's $P=0.97$). Similarly, no sex bias was detected between individuals in the bottom and top deciles of the number of lipid metabolism genes hit by one or more LGD variants (89.7% males in the bottom decile versus 86.9% males in the top decile; OR = 1.31, 95% CI = (0.73, 2.42), Fisher's $P=0.41$).

Furthermore, we used the healthcare claims data to examine relationships between dyslipidemia and exposure to commonly prescribed drugs in ASD that are thought to alter lipid levels. For this purpose, we used data from 6,621,118 individuals with ASD and/or dyslipidemia and at least 2 years of coverage. We repeated the analyses described above on the subpopulation of individuals with no record of atypical antipsychotic, anticonvulsant or antidiabetic drug prescriptions.

Similarly, we stratified the BCH EHR data on drug use, and compared blood lipid profiles among individuals with no prescription records for atypical antipsychotics, anticonvulsants or antidiabetics. We further stratified these data on metabolic state, and found that, among individuals with metabolic dysregulation, defined as a record of diabetes, obesity and/or metabolic syndrome X diagnoses, lipid profiles were significantly different in those with comorbid ASD (Fig. 2e and Supplementary Figs. 6 and 7).

Analysis of mouse phenotypes. To examine the functional overlap between ASD and dyslipidemia genes we compared all reported phenotypes of mouse models with targeted mutations in bona fide ASD and dyslipidemia genes. We also included models of SLOS as a gold standard intermediate, because SLOS is a rare syndrome of congenital malformations and intellectual disability caused by a cholesterol biosynthesis defect, with most cases meeting the diagnostic criteria for ASD[16].

We focused only on genes annotated by the authoritative OMIM compendium (https://www.omim.org) as ASD, dyslipidemia or SLOS genes. Their mapping to mouse genes was obtained from http://www.informatics.jax.org/downloads/reports/MGI_GeneOMIM.rpt. In all, we identified 34 ASD genes, 10 dyslipidemia genes and 1 SLOS gene implicated in the respective conditions by OMIM and modeled by at least one mouse. We next obtained the phenotypes of all mouse models with targeted mutations in these genes, using http://www.informatics.jax.org/downloads/reports/MGI_GenePheno.rpt, and mapped phenotype IDs to their descriptions using http://www.informatics.jax.org/downloads/reports/VOC_MammalianPhenotype.rpt. For each modeled gene, a phenotype was considered present if at least one mouse model of a targeted disruption to that gene had the phenotype reported in at least one publication. The results are summarized in a Boolean matrix in Supplementary Table 2. This matrix was clustered to understand the overall phenotypic similarities between ASD and dyslipidemia gene disruptions in an unsupervised manner. Both hierarchical and *k*-means clustering were used. For hierarchical clustering, dendextend v.1.13.4 and circlize v.0.4.9 were used in RStudio v.1.3.959 (https://cran.r-project.org/web/packages/dendextend/index.html and https://cran.r-project.org/web/packages/circlize/index.html, respectively). In addition, factoextra v.1.0.7 (https://cran.r-project.org/web/packages/factoextra/index.html) was used for *k*-means clustering and principal component analysis-based visualizations.

Statistical analysis. ASP analysis[42] was used to assess the significance of multiplex family variant sharing, and permutation tests[41] were used to assess the increased burden of neurodevelopmentally co-regulated, sex-differentially expressed LGD variation in probands compared with their unaffected siblings. These sources of evidence were combined using Fisher's method[44]. Limma[37] was used for sex-differential expression analysis. GSEA[46] was used for functional enrichment analysis. Student's *t*-tests were used to compare the number of LGD mutations in lipid-related genes with IQ and age, and Fisher's exact test was used to examine the association between LGD mutations in lipid-related genes and sex. Fisher's exact tests were further used to examine the association between ASD and altered lipid profiles, as well as the association between ASD and dyslipidemia diagnoses. Kruskal–Wallis tests were used to compare fasting triglycerides, total cholesterol and LDL-cholesterol in individuals with ASD and age-, gender- and metabolic state-matched controls. Generalized linear models were used to examine comorbid diagnoses in dyslipidemia-associated ASD, while controlling for age and sex.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Familial WES datasets can be obtained from https://ndar.nih.gov as Collections 1918, 2004 and 2042. The human neurodevelopmental transcriptome dataset is available at http://www.brainspan.org/api/v2/well_known_file_download/267666524. Functional annotations can be obtained from ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2go.gz and https://www.gsea-msigdb.org/gsea/downloads.jsp. EHRs and healthcare claims data used in the present study are not publicly available due to patient privacy concerns. Mouse phenotypes are available at http://www.informatics.jax.org/downloads/reports/MGI_GenePheno.rpt.

Code availability

The code used in the present study is available at https://github.com/yuanluo/autism_precision_medicine.

References

30. Kang, H. J. et al. Spatio-temporal transcriptome of the human brain. *Nature* **478**, 483–489 (2011).
31. Zhang, M. et al. Axonogenesis is coordinated by neuron-specific alternative splicing programming and splicing regulator PTBP2. *Neuron* **101**, 690–706 e610 (2019).
32. Su, C. H., D., D. & Tarn, W. Y. Alternative splicing in neurogenesis and brain development. *Front. Mol. Biosci.* **5**, 12 (2018).
33. Everitt, B. S. *The Cambridge Dictionary of Statistics* (Cambridge Univ. Press, 2006).
34. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *Inter Journal, Complex Systems*, 16951704 (2006).
35. Rosvall, M. & Bergstrom, C. T. Maps of random walks on complex networks reveal community structure. *Proc. Natl Acad. Sci. USA* **105**, 1118–1123 (2008).
36. Cleveland, W. S. & Devlin, S. J. Locally weighted regression: an approach to regression analysis by local fitting. *J. Am. Stat. Assoc.* **83**, 596–610 (1988).
37. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47–e47 (2015).
38. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.* **57**, 289–300 (1995).
39. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
40. Krumm, N. et al. Excess of rare, inherited truncating mutations in autism. *Nat. Genet.* **47**, 582–588 (2015).
41. Noreen, E. W. *Computer-Intensive Methods for Testing Hypotheses: An Introduction* (Wiley, 1989).
42. Neale, B., Ferreira, M. & Medland, S. *Statistical Genetics* (Taylor & Francis Group, 2012).
43. Lawrence, M. et al. Software for computing and annotating genomic ranges. *PLoS Comput. Biol.* **9**, e1003118 (2013).
44. Fisher, R. A. *Statistical Methods For Research Workers* (Cosmo Publications, 1925).
45. Dunn, O. J. Multiple comparisons among means. *J. Am. Stat. Assoc.* **56**, 52–64 (1961).
46. Subramanian, A. et al. Gene Set Enrichment Analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
47. Bateni, M. et al. Affinity clustering: hierarchical clustering at scale. *Adv. Neural Inf. Process. Syst.* **2017**, 6864–6874 (2017).
48. Kohane, I. S., Churchill, S. E. & Murphy, S. N. A translational engine at the national scale: informatics for integrating biology and the bedside. *J. Am. Med. Inform. Assoc.* **19**, 181–185 (2012).
49. Medicode. *ICD-9-CM: International Classification of Diseases, 9th Revision, Clinical Modification* (Medicode, 1996).
50. World Health Organization. *ICD-10: International Statistical Classification of Diseases and Related Health Problems* (World Health Organization, 2004).
51. Denny, J. C. et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat. Biotechnol.* **31**, 1102–1110 (2013).

Acknowledgements

Data analyzed in this manuscript reside in the National Institutes for Health (NIH)-supported NIMH Data Archive's NDAR as Collection nos. 1918, 2004 and 2042. We thank all the families at the participating Simons Simplex Collection sites, as well as the principal investigators (A. Beaudet, R. Bernier, J. Constantino, E. Cook, E. Fombonne, D. Geschwind, R. Goin-Kochel, E. Hanson, D. Grice, A. Klin, D. Ledbetter, C. Lord, C. Martin, D. Martin, R. Maxim, J. Miles, O. Ousley, K. Pelphrey, B. Peterson, J. Piggot, C. Saulnier, M. State, W. Stone, J. Sutcliffe, C. Walsh, Z. Warren and E. Wijsman). We thank SFARI Base for access to their phenotypic data. Approved researchers can obtain the Simons Simplex Collection population dataset described in the present study (https://base.sfari.org/ordering/phenotype/sfari-phenotype/download?code=11) by applying at https://base.sfari.org. We thank J. Eichler, D. Margulies and members of the Kohane lab for fruitful discussions. We thank somersault18:24 (www.somersault1824.com) for illustrations. Y.L. was supported by the US National Institutes of Health (1R21LM012618 and 5UL1TR001422). A.E., P.S. and I.S.K. were supported by the National Institute of Mental Health (P50MH106933). A.E. was supported by the Israeli Ministry of Science and Technology (grant no. 17708) and by the PrecisionLink Initiative at BCH. N.P. received funding support from Aetna Life Insurance Co. P.A. was supported by the US National Institutes of Health (U01HG007530, OT3OD025466, OT3HL142480, U54HG007963, 1U01TR002623-01 and 1U54HD090255-01).

Author contributions

Y.L., A.E. and I.S.K. designed the study and wrote the manuscript. Y.L., A.E., N.P., P.A. and A.L.-M. performed the analyses. P.S. and I.S.K. supervised the study. All authors contributed to the interpretation of the data.

Competing Interests

The authors declare no competing interests.

Additional information

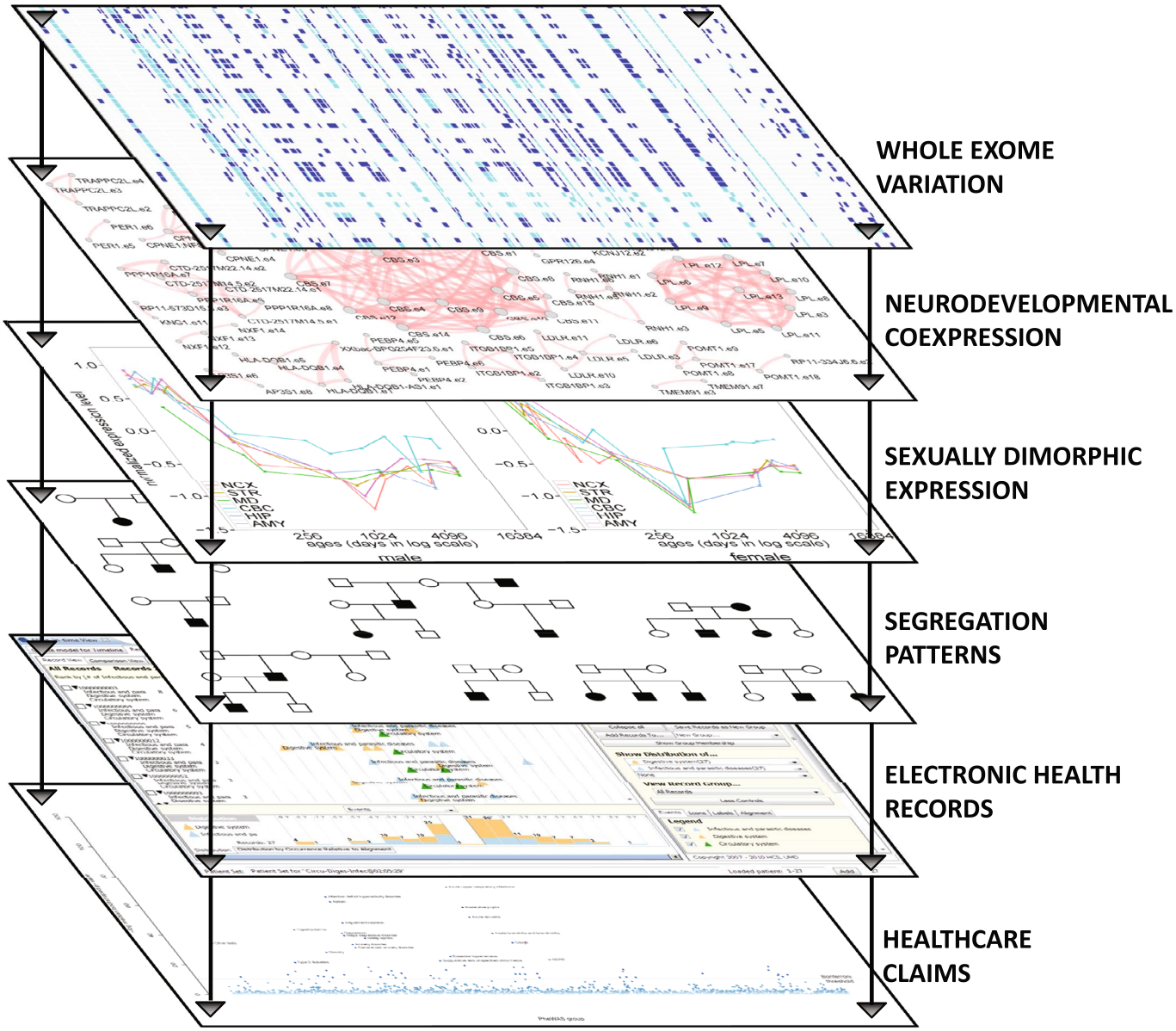
Extended data is available for this paper at https://doi.org/10.1038/s41591-020-1007-0.

Supplementary information is available for this paper at https://doi.org/10.1038/s41591-020-1007-0.

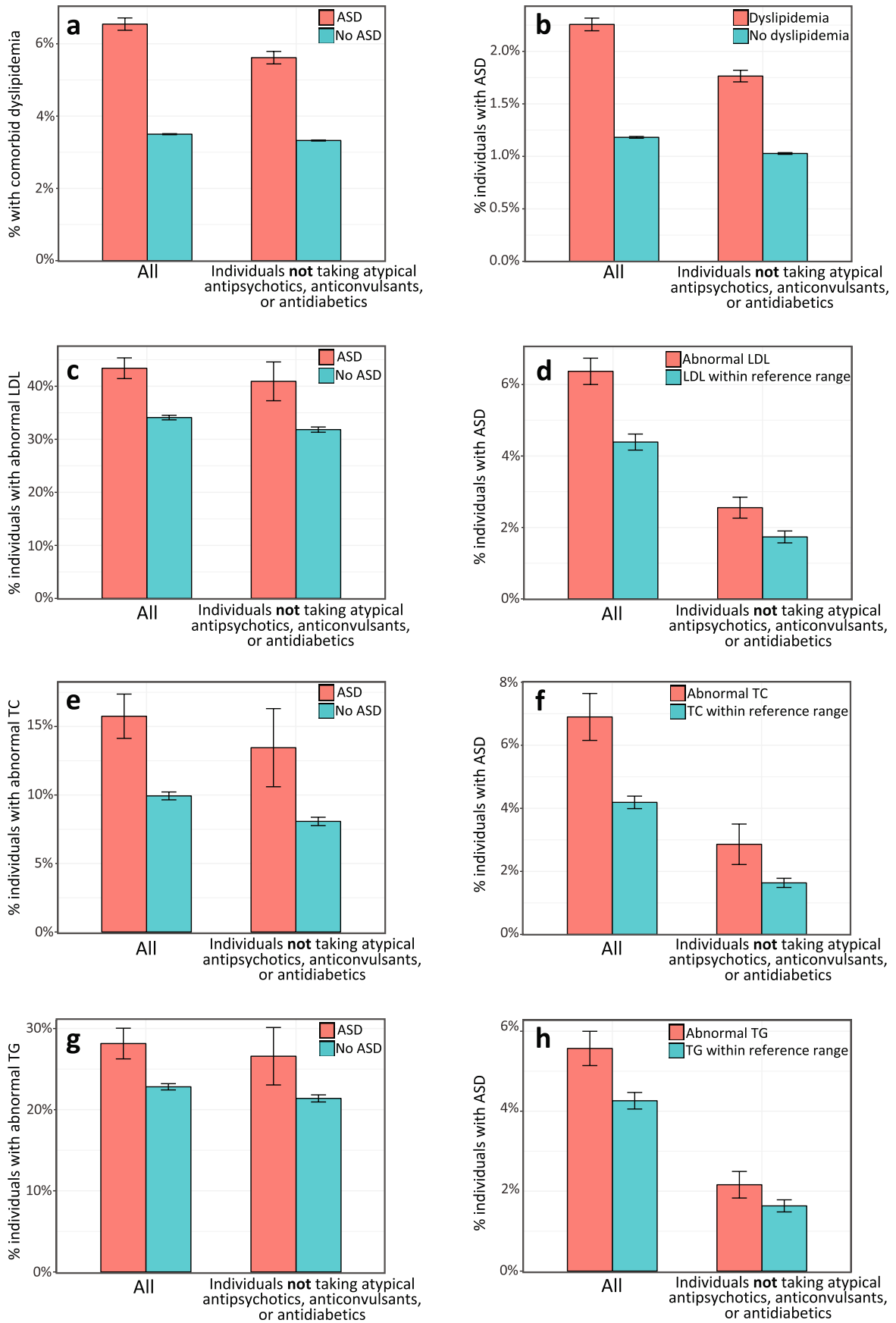
Correspondence and requests for materials should be addressed to I.S.K.

Peer review information Kate Gao was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Reprints and permissions information is available at www.nature.com/reprints.

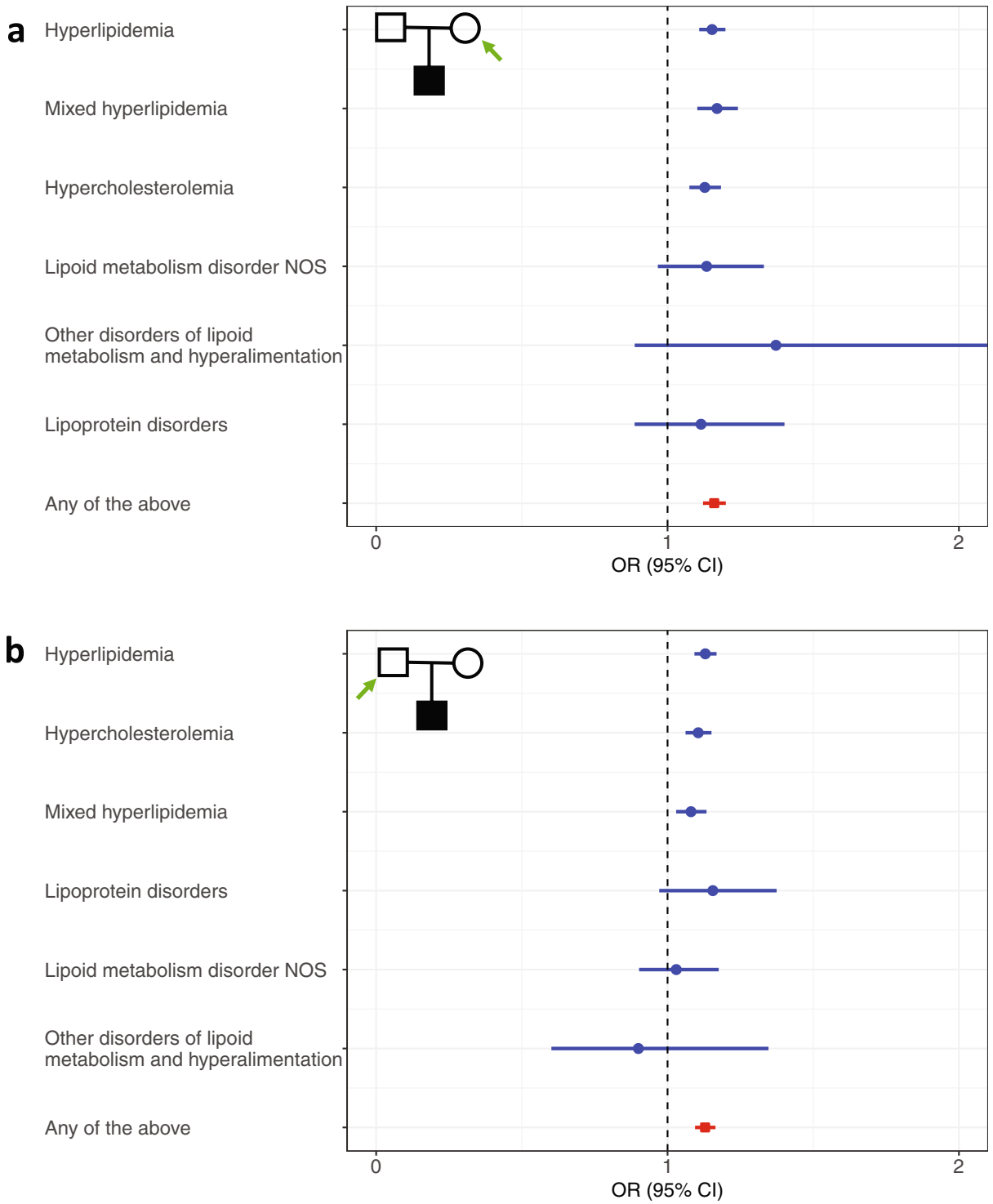


Extended Data Fig. 1 | Data modalities integrated in the present study.

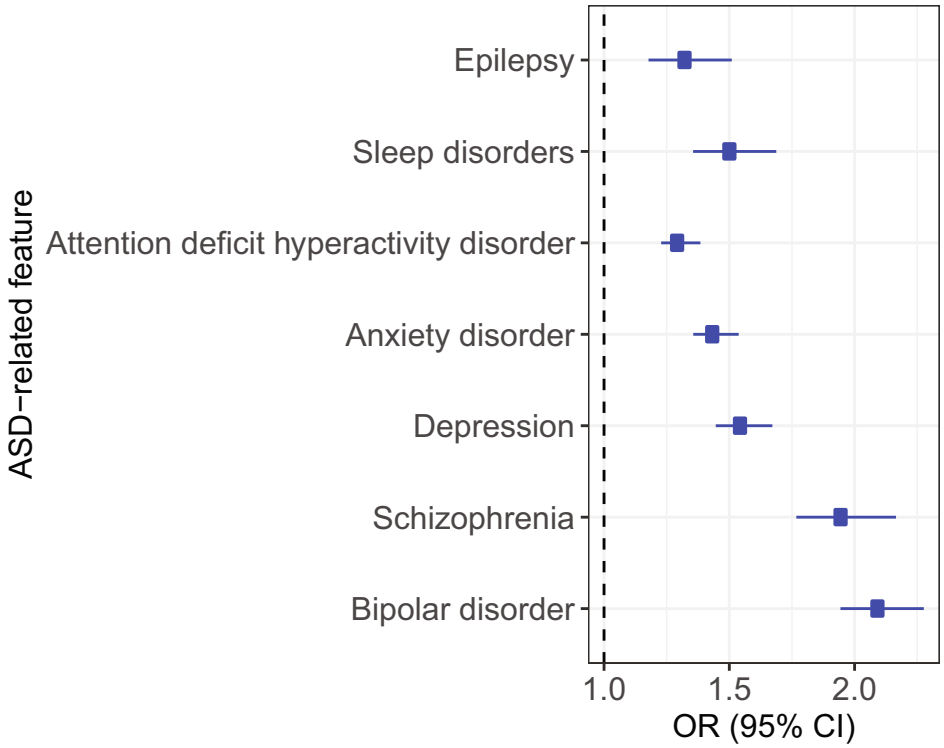


Extended Data Fig. 2 | See next page for caption.

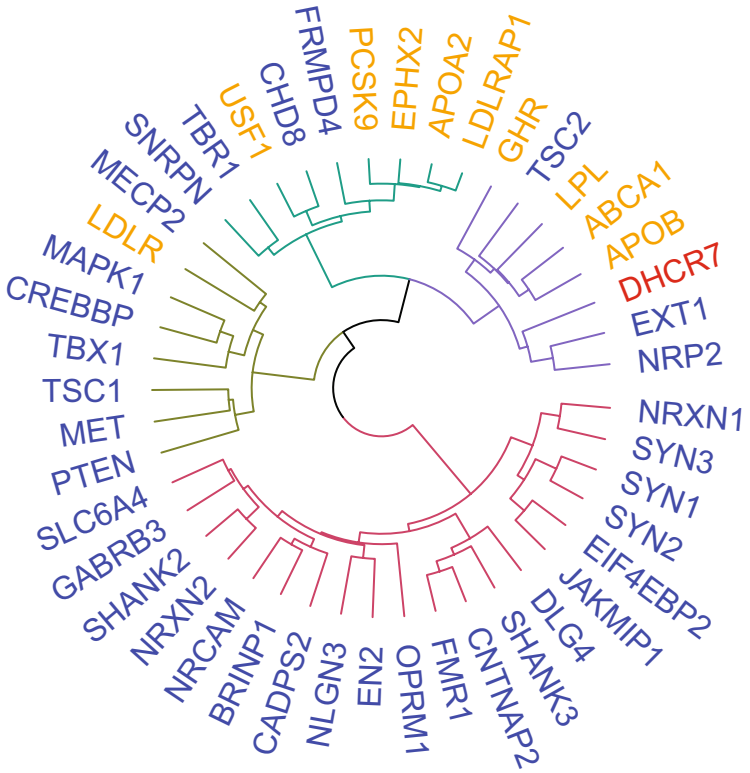
Extended Data Fig. 2 | Association between ASD and dyslipidemia, stratified by drug use. In addition to examining the relation between ASD and dyslipidemia in the entire cohorts, we also restricted our analyses to individuals with no prescription records of drugs commonly prescribed in ASD which are known to alter lipid levels, namely atypical antipsychotics, anticonvulsants, and antidiabetics. Error bars indicate the 95% CIs for the proportions. **a**, Rates of dyslipidemia diagnoses in individuals with ASD (red) and individuals with no ASD diagnosis (cyan), stratified by drug use (entire cohort OR = 1.93, 95% CI = (1.88, 1.99), Fisher's exact two-sided $P < 1 \times 10^{-323}$, $n = 6{,}621{,}118$ individuals; individuals not taking atypical antipsychotics, anticonvulsants, or antidiabetics OR = 1.73, 95% CI = (1.67, 1.79), Fisher's exact two-sided $P = 1.11 \times 10^{-201}$, $n = 6{,}488{,}315$). **b**, Rates of ASD diagnoses in individuals with dyslipidemia (red) and individuals with no dyslipidemia diagnosis (cyan), stratified by drug use (effect sizes as in a). **c**, Fraction of individuals with abnormal fasting LDL levels out of individuals with ASD and at least one fasting LDL test result (red), and individuals with no ASD diagnosis and at least one fasting LDL test result (cyan), stratified by drug use (entire cohort OR = 1.48, 95% CI = (1.36, 1.61), Fisher's exact two-sided $P = 1.06 \times 10^{-20}$, $n = 48{,}775$ individuals; individuals not taking atypical antipsychotics, anticonvulsants, or antidiabetics OR = 1.48, 95% CI = (1.27, 1.73), Fisher's exact two-sided $P = 6.16 \times 10^{-7}$, $n = 34{,}751$ individuals). **d**, Fraction of individuals with ASD out of individuals with abnormal fasting LDL (red), and individuals with all fasting LDL test results within the reference range (cyan), stratified by drug use (effect sizes as in c). **e-f**, Same as c-d but for fasting total cholesterol (TC). Entire cohort OR = 1.69, 95% CI = (1.49, 1.92), Fisher's exact two-sided $P = 7.14 \times 10^{-15}$, $n = 43{,}650$ individuals; individuals not taking atypical antipsychotics, anticonvulsants, or antidiabetics OR = 1.77, 95% CI = (1.36, 2.27), Fisher's exact two-sided $P = 2.00 \times 10^{-5}$, $n = 31{,}690$ individuals. **g-h**, Same as c-d but for fasting triglycerides (TG). Entire cohort OR = 1.33, 95% CI = (1.20, 1.46), Fisher's exact two-sided $P = 1.73 \times 10^{-8}$, $n = 47{,}650$ individuals; individuals not taking atypical antipsychotics, anticonvulsants, or antidiabetics OR = 1.33, 95% CI = (1.10, 1.60), Fisher's exact two-sided $P = 2.99 \times 10^{-3}$, $n = 39{,}165$ individuals.



Extended Data Fig. 3 | Enrichment of dyslipidemia diagnoses in parents of children with ASD (maternal OR = 1.16, 95% CI = (1.12, 1.20), Fisher's exact two-sided $P = 5.28 \times 10^{-18}$; paternal OR = 1.13, 95% CI = (1.09, 1.16), Fisher's exact two-sided $P = 1.92 \times 10^{-14}$; $n = 38{,}846$ families vs. repeatedly resampled matched controls from a total of 34,003,107 individuals). a, Association between maternal dyslipidemia and having a child with ASD. Shown is a forest plot detailing diagnosis-specific ORs by circles and their 95% CIs by horizontal lines. **b**, Association between paternal dyslipidemia and having a child with ASD. A diagnosis-specific forest plot is shown as in (**a**).



Extended Data Fig. 4 | Core ASD-related features associated with dyslipidemia in ASD. A forest plot depicts the association estimates for ASD-related clinical characteristics more common in individuals with ASD and dyslipidemia, as compared to individuals with ASD and no dyslipidemia (n = 80,714 individuals). ORs and their 95% CIs are shown by circles and horizontal lines, respectively.



Extended Data Fig. 5 | Phenotypic clustering of ASD (blue), dyslipidemia (orange), and SLOS (red) mouse models. Hierarchical clustering of 1,315 phenotypes measured in ASD ($n = 34$), dyslipidemia ($n = 10$), and SLOS ($n = 1$) mouse models identified four clusters. Three clusters (shown on top) include both dyslipidemia and ASD mice, with shared phenotypes such as seizures, abnormal synapse morphology, abnormal learning, abnormal brain size, and abnormal coordination. The fourth cluster (bottom) is ASD-specific. Thus, some ASD models are more similar to dyslipidemia models than to other ASD mice.

nature research

Corresponding author(s): Alal Eran, Isaac Kohane

Last updated by author(s): Jun 24, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Genomic data was compiled using the Genome Analysis Toolkit (GATK) version 3.3, as well as R 3.4 and Python 2.7 scripts. All code is available at https://github.com/yuanluo/autism_precision_medicine. Transcriptomic data was collected using Python 2.7 scripts, available at https://github.com/yuanluo/autism_precision_medicine. Electronic health records were accessed through the i2b2/tranSMART platform version 1.7.09. Healthcare claims data were stored on SQL Server 2017 and accessed using Microsoft R Open 3.3.3.
Data analysis	Custom R 3.4 and Python 2.7 code used for data analysis is available at https://github.com/yuanluo/autism_precision_medicine. The R igraph package version 1.2.0 was used for clustering co-expressed exons. The R stats package version 4.1.0 was used for locally weighted scatterplot smoothing. Limma version 3.42.2 was used for sex-differential expression analysis. Picard version 1.119 was used for BAM processing, and Samtools version 1.1 was used to shuffle BAM files and convert them back to fastq. BWA-MEM version 0.7.10 was used for aligning whole exome sequence fastqs to the hg19 reference human genome. Picard version 1.119 and GATK 3.3 were used for joint variant calling. The GenomicRanges toolkit version 1.30.3 was used for mapping variants to exon clusters. ANNOVAR version 2018Apr16 was used for variant annotations, as was gnomAD v2.1. Gene Set Enrichment Analysis (GSEA) version 4.0.3 was used for functional enrichment analysis. The PheWAS R package version 0.12 was used to map ICD9 codes to PheCodes. The dendextend R package version 1.13.4 and circlize R package version 0.4.9 were used in RStudio version 1.3.959 for clustering mouse phenotypes. The factoextra R package version 1.0.7 was used for K means clustering and visualization.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Familial whole exome sequence data can be obtained from https://ndar.nih.gov as Collections 1918, 2004, and 2042. The human neurodevelopmental transcriptome dataset is available at http://www.brainspan.org/api/v2/well_known_file_download/267666524. Functional annotations can be obtained from ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2go.gz and https://www.gsea-msigdb.org/gsea/downloads.jsp. Electronic health records and healthcare claims data used in this study are not publicly available due to patient privacy concerns. Mouse phenotypes are available at http://www.informatics.jax.org/downloads/reports/MGI_GenePheno.rpt

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No sample size calculations were performed, rather all available data was used. Post hoc power calculations showed that all analyses achieved power > 0.95.
Data exclusions	All data exclusion criteria were predetermined. In the familial whole exome data, siblings with significantly different sequence coverage were excluded, as determined by the GATK CallableLoci analysis. This was intended to avoid confounding by differential sequence coverage. In the healthcare claims and medical records data, individuals with duration of coverage spanning < 24 months were excluded. The rationale behind this was to ensure the availability of sufficient time for the detection of associated diagnoses or lab test results.
Replication	Two attempts at replication were successful, using orthogonal data types and multiple approaches. The reproducibility of our study was further enhanced by bootstrapping all the association tests performed in massive healthcare claims data 10,000 times.
Randomization	Mendelian randomization was the basis of the familial whole exome sequence analyses conducted. For other data types, individuals and samples were matched to optimally control for covariates, based on data availability. Specifically, in the transcriptomic data analysis, samples were matched based on brain region, age, sex, RNA integrity number (RIN), and postmortem interval (PMI). In the healthcare claims data, individuals were matched based on age, sex, duration of coverage, and zipcode as a proxy for socioeconomic status and access to care. In the institutional health records data, individuals were matched based on age, sex, availability of test results, drug use, and metabolic status.
Blinding	Blinding was not relevant to this study as it was based on population-level summary statistics.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging