Exploring Polygenic Contributors To Subgroups of Comorbid Conditions In Autism Spectrum Disorder

Louis Klein
University of New South Wales
Shannon D'Urso
University of Queensland
Valsamma Eapen
University of New South Wales
Liang-Dar Hwang
University of Queensland
Ping-I Lin (daniel.lin@unsw.edu.au)
University of New South Wales

Research Article

Keywords: ASD, comorbid, AGRE, pain

Posted Date: October 5th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-948090/v1

Version of Record: A version of this preprint was published at Scientific Reports on March 1st, 2022. See the published version at https://doi.org/10.1038/s41598-022-07399-7.

Exploring polygenic contributors to subgroups of comorbid conditions in autism spectrum disorder

Louis Klein[1,2,α], **Shannon D'Urso**[3,α], **Valsamma Eapen**[1,2], **Liang-Dar Hwang**[3,β], **and Ping-I Lin**[1,2,*,β]

[1]School of Psychiatry, University of New South Wales, Sydney NSW, Australia
[2]Mental Health Research Unit, Liverpool Hospital, SWS-LHD NSW, Australia
[3]Institute for Molecular Bioscience, University of Queensland, Brisbane QLD, Australia
[*]daniel.lin@unsw.edu.au
[α]these authors contributed equally to this work
[β]joint senior authors

ABSTRACT

Individuals with autism spectrum disorder (ASD) have heterogeneous comorbid conditions. This study examined whether comorbid conditions in ASD are associated with polygenic risk scores (PRS) of ASD or PRS of comorbid conditions in non-ASD specific populations. Genome-wide single nucleotide polymorphism (SNP) data were obtained from 1,386 patients with ASD from the Autism Genetic Resource Exchange (AGRE) study. After excluding individuals with missing clinical information concerning comorbid conditions, a total of 707 patients were included in the study. A total of 18 subgroups of comorbid conditions ('topics') were identified using a machine learning algorithm, topic modeling. PRS for ASD were computed using a genome-wide association meta-analysis of 18,381 cases and 27,969 controls. From these 18 topics, Topic 6 (over-represented by allergies) (p = 1.72x10^{-3}) and Topic 17 (over-represented by sensory processing issues such as low pain tolerance) (p = .037) were associated with PRS of ASD. For associated topics, we further assessed their associations with PRS of their corresponding comorbid conditions based on non-ASD specific populations. However, these two topics were not associated with the PRS of allergies and chronic pain disorder, respectively. Characteristics of the present AGRE sample and those samples used in the original GWAS for ASD, allergies, and chronic pain disorder, may differ due to significant clinical heterogeneity that exists in the ASD population. Additionally, the AGRE sample may be underpowered and therefore insensitive to weak PRS associations due to a relatively small samplesize. Findings imply that susceptibility genes of ASD may contribute more to the occurrence of allergies and sensory processing issues in individuals with ASD, compared with the susceptibility genes for their corresponding phenotypes. Since these comorbid conditions (i.e., allergies and pain sensation issues) cannot be attributable to the corresponding comorbidity-specific biological factors in non-ASD individuals, clinical management for these comorbid conditions may still depend on treatments for core symptoms of ASD.

Introduction

Psychiatric and medical comorbidities are a norm rather than an exception in autism spectrum disorder (ASD); a complex neurodevelopmental disorder characterized by social communication deficits and restricted/repetitive behaviors.[1] The importance of understanding medical comorbidities of ASD cannot be understated.[2] Appropriate management of comorbid medical conditions may lead to quality of life improvement for both children and their parents.[3] In this regard, understanding of shared etiologies—including potential genetic factors[4,5]—for ASD and comorbid conditions may be critical in management decisions. While research into the medical comorbidities of ASD has been ongoing[2], research into the genetic bases of ASD's psychiatric comorbidities is only now getting underway.[6–8] Understanding how genetic factors contribute to the comorbidities may provide novel insight into molecular mechanisms underlying heterogeneous clinical features of individuals with ASD. Such genetic components could be used to subgroup patients with ASD to generate clinical subtypes that reflect biological differences, which might provide opportunities for individualized treatment options.[9]

Prior studies that have investigated genetic contributions for comorbidities in ASD have implemented several different approaches. For instance, David and colleagues (2016) conducted an automated extraction of genes associated with ASD and its comorbid disorders, finding 1,031 genes associated with ASD—262 of these genes were involved in ASD only—while the remaining 779 genes were also associated with other comorbid disorders.[10] Their study results suggest that the majority of candidate genes for ASD have pleiotropic effects. Diaz-Beltran and colleagues (2017) used a two-fold systems biology approach to perform a comparative analysis of ASD with 31 frequently encountered comorbid disorders and determined a multi-comorbidity subtype of ASD, which led to the discovery of novel candidate genes of ASD.[11] Tylee and colleagues

(2018) used data from previous genome-wide association studies (GWAS) to determine whether commonly varying single nucleotide polymorphisms (SNPs) are shared between psychiatric and immune-related phenotypes, and found that ASD is most likely to correlate with allergy rather than all other major psychiatric disorders.[12] There is also evidence to suggest that ASD is often comorbid with the perception of pain, with ASD patients having a higher threshold of pain.[13,14] In this regard, Johnson and colleagues (2019) conducted a linkage disequilibrium score regression using GWAS associations for ASD and chronic pain to show that the two traits are genetically correlated.[15] Finally, a recent study used polygenic risk scores (PRS) derived from five psychiatric disorders, such as schizophrenia, major depressive disorder, attention deficit hyperactivity disorder, obsessive-compulsive disorders, and anxiety, and found that the polygenic contributions could distinguish Asperger syndrome (a diagnostic category in the Diagnostic and Statistical Manual (DSM) 4^{th} Edition although this term is no longer used in the updated DSM-5) from individuals with other non-Asperger subtypes of ASD.[16,17]

Despite these advances in understanding the relationships between the genetic bases for ASD and associated comorbidities, it is clear that these approaches require further development to fully understand the role of comorbid genetic risk within ASD.

In the present study, we adapted a recent approach implemented by McCoy and colleagues (2017) using topic modeling—a spatial clustering process tolerant of feature sparsity (for e.g., diagnostic features in low prevalence medical conditions)—to identify clusters of comorbid conditions in individuals with major depressive disorder to investigate which types of comorbid conditions are attributable to polygenic loading of major depressive disorder.[18] These authors observed that using standard phenome-wide association studies (PheWAS) to test across all possible predictors such as individual risk variants or genome-wide variants increases type I error as the process involves testing against all diagnostic codes (e.g. 1500+ codes), or type II error as the approach needs to correct for all these codes. This approach also does not take into consideration the correlation between individual diagnosis codes and the inconsistency and lack of reliability of the codes per se. By first using topic modeling on a corpus of ICD-9 diagnostic codes, McCoy et al were able to reduce the dimensionality of possible MDD comorbidity and thereby also mitigate risk of imprecision when examining associations with genetic indicants. However, it is important to note several limitations to this approach including that it cannot distinguish whether depression and associated co-morbidities were caused by shared versus unique but convergent genetic factors. Given that the number of topics was chosen arbitrarily, it is likely that an alternative number of topics would yield different results in terms of cluster composition. Moreover, it is possible to optimize the selection of the number of topics to cluster upon using a range of semantic and statistical indices developed by researchers in the applied topic modeling literature.

The present study extends the topic modeling approach applied by McCoy and colleagues in three ways: 1) We introduced a novel method for constructing document data from epidemiological data; 2) the number of topics was determined by a semi-unsupervised process that maximizes the trade-off between topic sensitivity and term specificity; 3) we jointly investigated associations between topics of ASD and three sets of PRSs of ASD and its clinically co-morbid conditions of interest, namely, chronic pain and allergies, respectively.

Results

We found a strong association between PRS_{ASD} and the Topic 6, the main allergy factor (the first 5 terms: past-diet, allergies, gluten-free, abnormal-gastro and casein-free) (Figure 2). The strongest associated PRS_{ASD} trench S5 ($p = 1.72\text{x}10^{-3}$) accounted for 1.41% of the variance in the Topic 6. We also found that Topic 17 (first 5 terms: *handed-right*, *low-acoustic*, *low-pain*, *low-tactile*, *abnorm-skin*, *diet-preferences*; minimum $p = 0.037$ for S4) accounted for 0.80% of the variance. Topic 14 was over-represented by maternal substance use and hence were excluded for the subsequent analyses (See Additional File 4 for full results).

We further assessed the association between Topic 6 and $PRS_{allergy}$ and found no evidence for an association (minimum $p = 0.083$ for $PRS_{allergy}$ S2). When fitting both PRS_{ASD} and $PRS_{allergy}$ into the model, the strengths of the association for the two PRSs slightly increased (PRS_{ASD} S5: $p = 1.48\text{x}10^{-3}$; $PRS_{allergy}$ S2: $p = 0.070$), but the overall model fit did not improve compared against the PRS_{ASD} only model ($p_{ANOVA} = 0.070$). Similarly, no evidence was found for an association between the Topic 17 and PRS_{pain} (minimum $p = 0.052$ for S5). However, combining PRS_{ASD} and PRS_{pain} improved the prediction of Topic 17 ($p_{ANOVA} = 0.031$; the two PRSs together accounts for 1.8% of the variance) and boosted the strength of associations for each of the two PRSs (PRS_{ASD} S4: $p = 0.022$; PRS_{pain} S5: $p = 0.031$). Overall, PRS_{ASD} contributed to a greater degree to the variance in topic 6 than $PRS_{allergy}$, and PRS_{ASD} contributed to a greater degree to the variance in Topic 17 than PRS_{pain} (see Figure 3).

We have identified 681 candidate genes for ASD, 374 candidate genes for allergies, and 346 candidate genes for pain-related disorder (e.g., abnormality in pain sensation or impaired pain sensation) (see Figure 4). There were 59 overlapping genes between the lists of candidate genes for ASD and allergies, and 135 overlapping genes between the lists of candidate genes for ASD and pain-related disorder. The p-value for the correlation between ASD and allergies was $1.19\text{x}10^{-27}$, while the p-value for the correlation between ASD and pain-related disorder was $1.11\text{x}10^{-117}$.

Discussion

The current study shows that polygenic loading in ASD may play a role in certain subgroups of comorbid conditions in ASD, such as allergy-related conditions, and sensory processing issues (e.g., low pain tolerance). We further found that these two subgroups (i.e. topics) of comorbid conditions could not be attributed to PRS of either allergies or chronic pain disorder. These findings suggest that ASD-associated genetic variants could contribute to ASD-related comorbidities including allergies and sensory processing issues such as low pain tolerance. In other words, these two types of comorbidities may share a proportion of risk variants with ASD. Notably, we found that both of these two subgroups (i.e. Topics) of comorbid conditions in individuals with ASD were not correlated with their corresponding PRS constructed using GWAS of non-ASD specific individuals.

Nevertheless, although both pain and pain sensation issues may share susceptibility genes with ASD,[19,20] our findings suggest that ASD may be more likely to correlate with abnormality in pain sensation at the gene level compared to allergies. This is consistent with our findings that PRS_{pain} might contribute, at least to a slightly higher degree, to the corresponding topic enriched with sensory processing issues than that of allergies and that combining PRS_{ASD} and PRS_{pain} strengthened the association with the corresponding comorbid conditions. These findings, to our knowledge, have not yet been published by other studies.

The current study has several limitations. First, characteristics of the present AGRE sample and those samples used in the original GWAS for ASD, allergies, and chronic pain disorder, may differ due to significant clinical heterogeneity that exists in the ASD population. This is further compounded by the significant changes in the diagnostic classification that has occurred in the past decade that would have potentially changed the ascertainment and phenotypic characterization of ASD. For example, in 2013 the Diagnostic and Statistical Manual 5^{th} edition combined the different subcategories such as Asperger Syndrome and Disintegrative Disorder etc. and collapsed it into a single category of ASD and further changes were made to the diagnostic criteria such as the inclusion of sensory issues[16], all of which would have impacted the ascertainment of the sample and the phenotypic characterization. Further, based on the developmental stage in which the sample was recruited there may be differences in the phenotypic characterization as development is a dynamic process and phenotypic manifestations may emerge later in life or change over the life course. In this regard, it is to be noted that the reference GWAS studies of ASD, allergies, and chronic pain disorder were conducted in the adult populations, while the test sample of ASD was a pediatric cohort. Specifically, some comorbid conditions may only manifest after certain ages and hence reference effect sizes based on the adult sample may lead to biased prediction using PRS values in the test sample consisting of children. The null association between PRS_{pain} and Topic 17 may indicate that age also played a substantial role in Topic 17; a cluster of comorbid conditions over-represented by sensory processing issues such as low pain tolerance, despite adjusting for age in the model. Second, each of the topics of comorbidities refers to a cluster of various correlated but distinct phenotypes, where the genetic architectures may not be well captured by PRS derived from a GWAS study of one single phenotype. Third, predictive values of PRS may substantially decrease if training (i.e. reference cohort) and testing data (i.e. scored cohort) sets are drawn from different populations.[21] Fourth, the AGRE sample may be underpowered to detect weak PRS associations due to its small samplesize.

Nonetheless, our research has several noteworthy strengths. First and foremost, we successfully implemented the methodology employed by McCoy and colleagues for ASD, which is an arguably more complex psychiatric category for diagnosis and detection than MDD[22,23]; the domain of mental illness in which this approach was first tested. Second, our implementation used a semi-supervised approach to the selection for the optimal number of topics based upon recent advances in the applied topic modeling literature. Thirdly, the use of a phenotypically well characterized sample to identify ASD subgroups using topic modeling is a substantial strength of the study. Finally, the way in which pseudo-EMRs were generated on the basis of clinical/diagnostic data is novel and may well be useful in other analytic contexts.

In summary, the current findings suggest that the occurrence of two subgroups of ASD-related comorbidities—allergies and chronic pain—may be driven by shared underlying genetic risk for ASD. Notably, these two types of comorbid conditions could not be attributable to genetic variants associated with either allergies or chronic pain disorder in non-ASD populations. Such findings suggest that genetic mechanisms of certain comorbid conditions such as allergies and sensory processing issues in patients with ASD could differ from those of the non-ASD population. Further, it is possible that there is a subgroup where the specific co-morbidities may indicate an alternate converging process such as a common immune pathway.[24] In this regard, immune system deficiencies and immune dysregulation in ASD may result in a wide variety of co-morbidities such as allergic sensitivities, asthma, rashes, gastro intestinal and skin sensitivities as well as sensory issues. Thus, the findings of genetic contributors for comorbid conditions in ASD may inform clinical management strategies. For example, treatment of comorbid allergies in persons diagnosed with ASD may still depend on the clinical management of core symptoms of ASD rather than allergy-specific therapies despite previously reported genetic correlations between ASD and allergies.[12] Alternatively the treatment may need to be targeted at any underlying immune related issues. Further, although emerging evidence suggests that peripheral somatosensory neurons are involved in tactile-related phenotypes in ASD[25,26], genetic variants associated with pain disorder seems at best to only play a limited role in abnormalities in pain sensation, in children with ASD. This deserves further exploration through research involving larger population samples as better understanding of the underlying pathogenetic

mechanisms involved in comorbid conditions in ASD may have clinical implications in the comprehensive assessment and management of these patients.

Methods

All methods were carried out in accordance with relevant guidelines and regulations.

Sample Description

The discovery sample of ASD comprised the whole-genome genotypic data retrieved from the Autism Genetic Resource Exchange (AGRE), of which the subject recruitment has been described elsewhere.[27] Briefly, AGRE is a joint effort of the Cure Autism Now (CAN) Foundation and the Human Biological Data Interchange (HBDI). The diagnosis was made by all of the NIH autism collaborative groups using the Autism Diagnostic Interview–Revised (ADI-R)[28] and the Autism Diagnostic Observational Schedule (ADOS)[29]. We have downloaded the clinical and SNP data (generated by the Affymetrix SNP 5.0 platform). We implemented the same data-cleaning algorithm used in the discovery sample. A total of 325,971 valid SNPs for 1,387 subjects, 97.3% had an European ancestry[30], diagnosed with ASD were obtained. The final dataset consisting of 707 subjects with comorbid physical and psychiatric features as well as perinatal factors were used to examine the patterns of comorbid conditions in the present study.

Data Preparation

Data preparation was performed using the R software v3.6.1[31] within the RStudio integrated development environment[32]. Full reproducible code and session information is available upon request.

Pseudo-EMRs

Documents were constructed as electronic medical record (EMR) analogues on the basis of a multistep process. Each psychiatric feature (for e.g., symptoms, diagnoses, historical presentations) was considered a candidate term for inclusion in the 'pseudo-EMR' for each subject. Categorical data were collapsed into the presence versus absence of features. Continuous data were then dichotomized using k-means clustering (i.e. $k = 2$).[33,34] Some continuous features required special consideration in relation to dichotomization, for e.g., features relating to developmental delays for which delay thresholds were drawn from clinical literature. Features were excluded where counts per feature were less than or equal to 1, as these were considered prohibitively sparse. Following dichotomization, psychiatric features were transcoded into labels indicating the presence of features by subject. Hyphenation was used to coerce features with complex clinical descriptions into single terms (e.g., 'floppy infant'; 'infant-floppy'). This is done to ensure that subsequent topic modeling did not tokenize features in ways disrupting correspondence with the observed data. Examples of the pseudo-EMRs produced by this process are shown in Additional File 1.

Topic Modelling

The optimal number of topics to model was explored prior to topic modeling.[35] Briefly, topic modeling, such as Latent Dirichlet Allocation (LDA), is used to identify abstract 'topics' that occur in a collection of documents.[36] Accordingly, a parallel process was run over a range of candidate models in order to evaluate the relationship between number of topics and model diagnostics according to recommendation by Chang et al., 2009 (Additional File 2).[37] The outcome of this procedure was the selection of 18 as the optimal number of topics to model given the observed data. LDA was then applied using Gibbs sampling with the following settings: 2000 iterations were performed with 500 iterations for thinning and a burn-in of 1000 iterations. Symmetric Dirichlet priors were applied to ensure topics would be well-separated; $\alpha = .1$, $\beta = .01$.[38] A threshold of 16 terms per topic was chosen following inspection of the resulting topic models (for term-to-topic probabilities, see Figure 1; for term frequencies, see Additional File 3). Topic modeling was conducted using the 'topicmodels' package v0.12.[39] For further background on topic modeling, please refer to Supplementary Information.

Genetic Association Analyses

Genotype and Imputation

All subjects were genotyped using the Affymetrix GeneChip Human Mapping 500K Array. We retained subjects with genotyping call rates exceeding 90% and single nucleotide polymorphisms (SNPs) with a call rate of 90% or greater, and Hardy-Weinberg equilibrium p-value $> 1\text{x}10^{-6}$. We remapped the raw genotype data from the GRCh35 to GRCh37 and conducted quality control by removing SNPs (i) with ambiguous alleles, (ii) with $> .2$ allele frequency difference from the reference panel, and (iii) not available in the reference panel using the Pre-imputations checks toolbox (https://www.well.ox.ac.uk/~wrayner/tools/) and the 1000 Genome European reference panel. Genotypes were next imputed using the Michigan Imputation Server implementing Minimac4[40], based on the European subset from the 1000 Genomes Phase 3 v5

(GRCh37/hg19) as reference panel with an imputation filter of $R^2 > .3$. Phasing of haplotypes was conducted using 'Eagle' v2.4.[41]

Polygenic Risk (PRS) Calculation

We generated polygenic risk scores (PRS) for ASD (n = 18,381 cases and 27,969 controls)[42], allergic disease (n = 180,128 cases and 180,709 controls)[6] and chronic pain (n = ~380,000)[15], using seven tranches of SNPs ($1x10^{-2}$, $1x10^{-3}$, $1x10^{-4}$, $1x10^{-5}$, $1x10^{-6}$, $1x10^{-7}$, $5x10^{-8}$, labelled as S2-S8) drawn from recent genome-wide association studies. The value for each p-value tranche represents the maximum p-value that is included in that tranche. This list was linkage-disequilibrium pruned using the 'clump()' function as implemented in 'PLINK' v1.9, with a 250kb window and minimum R^2 set at 0.5 by default.[43]

Statistical Analyses

The relationship between PRS tranches and topics across ASD, allergic disease and chronic pain were tested using a linear regression model. All model topics were inverse normal transformed prior analysis because they were not normally distributed. Covariates included age, sex, and the first 5 genetic principal components. We conducted analysis of variance (ANOVA) tests to compare the fits of models. Linear regression and ANOVA tests were conducted using the R language v3.6.1[31] within the RStudio IDE[32]. Further, we estimated the proportion of the variance of the dependent variable (i.e., topics of comorbid conditions) that could be explained by each predictor (i.e., PRS specific to ASD, PRS specific to the corresponding topic, age, and gender) using partial and semipartial correlation coefficients of a specified predictor – which were used to compare relative contributions of PRS specific to different phenotypes to the topics. Finally, we examined genetic correlations between ASD and comorbid conditions associated with PRS specific to ASD based on the lists of candidate genes extracted using the web tool, Genepanel.iobio.[44] The correlation was inferred based on the probability of detecting significant phenotype-phenotype associations by random chance calculated using the hypergeometric distribution.[45]

References

1. Hodges, H., Fealko, C. & Soares, N. Autism spectrum disorder: definition, epidemiology, causes, and clinical evaluation. *Transl Pediatr* **9**, S55–S65, DOI: 10.21037/tp.2019.09.09 (2020).
2. Al-Beltagi, M. Autism medical comorbidities. *WJCP* **10**, 15–28, DOI: 10.5409/wjcp.v10.i3.15 (2021).
3. Isaksen, J. *et al.* Children with autism spectrum disorders – The importance of medical investigations. *Eur. J. Paediatr. Neurol.* **17**, 68–76, DOI: 10.1016/j.ejpn.2012.08.004 (2013).
4. Ramaswami, G. & Geschwind, D. H. Genetics of autism spectrum disorder. In Geschwind, D. H., Paulson, H. L. & Klein, C. (eds.) *Handbook of Clinical Neurology*, vol. 147 of *Neurogenetics, Part I*, chap. 21, 321–329, DOI: 10.1016/B978-0-444-63233-3.00021-X (Elsevier B.V., 2018).
5. Yoo, H. Genetics of Autism Spectrum Disorder: Current Status and Possible Clinical Applications. *Exp Neurobiol* **24**, 257–272, DOI: 10.5607/en.2015.24.4.257 (2015).
6. Autism Spectrum Disorder Working Group of the Psychiatric Genomics Consortium *et al.* Identification of common genetic risk variants for autism spectrum disorder. *Nat Genet.* **51**, 431–444, DOI: 10.1038/s41588-019-0344-8 (2019).
7. Rylaarsdam, L. & Guemez-Gamboa, A. Genetic Causes and Modifiers of Autism Spectrum Disorder. *Front. Cell. Neurosci.* **13**, 385, DOI: 10.3389/fncel.2019.00385 (2019).
8. Solberg, B. S. *et al.* Patterns of Psychiatric Comorbidity and Genetic Correlations Provide New Insights Into Differences Between Attention-Deficit/Hyperactivity Disorder and Autism Spectrum Disorder. *Biol. Psychiatry* **86**, 587–598, DOI: 10.1016/j.biopsych.2019.04.021 (2019).
9. Masi, A., DeMayo, M. M., Glozier, N. & Guastella, A. J. An Overview of Autism Spectrum Disorder, Heterogeneity and Treatment Options. *Neurosci. Bull.* **33**, 183–193, DOI: 10.1007/s12264-017-0100-y (2017).
10. David, M. M. *et al.* Comorbid Analysis of Genes Associated with Autism Spectrum Disorders Reveals Differential Evolutionary Constraints. *PLoS ONE* **11**, e0157937, DOI: 10.1371/journal.pone.0157937 (2016).
11. Diaz-Beltran, L. *et al.* Cross-disorder comparative analysis of comorbid conditions reveals novel autism candidate genes. *BMC Genomics* **18**, 315, DOI: 10.1186/s12864-017-3667-9 (2017).
12. Tylee, D. S. *et al.* Genetic correlations among psychiatric and immune-related phenotypes based on genome-wide association data. *Am. J. Med. Genet.* **177**, 641–657, DOI: 10.1002/ajmg.b.32652 (2018).
13. Clarke, C. Autism Spectrum Disorder and Amplified Pain. *Case Reports Psychiatry* **2015**, 1–4, DOI: 10.1155/2015/930874 (2015).

14. Gu, X. *et al.* Heightened brain response to pain anticipation in high-functioning adults with autism spectrum disorder. *Eur J Neurosci* **47**, 592–601, DOI: 10.1111/ejn.13598 (2018).

15. Johnston, K. J. A. *et al.* Genome-wide association study of multisite chronic pain in UK Biobank. *PLoS Genet.* **15**, e1008164, DOI: 10.1371/journal.pgen.1008164 (2019).

16. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders* (American Psychiatric Association, 2013), fifth edition edn.

17. González-Peñas, J. *et al.* Psychiatric comorbidities in Asperger syndrome are related with polygenic overlap and differ from other Autism subtypes. *Transl Psychiatry* **10**, 258, DOI: 10.1038/s41398-020-00939-7 (2020).

18. McCoy, T. H. *et al.* Polygenic loading for major depression is associated with specific medical comorbidity. *Transl Psychiatry* **7**, e1238–e1238, DOI: 10.1038/tp.2017.201 (2017).

19. Australian Asthma Genetics Consortium (AAGC) *et al.* Meta-analysis of genome-wide association studies identifies ten loci influencing allergic sensitization. *Nat Genet.* **45**, 902–906, DOI: 10.1038/ng.2694 (2013).

20. Brown, C. O., Uy, J. & Singh, K. K. A mini-review: Bridging the gap between autism spectrum disorder and pain comorbidities. *Can. J. Pain* **4**, 37–44, DOI: 10.1080/24740527.2020.1775486 (2020).

21. Gola, D. *et al.* Population Bias in Polygenic Risk Prediction Models for Coronary Artery Disease. *Circ: Genomic Precis. Medicine* **13**, DOI: 10.1161/CIRCGEN.120.002932 (2020).

22. Happé, F., Ronald, A. & Plomin, R. Time to give up on a single explanation for autism. *Nat Neurosci* **9**, 1218–1220, DOI: 10.1038/nn1770 (2006).

23. Waterhouse, L. *Rethinking Autism* (Elsevier, Amsterdam, 2013).

24. Mead, J. & Ashwood, P. Evidence supporting an altered immune response in ASD. *Immunol. Lett.* **163**, 49–55, DOI: 10.1016/j.imlet.2014.11.006 (2015).

25. Orefice, L. L. *et al.* Peripheral Mechanosensory Neuron Dysfunction Underlies Tactile and Behavioral Deficits in Mouse Models of ASDs. *Cell* **166**, 299–313, DOI: 10.1016/j.cell.2016.05.033 (2016).

26. Orefice, L. L. *et al.* Targeting Peripheral Somatosensory Neurons to Improve Tactile-Related Phenotypes in ASD Models. *Cell* **178**, 867–886.e24, DOI: 10.1016/j.cell.2019.07.024 (2019).

27. Geschwind, D. H. *et al.* The Autism Genetic Resource Exchange: A Resource for the Study of Autism and Related Neuropsychiatric Conditions. *The Am. J. Hum. Genet.* **69**, 463–466, DOI: 10.1086/321292 (2001).

28. Lord, C., Rutter, M. & Le Couteur, A. Autism Diagnostic Interview-Revised: A revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *J Autism Dev Disord* **24**, 659–685, DOI: 10.1007/BF02172145 (1994).

29. Lord, C. *et al.* The autism diagnostic observation schedule-generic: a standard measure of social and communication deficits associated with the spectrum of autism. *J Autism Dev Disord* **30**, 205–223 (2000).

30. Exchange, A. G. R. AGRE: Autism Genetic Resource Exchange (2008).

31. RCore. R: A language and environment for statistical computing (2018).

32. RStudio. RStudio: Integrated Development for R (2020).

33. Hartigan, J. A. & Wong, M. A. Algorithm AS 136: A K-Means Clustering Algorithm. *Appl. Stat.* **28**, 100, DOI: 10.2307/2346830 (1979).

34. MacQueen, J. Some methods for classification and analysis of multivariate observations. *Proc. fifth Berkeley symposium on mathematical statistics probability* **1**, 281–297 (1967).

35. De Battisti, F., Ferrara, A. & Salini, S. A decade of research in statistics: a topic model approach. *Scientometrics* **103**, 413–433, DOI: 10.1007/s11192-015-1554-1 (2015).

36. Blei, D., Ng, A. & Jordan, M. Latent Dirichlet allocation. *The J. Mach. Learn. Res.* **3**, 601–608 (2001).

37. Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C. & Blei, D. Reading tea leaves: How humans interpret topic models. *Neural information processing systems* **32**, 288–296, DOI: https://doi.org/10.5555/2984093.2984126 (2009).

38. Tang, J., Meng, Z., Nguyen, X., Mei, Q. & Zhang, M. Understanding the limiting factors of topic modeling via posterior contraction analysis. In *31st International Conference on Machine Learning (ICML 2014)*, 190–198 (Stroudsburg, PA, 2014).

39. Grün, B. & Hornik, K. topicmodels: An R Package for Fitting Topic Models. *J. Stat. Soft.* **40**, DOI: 10.18637/jss.v040.i13 (2011).

40. Fuchsberger, C., Abecasis, G. R. & Hinds, D. A. minimac2: faster genotype imputation. *Bioinformatics* **31**, 782–784, DOI: 10.1093/bioinformatics/btu704 (2015).

41. Loh, P.-R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. *Nat Genet.* **48**, 1443–1448, DOI: 10.1038/ng.3679 (2016).

42. 23andMe Research Team *et al.* Shared genetic origin of asthma, hay fever and eczema elucidates allergic disease biology. *Nat Genet.* **49**, 1752–1757, DOI: 10.1038/ng.3985 (2017).

43. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaSci* **4**, 7, DOI: 10.1186/s13742-015-0047-8 (2015).

44. Ekawade, A. *et al.* Genepanel.iobio - an easy to use web tool for generating disease- and phenotype-associated gene lists. *BMC Med Genomics* **12**, 190, DOI: 10.1186/s12920-019-0641-1 (2019).

45. Garcia-Albornoz, M. & Nielsen, J. Finding directionality and gene-disease predictions in disease associations. *BMC Syst Biol* **9**, 35, DOI: 10.1186/s12918-015-0184-9 (2015).

Acknowledgements

We appreciate all patients and families that provided their clinical data and biospecimen for this collaborative multi-site database. This project was not supported by any grant.

Author contributions statement

LK, SD, L-DH, and P-IL designed and drafted the manuscript. LK, SD, L-DH, and P-IL conducted analyses. P-IL conceived the study and provided access to the data. VE provided expert advice and review of clinical implications. LK and SD have equal contributions. L-DH and P-IL are joint senior authors on this work.

Additional information

The authors declare that they have no competing interests. De-identified genetic and clinical data used in the current study can be accessed at `https://figshare.com/articles/dataset/Autism_GWAS_data/14253230`. Full reproducible code and session information is available upon request.

Figures

Figure 1. 'Terms-to-topic' probabilities for patients with autism spectrum disorder (ASD)
Names for each panel indicate the topic number from which associated data are drawn. *x*-axes indicate beta weights (β) or 'term-by-topic' probabilities. *y*-axes indicate the most probable (n = 10) terms rank-ordered by β weight. Dashed lines indicate β = .05.

Figure 2. Heatmap of the associations between all topics and polygenic risk score (PRS) tranches of autism spectrum disorder (ASD)
Topic 6 is over-represented by terms relating to allergies while Topic 17 is over-represented by sensory processing issues.

Figure 3. Relative contributions of polygenic risk scores to the two topics of comorbid conditions
Relative contributions were calculated using the squared values from semi-parametric correlation tests. Topic 6 is over-represented by allergies while Topic 17 is over-represented by sensory processing issues. PRS_{ASD} indicates the PRS values derived from the GWAS study of ASD, $PRS_{allergy}$ indicates the PRS values derived from the GWAS study of allergies, and PRS_{pain} indicates the PRS values derived from the GWAS study of chronic pain disorder.

Figure 4. Genetic correlations between ASD and the other two comorbid phenotypes (allergy, pain)
The size of the circle is proportional to the number of candidate genes.

Additional Files

Additional File 1 — *Pseudo-EMRs constructed in preparation for topic modelling by latent Dirichlet allocation*

Numbers 1-6 indicate unique subjects with ID masked. Observed psychiatric and sociodemographic features have been transformed into a pseudo-EMR for each subject.

Additional File 2 — *Evaluation of various model diagnostics in relation to selecting the optimal number of topics to model given the observed data*

Top left: Held-out likelihood indicates perplexity (higher is better). Top right: Lower bound indicates model convergence (higher is better). Bottom left: Residuals indicate model saturation (lower is better). Bottom right: Semantic coherence indicates co-occurrence of probable terms per topic (higher is better).

Additional File 3 — *Descriptive table of counts and percentages for clinical features included in topic modeling procedure*

Each count represents the presence of the token psychiatric feature for a single subject.

Additional File 4 — *Association between all topics and polygenic risk score (PRS) tranches of autism spectrum disorder (ASD)*

Seven tranches of SNPs ($1x10^{-2}$, $1x10^{-3}$, $1x10^{-4}$, $1x10^{-5}$, $1x10^{-6}$, $1x10^{-7}$, $5x10^{-8}$) are labelled as S2-S8.

Figures

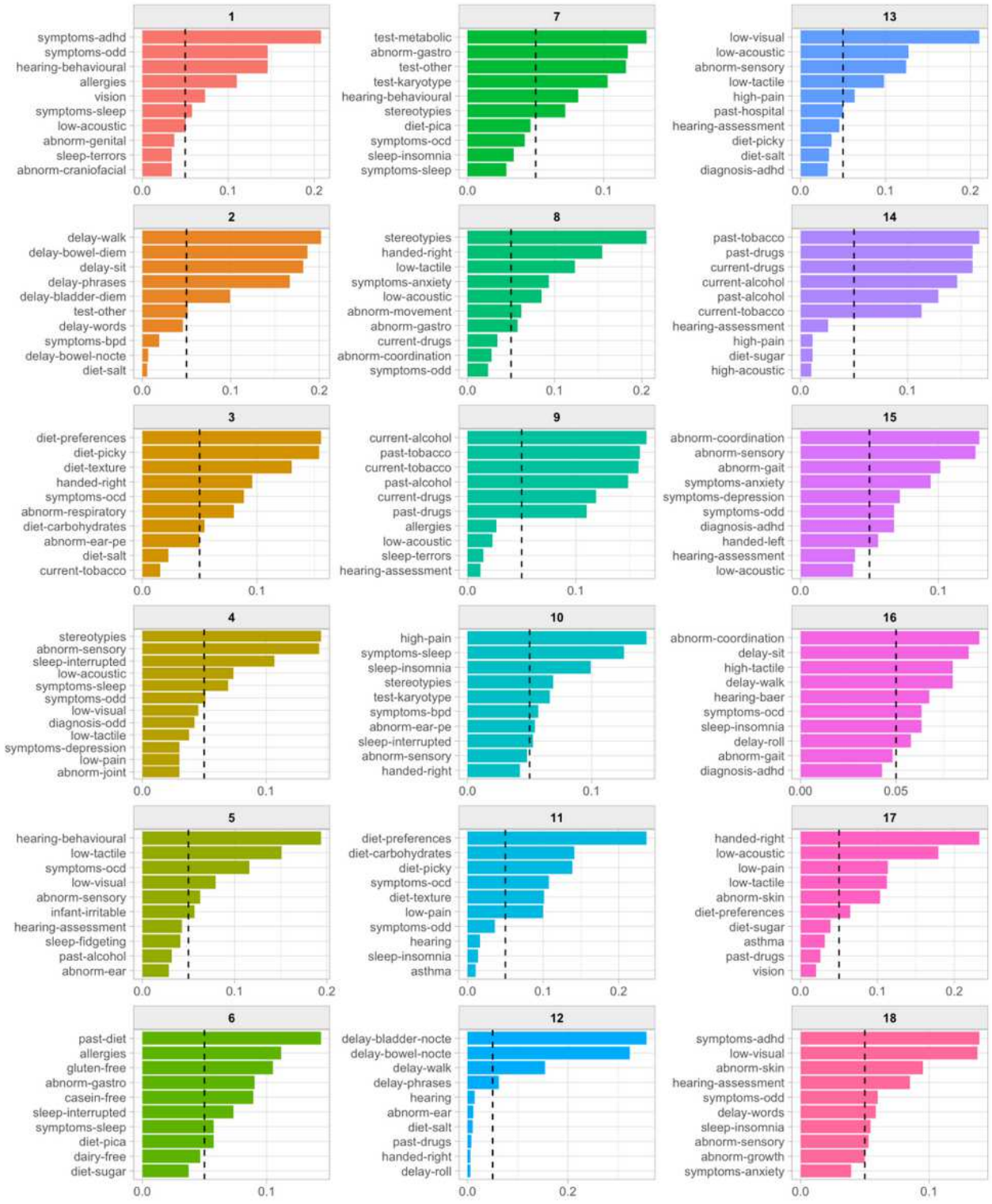

Figure 1

'Terms-to-topic' probabilities for patients with autism spectrum disorder (ASD) Names for each panel indicate the topic number from which associated data are drawn. x-axes indicate beta weights (β) or'term-

by-topic' probabilities. y-axes indicate the most probable (n = 10) terms rank-ordered by β weight. Dashed lines indicate β = .05.

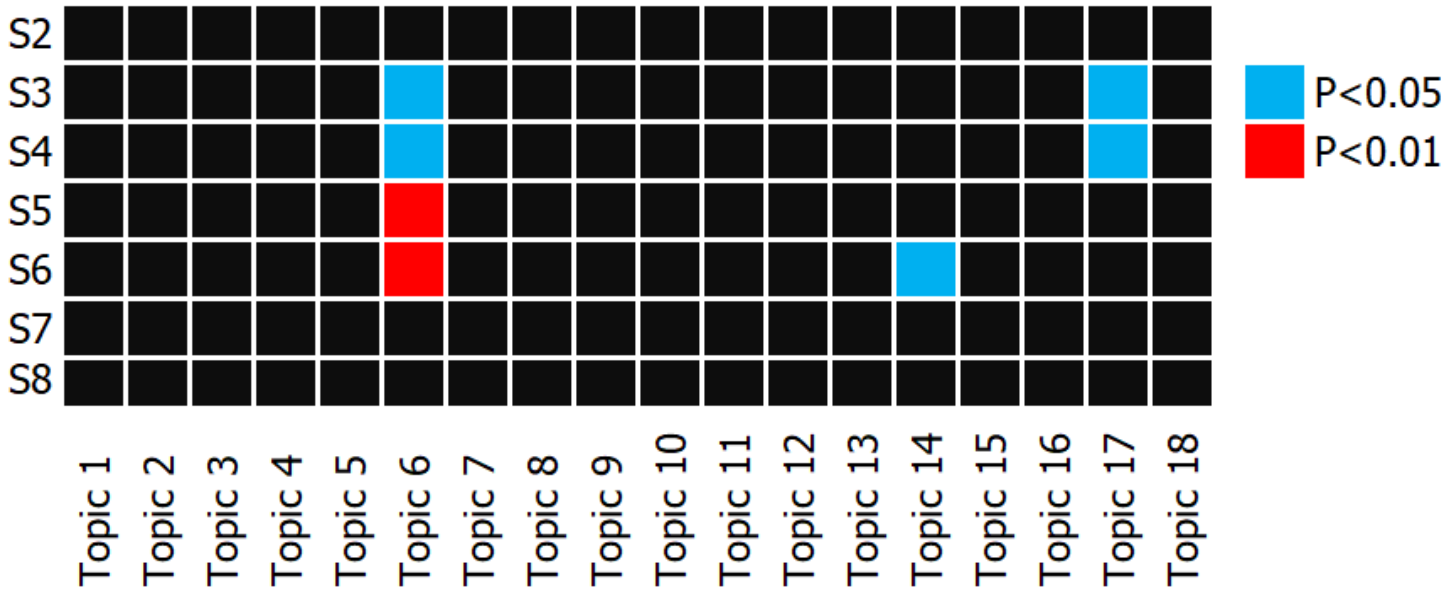

Figure 2

Heatmap of the associations between all topics and polygenic risk score (PRS) tranches of autism spectrum disorder (ASD)Topic 6 is over-represented by terms relating to allergies while Topic 17 is over-represented by sensory processing issues

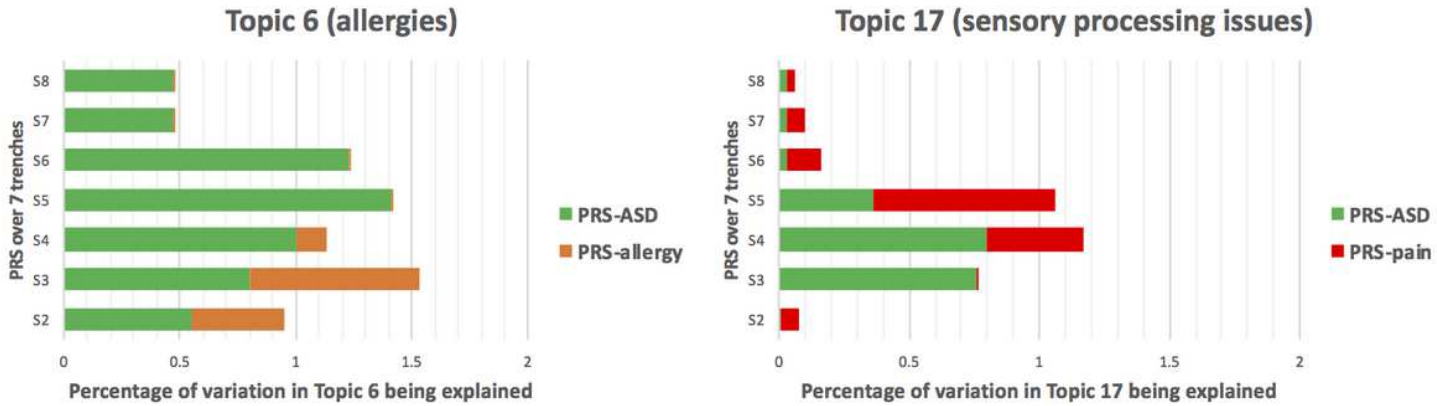

Figure 3

Relative contributions of polygenic risk scores to the two topics of comorbid conditions Relative contributions were calculated using the squared values from semi-parametric correlation tests. Topic 6 is over-represented by allergies while Topic 17 is over-represented by sensory processing issues. PRSASD indicates the PRS values derived from the GWAS study of ASD, PRSallergy indicates the PRS values

derived from the GWAS study of allergies, and PRSpain indicates the PRS values derived from the GWAS study of chronic pain disorder

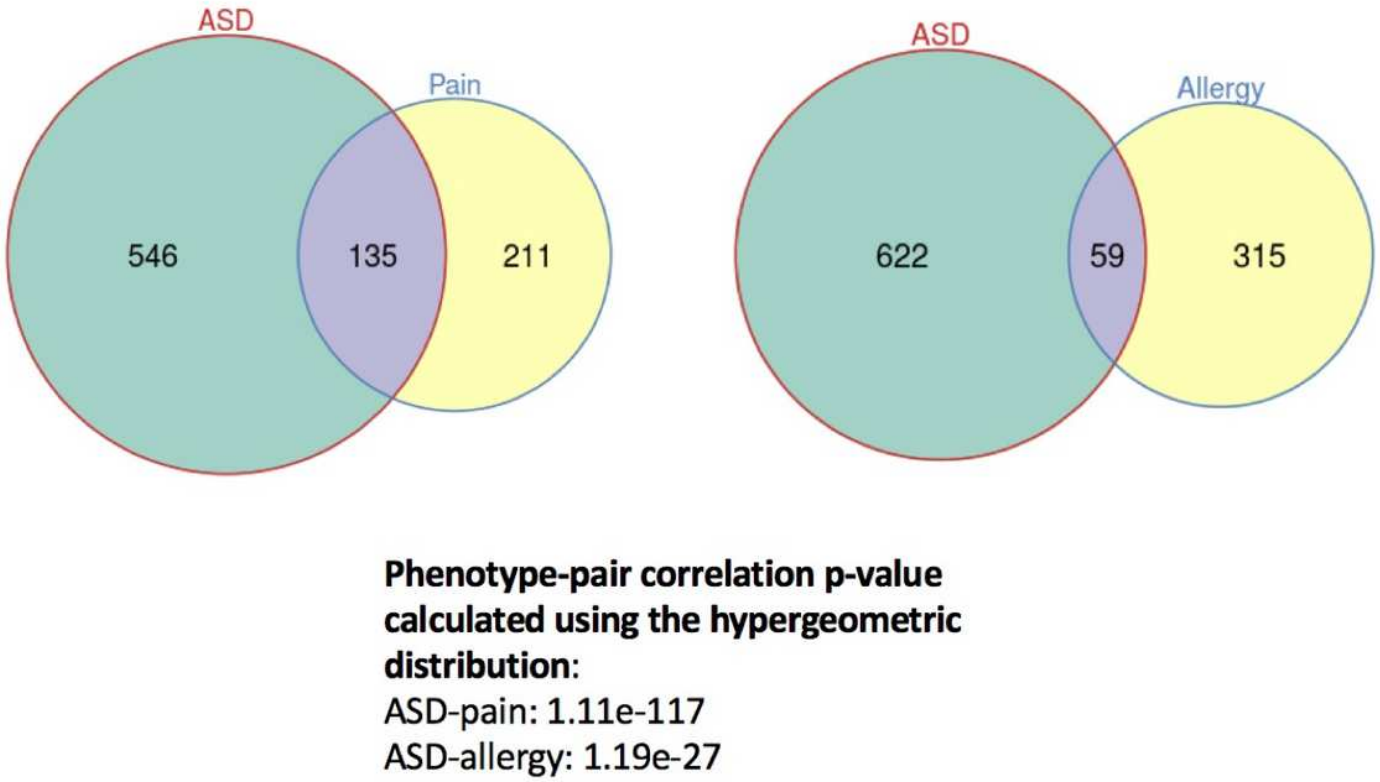

Figure 4

Genetic correlations between ASD and the other two comorbid phenotypes (allergy, pain) The size of the circle is proportional to the number of candidate genes.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AdditionalFile1.jpeg
- AdditionalFile2.jpeg
- AdditionalFile3.jpeg
- AdditionalFile4.xls
- SupplementaryInformation.docx