A 3D approach to understanding heterogeneity in early developing autisms

Veronica Mandelli[1†], Ines Severino[1,2†], Lisa Eyler[3,4], Karen Pierce[5], Eric Courchesne[5] and Michael V. Lombardo[1*]

Abstract

Background Phenotypic heterogeneity in early language, intellectual, motor, and adaptive functioning (LIMA) features are amongst the most striking features that distinguish different types of autistic individuals. Yet the current diagnostic criteria uses a single label of autism and implicitly emphasizes what individuals have in common as core social-communicative and restricted repetitive behavior difficulties. Subtype labels based on the non-core LIMA features may help to more meaningfully distinguish types of autisms with differing developmental paths and differential underlying biology.

Methods Unsupervised data-driven subtypes were identified using stability-based relative clustering validation on publicly available Mullen Scales of Early Learning (MSEL) and Vineland Adaptive Behavior Scales (VABS) data (n=615; age=24–68 months) from the National Institute of Mental Health Data Archive (NDA). Differential developmental trajectories between subtypes were tested on longitudinal data from NDA and from an independent in-house dataset from UCSD. A subset of the UCSD dataset was also tested for subtype differences in functional and structural neuroimaging phenotypes and relationships with blood gene expression. The current subtyping model was also compared to early language outcome subtypes derived from past work.

Results Two autism subtypes can be identified based on early phenotypic LIMA features. These data-driven subtypes are robust in the population and can be identified in independent data with 98% accuracy. The subtypes can be described as Type I versus Type II autisms differentiated by relatively high versus low scores on LIMA features. These two types of autisms are also distinguished by different developmental trajectories over the first decade of life. Finally, these two types of autisms reveal striking differences in functional and structural neuroimaging phenotypes and their relationships with gene expression and may highlight unique biological mechanisms.

Limitations Sample sizes for the neuroimaging and gene expression dataset are relatively small and require further independent replication. The current work is also limited to subtyping based on MSEL and VABS phenotypic measures.

Conclusions This work emphasizes the potential importance of stratifying autism by a Type I versus Type II distinction focused on LIMA features and which may be of high prognostic and biological significance.

Keywords Stratification, Subtypes, Clustering, fMRI, Gene expression

†Veronica Mandelli and Ines Severino have equally contributions to this work.

*Correspondence:
Michael V. Lombardo
michael.lombardo@iit.it
Full list of author information is available at the end of the article

Background

The diagnostic concept of autism has been anything but stationary over time [1]. To illustrate this non-stationarity, consider that the autism of several decades ago was a population largely comprised of individuals with the core social-communication (SC) and restricted repetitive behavior (RRB) issues, but also frequently possessed profound issues in non-core features such as language, intellectual, motor, and adaptive functioning (features that are henceforth referred to as acronym 'LIMA'). In contrast, the autism of today is largely comprised of individuals with varying and sometimes subtle core SC and RRB issues, but whom are much more likely to not have such profound LIMA difficulties [2]. As a likely result of these changes over time, we have seen phenomena such as increasing heterogeneity [3], prevalence [4], decreasing effect sizes [5], and a change in views about autism, particularly from the neurodiversity perspective. Thus, it is no surprise that the most contentious debates in the field today deal with topics that ultimately boil back down to contrasting the relative importance of core SC and RRB versus non-core LIMA features. This can be seen in debates contrasting medical versus social perspectives [6], vocabulary and language use [7], and debate regarding different subtyping approaches (e.g., prototypical autism, profound autism) [8–10] versus 'spectrum'/continuous models (e.g., autistic traits, polygenic risk) [11, 12].

Our contention is that the autism spectrum has become too wide to be meaningful as a whole when it comes to finding explanations about differential **b**iology, **o**utcomes, and responses to **t**reatment (i.e. an acronym we label as the 'BOT' objectives [1]). Thus, at least for the purposes of the BOT objectives, we suggest that autism must be split into different types of autisms distinguished by non-core LIMA features. Owing to terminology already in use in the literature regarding dual perspectives of autism as 'disability' (i.e. the medical model) versus a 'difference' (i.e. social models such as neurodiversity), we propose that an initial split of autism be one of two highly differentiated early developmental and behavioral types—that is, one of profound 'disability' (Type I) versus one of a 'difference' (Type II). Furthermore, these contrasting types are hypothesized to be most prominently differentiated over 'development'. Putting together these 3 'D' words— 'disability' versus 'difference' over 'development'—we propose a framework called AUTISMS-3D (henceforth referred to as A3D). In simple terms, the A3D model suggests that when one focuses on the non-core LIMA features, autism can be thought of as at least two different types that contrast disability versus difference over development.

A key distinction separating A3D from other approaches [8, 9, 13] is the idea that these theorized autism subtypes are evident and detectable with unsupervised data-driven approaches in early development. In past work utilizing just the domains of the Vineland Adaptive Behavior Scales (VABS), we showed that stability-based relative clustering validation could be applied to separate the predicted Type I versus Type II autisms with very high generalizability accuracy in new unseen datasets [14]. Thus, unsupervised data-driven discoveries can be immediately translated into supervised prediction models that can generalize with very high accuracy and highlight the A3D subtype distinction. This attribute of the A3D approach separates it from past issues that have been present with subtyping in autism [15]. One of the primary reasons for abandoning DSM-IV subtype labels of PDD-NOS, Asperger Syndrome, high- and low-functioning autism was because of the lack of agreement between expert clinicians in making such labels [16]. Even with more recent suggestions such as the label of 'profound autism' [9], it could be argued that the defining boundaries for that label are still based on potentially arbitrary and/or hard cutoffs for age (e.g.,>8 years) and IQ (e.g., IQ<50). Furthermore, the Lancet commission suggested the use of a label like profound autism be primarily administrative [9], leaving a gap with regard to how research might be able to utilize a label that picks up on a similar kind of conceptual distinction. In contrast to the approach of focusing on 'prototypical autism', A3D again separates itself by not relying on subjective expert clinical judgement for making such a distinction. While there is no substitute for expert clinical judgement as the gold standard for making the autism diagnosis itself, for making these types of subtype distinctions, A3D makes the prediction that subtypes based on LIMA features will be evident in robust statistical differences that unsupervised data-driven models will be able to detect with very high generalizability accuracy. In this work, we go beyond past work utilizing VABS domains [14] to measure early developing skills in all four pillars of non-core LIMA features. We predict that Type I versus Type II autisms will be evident via stability-based relative clustering validation [17]. Furthermore, we predict that the theorized Type I versus Type II autisms will be heavily differentiated in developmental trajectories tested with longitudinal data over the first 5–6 years of life.

Given one of the field's over-arching research priorities is to better understand differential biology behind many autisms (i.e. the 'B' in the 'BOT' objectives) [1], another goal of the A3D model is to test hypotheses regarding whether these two phenotypically differentiated types are also sensitive indicators of differential underlying neurobiology. In related past work, we utilized a stratification

Mandelli *et al. Molecular Autism*   (2024) 15:41

Page 3 of 16

approach based on early language outcomes (ELO) and showed that the distinction between good versus poor ELO is indicative of types with differential underlying early neurobiology [18–21]. The ELO stratification model was used here as a benchmark to test the performance of the A3D model in explaining biological variation. The ELO stratification approach leverages only one of the four pillars of non-core LIMA features we hypothesize to be relevant (i.e. language), while A3D takes into account the full set of LIMA features. Thus, in the current work, we seek to evaluate how autism subtypes predicted from A3D may match up (or outperform) in terms of early functional neural systems response to speech and structural neuroimaging phenotypes (e.g., surface area (SA) and cortical thickness(CT)) that had been previously identified to be highly important in ELO subtypes [18–20].

Methods

NDA dataset

The primary dataset used in this work for the purposes of building a generalizable stratification model comes from the National Institute of Mental Health Data Archive (NDA) (https://nda.nih.gov/). NDA includes several datasets that measure two standardized early clinical measures that are integral for this work—the Mullen Scales of Early Learning (MSEL) and Vineland Adaptive Behavior Scales (VABS). All MSEL and VABS data from NDA was downloaded in February 2022 for further analysis in this work. MSEL and VABS data from NDA was then filtered, preprocessed and analyzed for the existence of stable and highly replicable theoretically predicted autism subtypes (e.g., Type I and Type II) at early ages via unsupervised data-driven approaches (i.e., *reval*) [17]. We downloaded all available MSEL [22] and VABS (i.e., VABS 2nd edition parent and caregiver form or survey form; VABS 3rd edition) [23, 24] within the age range of 0–68 months. VABS standard scores from communication, daily living skills, socialization, and motor domains were extracted for use in further data analysis. For MSEL data, we extracted chronological age at MSEL assessment as well as age-equivalent scores for visual reception (VR), fine motor (FM) and receptive and expressive language (RL; EL) subscales. Next, we implemented a series of data cleaning steps that entailed selecting only individuals with an autism diagnosis, dropping duplicate data, extracting the earliest available time point for individuals with longitudinal data, selecting only individuals whose MSEL and VABS data were assessed not more than one year apart, and finally drop individuals with more than one missing subscale for each of the instruments used (Fig. 1A). After data cleaning, a total of n=615 autistic individuals remained for further downstream analysis. Data from individuals with more than one data point
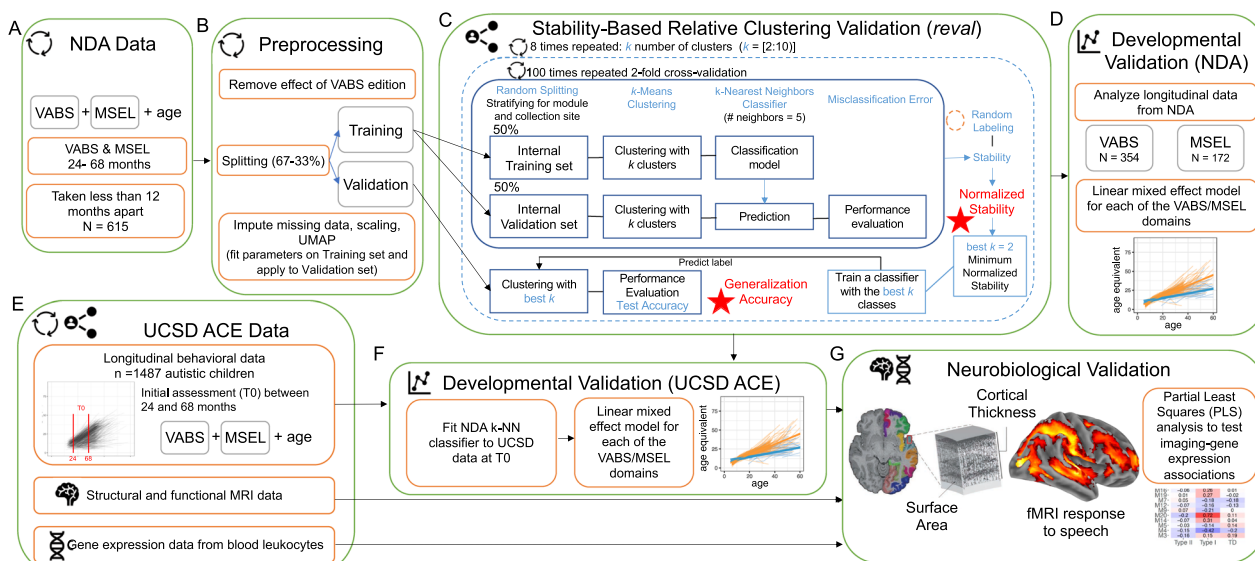

Fig. 1 Schematic of data analysis workflow/pipeline. This figure shows a workflow of the analyses conducted in this work. Panels **A** and **B** depict how we select, clean, and preprocess data from the National Institute of Mental Health Data Archive (NDA). To identify subtypes in a data-driven manner, we used stability-based relative clustering validation analysis (*reval*) (panel **C**) on a final sample of 615 autistic children aged 24–68 months. Subtypes are tested for their prognostic validity (panel **D**) utilizing longitudinal data from NDA. The stratification model was tested again for developmental trajectory differences in an independent longitudinal dataset from UCSD ACE (panels **E**–**F**). Subtypes were also tested for neurobiological differentiation using structural and functional neuroimaging data and relationships to blood leukocyte gene expression patterns (panel **G**)

Mandelli *et al. Molecular Autism*   (2024) 15:41

Page 4 of 16

were also utilized for the subsequent longitudinal analysis that sought to test the hypothesis that the subtypes' developmental trajectories heavily differentiate over the first 5–6 years of life (MSEL n = 179; VABS n = 354).

Before doing the primary downstream *reval* clustering analysis on the NDA dataset, we had to address the batch effect caused by the presence of different VABS versions (e.g., VABS Edition 2 which comprises either the caregiver or survey forms, as well as VABS 3rd edition). For each of the VABS domains, we apply a batch correction factor to adjust the raw data. These factors correspond to the beta coefficients estimated in a linear model with VABS standard scores as the dependent variable and VABS version as the independent variable with age and sex as covariates). Batch correction factors are taken from Mandelli et al., [14] since they were estimated in a bigger sample size within the age range of interest for this work (n = 604, age 0–72 months). To project out batch-related variability brought on by the VABS version, the correction factors (i.e., beta coefficients) were multiplied against the raw data and then the difference was removed from the raw VABS domain standard scores. Subsequent *reval* clustering analyses were carried out with version-corrected VABS data.

UCSD dataset

To test the stratification model for differences in developmental trajectories using a dataset that is independent of the NDA dataset used to build the stratification model, we utilized a longitudinal phenotypic dataset from the University of California, San Diego (henceforth referred to as UCSD ACE). This study was approved by the Institutional Review Board at University of California, San Diego. This dataset comprises longitudinal VABS and MSEL measurements from n = 1487 autistic children. The enrollment of most individuals into the UCSD ACE dataset started from ~12 months using a combination of population-based autism risk screening and community referrals [25]. Longitudinal data was collected every six months from the initial intake assessment until the age of 5–6 years of age (Supplementary Table 1), and included a comprehensive battery of assessments such as VABS, MSEL, and the Autism Diagnostic Observation Schedule (ADOS-2) [26]. Clinical diagnosis of autism was made at by expert clinicians at the UCSD ACE center. To match the age range of the NDA subtyping model, we only included a subset of children with an assessment between the ages of 24 and 68 months in the current study. The NDA subtype prediction model was used to predict subtype labels for the UCSD ACE dataset based on VABS—MSEL scores at the first assessment time point between 24–68 months.

A subset of autistic and non-autistic individuals from the UCSD dataset (n = 81 autism; n = 37 typically developing (TD); n = 19 language or developmental delays (LD/DD)) with structural (MRI) and functional MRI (fMRI) and blood leukocyte gene expression data was also analyzed. This dataset was previously reported in past papers examining heterogeneity in early language outcomes (ELO) [18–20]. Here we applied the A3D prediction model to label autistic individuals in this imaging dataset as either Type I and Type II and then test if such labels are sensitive to differential underlying neurobiology. For details on data collection and processing of UCSD neuroimaging and gene expression data, see the section on neuroimaging and gene expression data.

Phenotypic measures utilized in NDA and UCSD ACE datasets

Vineland Adaptive Behavior Scales (VABS)

The VABS is a widely used standardized and semi-structured caregiver interview for assessing adaptive functioning in typical and clinical populations throughout the lifespan [23, 24]. For children within the age range included in this work (≤68 months) the VABS assesses adaptive functioning skills within four domains: communication, daily living skills, socialization, and motor skills. In this work, we used standardized scores for each domain. Standardized scores are computed to indicate where an individual stands relative to typically-developing age-appropriate norms, whereby for each standardized score, the mean is 100 and the standard deviation is 15.

Mullen Scales of Early Learning (MSEL)

The MSEL is a standardized assessment for non-verbal cognitive, motor, and language skills that can be administered in children aged 0–68 months [22]. In this work, four MSEL subscales were utilized which are relevant for assessing non-core autism characteristics within the A3D framework: visual reception (VR), fine motor (FM), receptive language (RL), and expressive language (EL). For each subscale, developmental quotients (DQ) were computed as (age-equivalent/chronological age)*100, to have age-standardized scores centered around a value of 100. Computation of MSEL DQ scores eliminates some known issues with floored MSEL t-scores but also allows for MSEL to be used in further data analysis for clustering with compatible age-standardized scores consistent with VABS (e.g., age-standardized scores centered at a value of 100 to indicate normative age-appropriate skill level). For evaluating MSEL developmental trajectories, we utilized MSEL age-equivalent scores.

Stability-based relative clustering validation analysis (reval)
Our primary stratification model utilizes an unsupervised data-driven clustering approach that seeks to identify the optimal number of clusters with optimal level of generalizability in new data. This approach is based on stability-based relative clustering validation [17, 27], whereby we transform the unsupervised learning framework into a classification problem which allows for immediate translation of data-driven clustering solutions into supervised classification models that can be empirically demonstrated to be robust and generalizable partitions of the autism population. Stability-based relative clustering validation analysis was implemented with our Python library called *reval* [17]. The *reval* algorithm is shown in Fig. 1C and can be described in the following series of steps. First, the dataset is split into independent training (dataset X_{tr}) and validation (dataset X_{val}) sets. Within the training set (X_{tr}), *reval* splits the data into internal training (X_{tr_tr}) and test (X_{tr_ts}) sets, with the goal to identify the optimal number of clusters (k). Clustering algorithm A_k is then fit to both X_{tr_tr} and X_{tr_ts} and produces clustering labels Y_{tr_tr} and Y_{tr_ts} respectively. A classifier (Φ_{tr_tr}) is then trained on $[X_{tr\_tr}, Y_{tr\_tr}]$ and then fit to X_{tr_ts} to predict labels $Y_{tr\_ts\_pred}$. Misclassification error is computed by comparing the A_k's clustering labels (Y_{tr_ts}) to a classifier's (Φ_{tr_tr}) predicted labels ($Y_{tr\_ts\_pred}$). The *reval* algorithm iterates this procedure over a range of possible k clusters, and then identifies the optimal k that minimizes misclassification error on the internal test set (X_{tr_ts}). Misclassification error is the normalized Hamming distance between the actual clustering labels (Y_{tr_ts}) versus the classifier's (Φ_{tr_tr}) predicted labels ($Y_{tr\_ts\_pred}$). However, because of the non-uniqueness of clustering labels, *reval* permutes the labels (Y_{tr_ts}) and uses the Kuhn-Munkres algorithm to identify the labeling solution that minimizes misclassification error. This final measure of misclassification error is called 'clustering stability' [27] and this index ranges from 0–1, with lower values indicating more stable and reproducible clustering solutions. Stability is then normalized to the stability of random labelings to arrive at the final measure of normalized stability. Thus, the overall first goal in *reval* is to identify the clustering solution A_k that minimizes normalized stability. Once the optimal k is identified from the training set, we then train a classifier (Φ_{tr}) on the entire training set at the optimal k, and then apply it to held-out validation set (X_{val}). Clustering labels (Y_{val}) obtained on the validation set X_{val} are then compared to the classifier's (Φ_{tr}) predicted labels (Y_{val_pred}) and we compute a measure of generalizability accuracy operationalized as 1-misclassification error. Clustering solutions with high generalizability accuracy are the solutions we seek, as they are informative to solutions that have high potential to generalize well on the larger autism population. Furthermore, if generalization accuracy is very high, the output of *reval* is a classifier (i.e. prediction model) which we can use on new independent datasets to predict subtype labels (e.g., Type I vs Type II).

Before running *reval* we first select data and features from the NDA dataset for use in clustering. For selecting NDA data, we utilized all available data from participants that had both MSEL and VABS between 24–68 months, and whereby the MSEL and VABS are sampled within 12 months of each other. For individuals with repeated MSEL and VABS measures (i.e. longitudinal data) the very first timepoint of MSEL and VABS was utilized for clustering. The features selected comprise all MSEL subscales (EL, RL, VR, FM) and VABS domains (Socialization, Communication, Daily Living Skills, Motor). Standardized scores from both measures were utilized for clustering, alongside the age at time of assessment.

Before entering the *reval* clustering analysis, the data is split into independent training (X_{tr}) and validation (X_{val}) sets (training set = 67%; test set = 33%) while balancing the splits for variables such as age, sex, and VABS version. Three preprocessing steps are then applied. First, we impute missing values with a k-nearest neighbors imputation algorithm where the number of neighbors is identified via grid search (*sklearn.impute.KNNImputer*). The next two steps are ones where the parameters are fit on the training set and then applied to the validation set—scaling each feature in the dataset to mean of 0, standard deviation (SD) of 1 (*sklearn.preprocessing.StandardScaler)* and applying the dimensionality reduction technique Uniform Manifold Approximation and Projection (UMAP) [28] (https://umap-learn.readthedocs.io/en/latest; n_neighbors = 30, min_dist = 0.0, n_components = 2, random_state = 42, metric = Euclidean) to reduce the number of features from 8 to 2 (Fig. 1B). The clustering algorithm used for this work was k-means clustering and was chosen on the basis of being one of the simplest and easier to understand clustering algorithms to use and because it is the same algorithm used in further validation analyses we run such as SigClust [29]. The classification model used throughout *reval* for this work was chosen to be k-nearest neighbors. This type of classifier was selected based on prior evidence that it typically performs best in combination with k-means clustering in several validation analyses across multiple types of datasets, as shown in the original *reval* paper [17].

While the *reval* clustering approach is meant to ensure the stability of clustering solutions, the optimal k that is determined by *reval* must still be tested for whether it is a solution indicative of a situation where actual true clusters exist. To test the solution against the null hypothesis, we utilized a Monte Carlo simulation framework called

SigClust [29], meant to test whether the data significantly deviate from the null hypothesis that the data originates from a single multivariate Gaussian null distribution. This test was implemented with the *sigclust* library in R [29] with number of simulations set to 10,000. On each iteration of the simulation, SigClust generates a single multivariate Gaussian null distribution, applies k-means clustering at a specific optimal k, and then estimates a test statistic, CI, operationalized as the ratio of within-class sum of squares divided by the total sum of squares relative to the mean. After all iterations of the simulation are complete, the actual CI from the real data is computed and a p-value is calculated as the proportion of CI values from the simulation that are smaller than the actual CI from the real dataset.

Longitudinal analyses

The *reval* modeling identifies subtypes from a single snapshot of earliest MSEL and VABS scores that represent LIMA features. To test whether the subtypes are different in later developmental trajectories in language, cognitive, motor, and adaptive domains, we ran longitudinal analyses on a subset of NDA individuals with repeated measures over time (see Supplementary Table 1 for sample sizes) (Fig. 1D). The reasoning behind conducting this analysis is that the subtypes within the A3D model are predicted to have different developmental trajectories and outcomes. Longitudinal modeling was implemented using a linear mixed effect model (e.g., *lmer* from the *lme4* R library), whereby MSEL age-equivalent or VABS standard scores were the dependent variable and age, subtype and the age*subtype interaction were fixed effects and with random effects of age with random intercepts and slopes grouped by subject. We also ran longitudinal modeling with a second independent dataset from UCSD (see Supplementary Table 1 for sample sizes). Here we applied the NDA subtyping model to earliest intake (between 24–68 months) MSEL and VABS scores from the UCSD dataset to generate subtype labels (Fig. 1E). Similar linear mixed effect modeling was used to test age, subtype, and age*subtype interactions as fixed effects and with age as random intercepts and slopes group by subject.

Neuroimaging and gene expression data acquisition and preprocessing

All structural and functional neuroimaging data utilized in this work has been previously published in work examining ELO stratification (i.e. fMRI [18, 19], sMRI [20]). In this work, we take all data utilized in this past work and re-analyze it with different autism subtype labels (i.e. A3D Type I vs. Type II). To briefly recap characteristics of this data and how it was initially collected, all neuroimaging

data was collected during natural sleep at UCSD on a 1.5 Tesla General Electric MRI scanner. High-resolution T1-weighted IR-FSPGR (inversion recovery fast-spoiled prepared gradient recalled) anatomical scans were used for analyses of structural morphometry (surface area, cortical thickness) and for use during normalization steps in preprocessing fMRI data (IR-FSPGR parameters: TE (echo time)=2.8 ms, TR (repetition time)=6.5 ms, flip angle=12°, bandwidth=31.25 kHz, field of view=24 cm, and slice thickness=1.2 mm). Cortical morphometric measurements such as surface area (SA) and cortical thickness (CT) were derived from Freesurfer processing of the anatomical T1 data [20]. These morphometric measures were estimated from a genetically-sensitive parcellation known as GCLUST [30, 31]. All SA and CT data analyzed here are identical to those reported in prior work [20]. fMRI data used in this work is identical to the data in previous papers on ELO [18, 19], and thus all details of the language paradigm and imaging acquisition parameters apply (i.e. single-echo echoplanar imaging (EPI) sequence; echo time=30 ms; repetition time=2500 ms; flip angle=90°; bandwidth=70 kHz; field of view=25.6 cm; in-plane resolution=4×4 mm; slice thickness=4 mm; 31 slices). Preprocessed data was then input into single-subject general linear models (GLM) and was utilized for re-analysis [18], but whereby new tests in this work specifically examine group differences using the new A3D stratification model (e.g., Type I vs Type II) and to compare the A3D stratification model to the previously reported ELO stratification model [18, 19]. Both whole-brain activation mapping per each group and region of interest (ROI) analyses for between-group comparisons were used. In both whole-brain and ROI analysis models, sex and age at scan were included as covariates of no interest. Whole-brain mapping was thresholded using a voxel-wise FDR correction at q < 0.05 [32]. The ROIs utilized in this analysis were the same 4 meta-analytic ROIs used in prior work [18, 19], which were derived from Neurosynth for the term 'language' and comprise left and right hemisphere temporal and lateral frontal cortices. Model comparison between ELO and A3D models was implemented using the Akaike Information Criterion (AIC) statistic. Models with lower AIC are considered better and we considered the model sufficiently better if the difference in AICs was greater than 10 [33].

In this work, we also re-analyzed neuroimaging-gene expression relationships with partial least squares (PLS) analysis [34], similar to previously reported papers on ELO stratification [19, 20]. The neuroimaging and blood leukocyte gene expression data re-analyzed was identical to those used in this past work [19, 20]. However, in the current work we examine these relationships in

the A3D stratification model (e.g., Type I vs. Type II). To briefly recap how this data was initially collected and analyzed, all blood gene expression data was collected at UCSD. Four to six milliliters of blood was collected into EDTA-coated tubes from toddlers on visits when they had no fever, cold, flu, infections, or other illnesses, or use of medications for illnesses 72 h before the blood draw. Blood samples were passed over a LeukoLOCK filter (Ambion) to capture and stabilize leukocytes and immediately placed in a -20 ℃ freezer. Total RNA was extracted according to standard procedures and manufacturer's instructions (Ambion). LeukoLOCK disks (Ambion cat no. 1933) were freed from RNAlater, and Tri-reagent (Ambion cat no. 9738) was used to flush out the captured lymphocytes and lyse the cells. RNA was subsequently precipitated with ethanol and purified though washing and cartridge-based steps. The quality of the mRNA samples was quantified according to the RNA Integrity Number (RIN). RIN values of 7.0 or greater were considered acceptable [35] and all processed RNA samples passed RIN quality control. Quantification of RNA was performed with a NanoDrop spectrophotometer (Thermo Scientific). Samples were prepared in 96-well plates at a concentration of 25 ng/µl.

RNA was assayed at Scripps Genomic Medicine for labeling, hybridization, and scanning with the Illumina BeadChip pipeline (Illumina) according to the manufacturer's instructions. All arrays were scanned with the Illumina BeadArray Reader and read into Illumina GenomeStudio software (version 1.1.1). Raw data were exported from Illumina GenomeStudio, and data preprocessing was performed with the *lumi* package [36] for R (http://www.R-project.org/) and Bioconductor (http://www.bioconductor.org/) [37]. Raw and normalized data are part of larger sets deposited in the Gene Expression Omnibus database (GSE42133 and GSE111175).

A larger primary dataset of blood leukocyte gene expression was available from 383 samples from 314 toddlers 1–4 years old. The samples were assayed with the Illumina microarray platform in three batches. The datasets were combined by matching the Illumina Probe ID and probe nucleotide sequences. The final set included a total of 20,194 gene probes. Quality-control analysis was performed to identify and remove 23 outlier samples from the dataset. Samples were marked as outliers if they showed low signal intensity (average signal two SD lower than the overall mean), deviant pairwise correlations, deviant cumulative distributions, deviant multidimensional scaling plots, or poor hierarchical clustering, as described elsewhere [38]. The high-quality dataset included 360 samples from 299 toddlers. High reproducibility was observed across technical replicates (mean Spearman correlation of 0.97 and median of 0.98). Thus,

we randomly removed one of each of two technical replicates from the primary dataset. The 20,194 probes were then collapsed to ~14,000 genes (14,313 for the fMRI-gene expression dataset [19]; 14,426 for the structural MRI-gene expression dataset [20]) on the basis of picking the probe with maximal mean expression across samples. Data were quantile normalized and then adjusted for batch effects, sex, and RIN. The batch-, sex-, and RIN-adjusted data were used in all further downstream analyses.

Weighted gene co-expression network analysis (WGCNA) analysis was used to reduce several thousands of highly correlated genes down to a less redundant number of gene co-expression modules to utilize in the PLS analysis. The WGCNA analysis is described in past papers [19, 20], and is briefly recapped here. WGCNA is implemented with the *WGCNA* library in R [39]. Correlation matrices are estimated with the robust correlation measure of biweight midcorrelation and then converted into adjacency matrices that retained the sign of the correlation. These adjacency matrices were then raised to a soft power of 16. This soft power was chosen by finding the first soft power for which a measure of R^2 scale-free topology-model fit saturated at least above $R^2 > 0.8$ [40], and the slope was between -1 and -2 [41]. The soft-power-thresholded adjacency matrix was then converted into a topological overlap matrix (TOM) and then a TOM dissimilarity matrix (1—TOM). The TOM dissimilarity matrix was then input into agglomerative hierarchical clustering with the average linkage method. Gene modules were defined from the resulting clustering tree, and branches were cut with a hybrid dynamic tree-cutting algorithm (deepSplit parameter=4). Modules were merged at a cut height of 0.2, and the minimum module size was set to 100. Only genes with a module membership of $r > 0.3$ were retained within modules. For each gene module, a summary measure called the module eigengene (ME) was computed as the first principal component of the scaled (standardized) module expression profiles, and these ME values were utilized in further PLS analyses. We also computed module membership for each gene and module. Module membership indicated the correlation between each gene and the ME. Genes that could not be clustered into any specific module were left within the M0 module, and this module was not considered in any further analyses.

Partial least squares (PLS) analysis [34] was utilized to examine neuroimaging-gene expression relationships and was implemented in an identical manner as prior work [19, 20]. To briefly recap, PLS allows for identification of latent imaging-gene expression variable pairs (LV) that maximally explained covariation in the dataset and which are uncorrelated with other latent variable pairs.

Mandelli *et al. Molecular Autism*   (2024) 15:41

Page 8 of 16

The strength of such covariation was denoted by the singular value (*d*) for each brain–gene expression LV, and hypothesis tests were made via permutation tests on the singular values. Furthermore, identifying brain regions that most strongly contribute to each LV pair was made via bootstrapping, whereby a bootstrap ratio was created for each voxel and represented the reliability of that voxel for contributing strongly to the LV pattern identified. The bootstrap ratio was roughly equivalent to a *Z* statistic and could be used to threshold data to find voxels or parcels that reliably contribute to an LV pair. The PLS analyses reported here were implemented within the *plsgui* MATLAB toolbox (http://www.rotman-baycrest.on.ca/pls/). The imaging data for fMRI was single-subject t-statistic maps, while for structural MRI imaging data GCLUST parcellated surface area and cortical thickness measures were used. WGCNA ME values were inserted for the gene expression data in the PLS analysis. For statistical inference on identified imaging-gene expression LV pairs, a permutation test was run with 10,000 permutations. To identify reliably contributing voxels/parcels for imaging-gene expression LVs and to compute 95% confidence intervals (CI) on correlations, bootstrapping was used with 10,000 resamples. Bootstrap ratios (BSR) were used to visualize patterns across the brain, with larger BSRs indicating stronger and more reliable contributions to significant imaging-gene expression LVs. Gene coexpression modules in which the 95% bootstrap CIs did not encompass 0 were denoted 'non-zero' association modules. All other modules in which the 95% CIs included 0 were denoted 'zero' modules.

Gene set enrichment analyses

Our PLS analyses highlight non-zero modules that significantly contribute the imaging-gene expression association effects. Thus, we utilized genes from non-zero modules and then tested them for enrichment with an independent gene list of attenuated transcriptomic regional identity (ARI) genes, isolated in post-mortem cortical tissue of autistic individuals [42]. This analysis was done to test how genes isolated assayed from blood leukocytes and associated with imaging phenotypes in autism may overlap with genes isolated as dysregulated in their cortical patterning directly in cortical tissue of autistic patients. Gene set enrichment analyses were implemented using custom R code written by MVL that computes hypergeometric *P* values and enrichment odds ratios (ORs). The background pool for these enrichment tests was set to the number of total genes assessed in our prior papers (e.g., 14,313 [19]; 14,426 [20]). After all enrichment tests were computed, results are interpreted only if the enrichment was statistically significant after

false discovery rate (FDR) correction for multiple comparisons at a threshold of $q < 0.05$.

Results

Detection of Type I versus Type II autisms in early development

The first objective of this work was to test whether it is possible to detect autism Type I and Type II subtypes in early development, as predicted by the A3D model. Stability-based relative clustering validation (*reval*) [17] identified a two-cluster solution as the optimal solution and which generalizes with 98% accuracy in the independent NDA validation dataset (Fig. 2A). SigClust analysis confirms that this solution is indeed indicative of true clusters that heavily deviate from the null hypothesis of that the data come from one originating multivariate Gaussian distribution ($p = 9.99e-5$; Fig. 2B). The subtypes conform to the predicted Type I ('disability') versus Type II ('difference') distinction, as they show a marked separation between high versus very low scores across all LIMA features (Fig. 2C). Notably, the distinction here is not based on hard cutoffs. Rather, distributions for all subscales/domains overlap to some extent, but effect sizes remain rather large throughout (Cohen's d > 0.96) (Supplementary Table 3). The fact that distributions overlap and yet a classifier can still make accurate labelings with 98% accuracy indicates the robustness of such an approach over-and-above other heuristic methods typically used in clinical practice such as hard cutoff thresholding.

Differential developmental trajectories for Type I vs Type II autisms

We next tested if A3D Type I versus Type II autisms differed in terms of subsequent developmental trajectories measured over the first decade of life (Fig. 2D–E). For each MSEL subscale, we found highly significant age*subtype interactions in both NDA and UCSD ACE datasets (Supplementary Tables 4–5). This interaction effect can be described by slower and less steep trajectories for Type I compared to Type II and indicates that Type II autism develops language, fine motor, and non-verbal cognitive skills at a much faster rate. For VABS trajectories we did not find any significant age*subtype interactions in the NDA dataset (Supplementary Table 4). In contrast, significant age*subtype interactions appear in the UCSD ACE dataset for all VABS domains except for the motor domain (Supplementary Table 5). The lack of an effect for VABS motor compared to a conceptually related MSEL FM subscale may be explained by differences in how MSEL versus VABS data are measured (child's performance versus parent-report), true non-shared overlap in what VABS motor versus MSEL
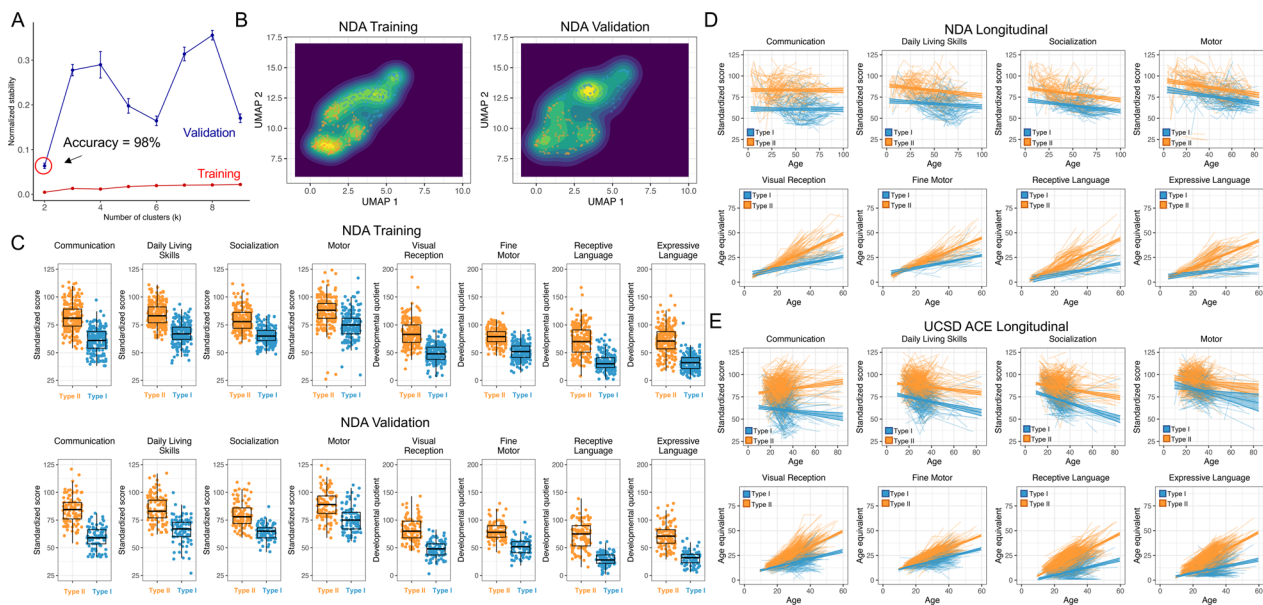
Mandelli *et al. Molecular Autism*   (2024) 15:41

Page 9 of 16

Fig. 2 Robust and highly generalizable early distinctions between Type I versus Type II autisms. Panel **A** shows that a 2-subtype solution is the optimal clustering solution that minimizes normalized stability via stability-based relative clustering validation (*reval*). This solution generalizes with 98% accuracy in the held-out NDA validation dataset. Panel **B** shows UMAP density plots that depict separated peaks of distributions for Type I (blue) and Type II (orange) autism subtypes. Panel **C** plots the standardized scores across all MSEL and VABS features to describe how subtypes manifest as relatively low (Type I; blue) versus high (Type II; orange) scores across all features. Panels D-E show plots of developmental trajectories for the Type I (blue) and Type II (orange) autisms, respectively for NDA (**D**) and UCSD ACE (**E**) datasets

FM measure, or differences in the age range analyzed between VABS and MSEL. Notably, the larger longitudinal sample size for the UCSD ACE dataset (n = 1487) may imply that such subtle yet significant effects here were not detected in NDA because of lower statistical power. VABS trajectory differences can be described as Type II autism having much less steep decline in skills over time compared to Type I autism. Overall, these results showcase that Type I and Type II autisms can be characterized by not just initial large differences in early LIMA features, but also by differential rates of development over the first decade of life. Thus, if Type I and Type II autisms are a distinction between profound disability versus difference, these results show that those descriptions are increasingly more prominent as development progresses.

Differential functional neural responses to speech
In the next analysis we examined fMRI responses to speech between A3D Type I and Type II autisms. Prior work using early language outcome (ELO) subtype labels showed that autism with poor ELO does not robustly activate language-sensitive superior temporal cortices [18, 19]. Since the same fMRI data was re-analyzed in this work, we first examined how well the A3D labels might correspond to the ELO subtype labels. In contrast to the ELO model, the A3D model breaks down into a somewhat different distinction ($\chi^2 = 14.21$, $p = 0.0001$)

(Supplementary Table 6). Type I individuals in A3D are about 17% of all autistic individuals in the fMRI dataset, and all Type I individuals in A3D would be labeled as the Poor subtype of the ELO model. In contrast, all individuals labeled in the Good subtype under the ELO model fell into the Type II A3D subtype (Supplementary Table 6). This contrast between models indicates that the Type I label from A3D is a much more conservative label than the 'Poor' label from the ELO model.

Despite the considerable difference in subtype labels between the models, whole-brain maps show a quite similar activation patterns in each subtype between ELO and A3D models. The robust activation patterns observed in TD and LD/DD non-autistic comparison groups are largely preserved, particularly in left hemisphere superior temporal cortex, in the ELO Good and A3D Type II subtypes. In contrast, very little is identified in the ELO Poor and A3D Type I subtypes (Fig. 3A). This qualitatively may suggest that very similar fMRI differences are apparent across the ELO and A3D models. To test this in a more targeted way for specific brain regions sensitive to language, we conducted region of interest (ROI) analyses for 4 regions (bilateral temporal and frontal cortex regions) that meta-analytic evidence shows are sensitive to language processing [18, 19]. Here we find robust evidence that
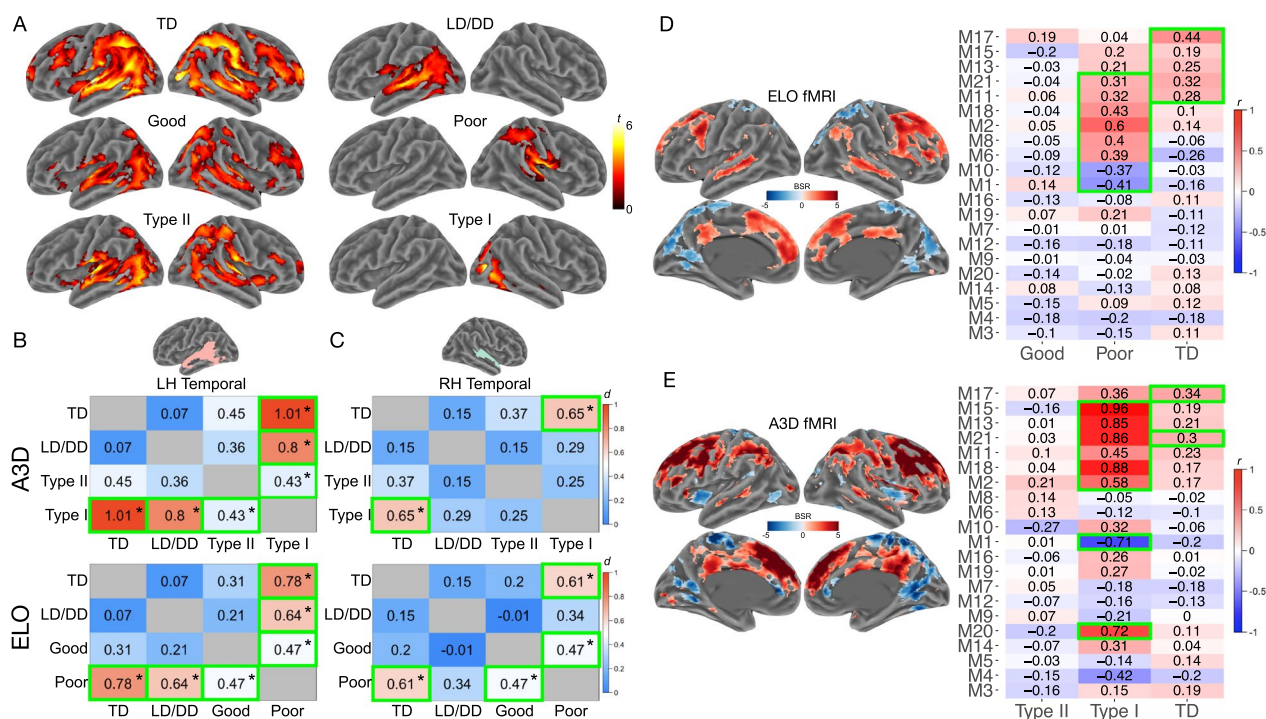
Fig. 3 Functional neural response to speech and associations with gene expression. Panel **A** depicts whole-brain activation analysis maps for each of the groups under study. All maps are thresholded at voxel-FDR q < 0.05, except for LD/DD, Poor and Type I groups, which are thresholded at cluster-forming p < 0.05 and cluster FDR q < 0.05 for visualization purposes. Panel **B** depicts standardized effect sizes (Cohen's d) for the LH temporal ROI for all pairwise group comparisons for A3D (top) and ELO (bottom) stratification models. The asterisks and green outlined cells indicate statistical significance at p < 0.05. Similar effect size heatmaps are shown in panel **C** for the RH temporal ROI. Panels **D** and **E** show results for the PLS analysis of the statistically significant LV1 fMRI-gene expression relationship identified in the ELO (**D**) and A3D (**E**) models. The whole-brain map in these panels show brain bootstrap ratios (BSR) that indicate which voxels most strongly contribute to the fMRI-gene expression LV1 relationship. The heatmaps to the right on each panel shows each gene co-expression module along the rows and groups along the columns. The modules of importance are highlighted with green outlines (non-zero modules) as they contain a correlation with fMRI response to speech whose 95% bootstrap CIs do not encompass a correlation of zero

the ELO pattern of the Poor subtype showing a lack of left hemisphere superior temporal cortex (LH temporal) response to speech, is similarly evident in the Type I subtype from A3D, but with somewhat larger effect sizes in A3D (Fig. 3B; Supplementary Table 7A). Model comparisons with the Akaike Information Criterion (AIC) show that ELO and A3D models are equally good at explaining variance in functional neural responses to speech (*ELO AIC* = − 446.59; *A3D AIC* = − 447.92, Supplementary Table 7B). The slightly lower AIC for the A3D model may be due to marginally stronger effect sizes for TD vs. Type I and LD/DD vs. Type I, compared to the ELO TD vs. Poor and LD/DD vs. Poor comparisons (Fig. 3B). Similar effects also emerge with ELO and A3D models showing very similar AIC values in RH superior temporal cortex (RH temporal) (Fig. 3C; Supplementary Table 7B), while little to no group differences are apparent in LH and RH frontal ROIs for either ELO or A3D (Supplementary Table 7B).

Differential large-scale associations between fMRI response and gene expression

We next compared how functional neural response to speech is related to gene expression patterns measured in blood leukocytes. In our prior work on the ELO model, we found evidence for a large-scale fMRI-gene expression relationship whereby such relationships are relatively distinct between TD and autism ELO subtypes [19] (Fig. 3D). The same analysis applied to the A3D model reveals a very similar relationship (LV1: *d* = 110.73, *p* = 1.99e−4, *percentage covariance explained* = 34.16%), comprising many of the same frontal and temporal regions, but additionally showing prominent contributions of lateral parietal areas connected by the arcuate fasciculus in language-sensitive areas [43] (Fig. 3E). Additionally, the correlations between specific non-zero gene co-expression modules apparent in the Type I subtype (Fig. 3E) are much stronger than similar relationships observed in the ELO Poor subtype (Fig. 3D). Thus, much like the prior analysis showing similar differences in fMRI

activation response to speech, associations with functional genomic mechanisms also appear to be quite similar in A3D vs ELO, albeit with slightly stronger effects in A3D.

Differential genomic cortical patterning of cortical surface area

In prior work we also reported that genomic cortical patterning of surface area and cortical thickness manifest differently in ELO subtypes [20]. In the next analysis, we revisit that same dataset and apply the A3D model. Similar to the comparison of fMRI ELO vs. A3D models in terms of sample sizes, in this analysis of structural imaging phenotypes the A3D model was again more conservative in labeling individuals as Type I (17% of all autistic individuals) compared to the ELO model Good vs Poor distinction ($\chi^2 = 13.36$, $p = 0.0002$) (Supplementary Table 6). PLS analyses of surface area in the ELO model revealed on one significant LV pair, interpreted as a normative cortical patterning effect with similar associations found in the TD and autism Good subtype, whereas such patterns were absent in the autism Poor ELO subtype [20] (Fig. 4A). In contrast, analysis of the A3D model reveals 2 significant LV pairs, with LV1 being an effect exclusive to the Type I subtype (LV1: $d = 4.22$, $p = 1.99e-4$,

percentage covariance explained $= 26.78\%$; Fig. 4B). The LV2 effect recapitulates the normative surface area patterning effect found in the ELO model (LV2: $d = 3.57$, $p = 1.99e-4$, *percentage covariance explained* $= 19.21\%$; Fig. 4C), whereby similar correlations are observed in TD and Type II autism. However, LV2 also reveals opposite effects occurring for Type I autism in some modules (e.g., M4, M6), and additional modules with important nonzero effects that are exclusive to Type I (e.g., M11, M12). Thus, although retaining a similar normative surface area patterning effect, the A3D model is additionally sensitive to an effect for how surface is atypically patterned specific to the Type I autism subtype.

Differential genomic cortical patterning of cortical thickness

Next, we looked at genomic patterning of cortical thickness. The ELO model had previously revealed 2 different effects on cortical thickness—LV1 being a normative effect with similar associations in TD and the autism Good ELO subtype (Fig. 5A) and LV2 being an atypical effect on cortical thickness specific to the autism Poor ELO subtype [20] (Fig. 5B). The A3D model reveals only one significant LV pair (A3D CT LV1: $d = 4.98$, $p = 9.99e-5$, *percentage covariance explained* $= 35.33\%$;
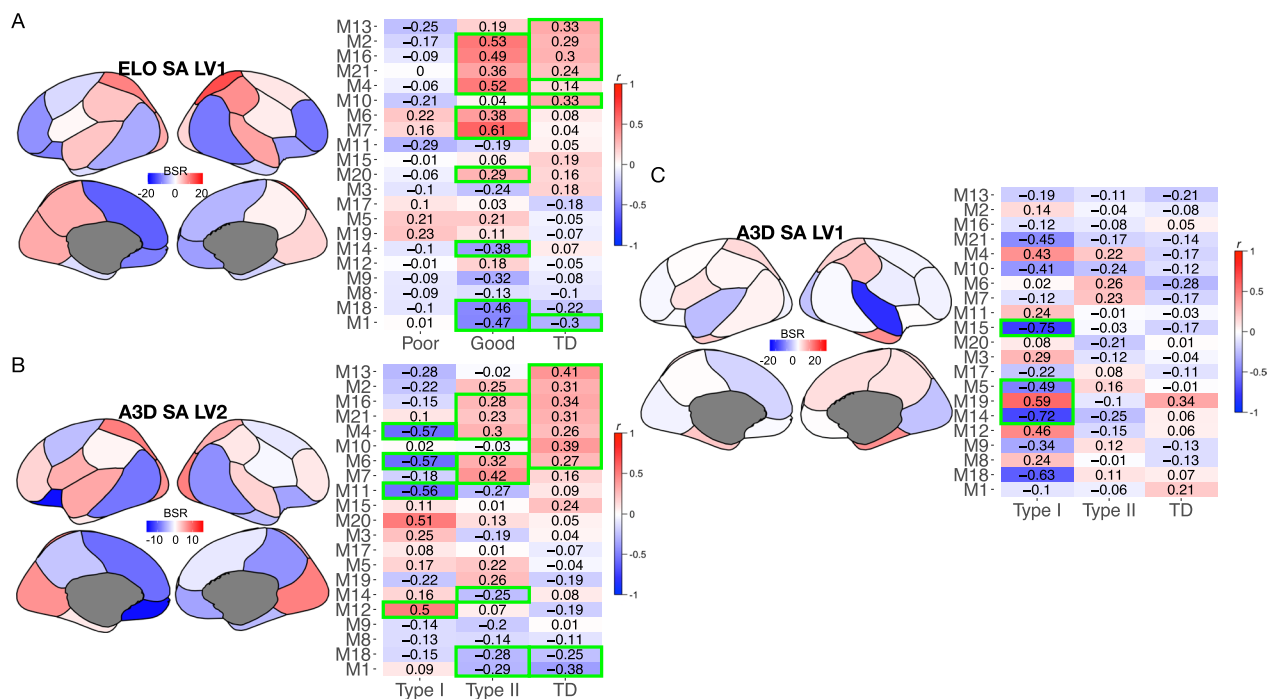

Fig. 4 Gene expression associations with cortical surface area. Results for PLS analyses on the phenotype of cortical surface area for the ELO (**A**) or A3D (**B**, A3D SA LV2; **C**, A3D SA LV1) models. The parcellated brain maps show brain bootstrap ratios (BSR), whereby more extreme BSRs indicate brain regions that more strongly contribute to the LV relationship. The heatmap on the right of each panel shows correlations with each gene co-expression module along the rows and each group along the columns. The important non-zero modules are highlighted as cells with green outlines
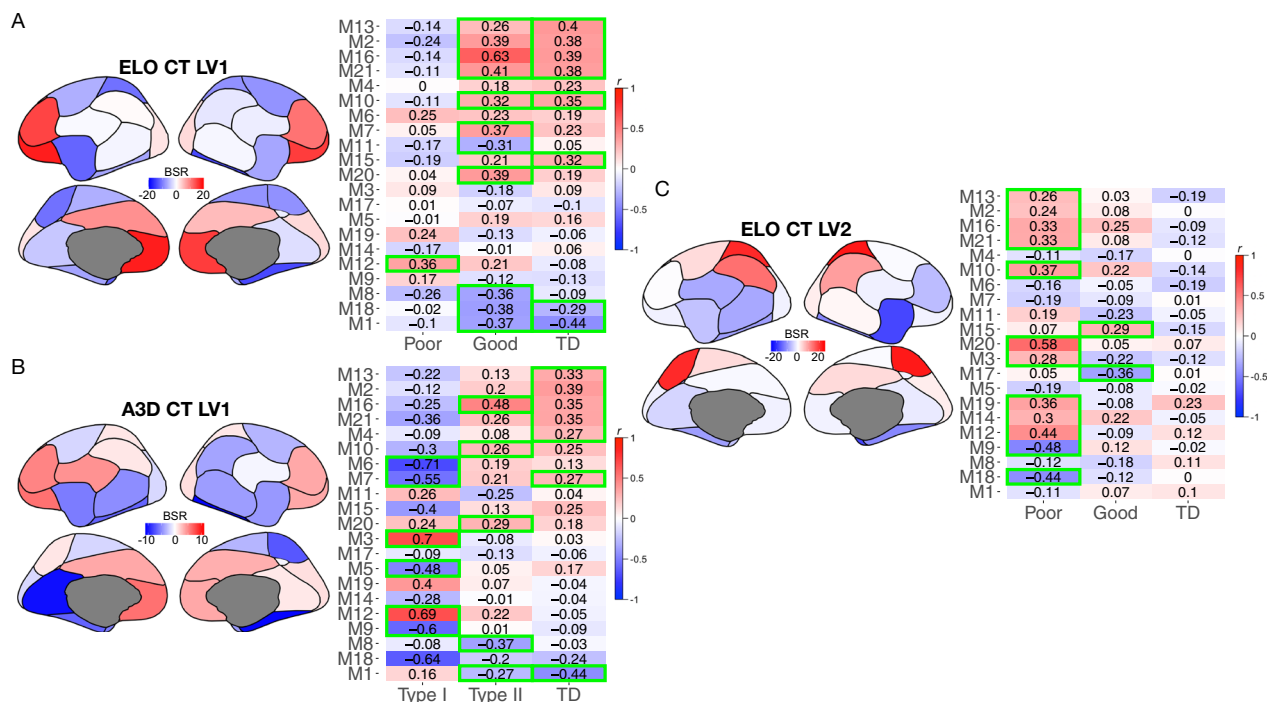
Fig. 5 Gene expression associations with cortical thickness. Results for PLS analyses on the phenotype of cortical thickness for the ELO (**A**, ELO CT LV1; **C**, ELO CT LV2) or A3D (**B**, A3D CT LV1) models. The parcellated brain maps show brain bootstrap ratios (BSR), whereby more extreme BSRs indicate brain regions that more strongly contribute to the LV relationship. The heatmap on the right of each panel shows correlations with each gene co-expression module along the rows and each group along the columns. The important non-zero modules are highlighted as cells with green outlines

Fig. 5C). However, A3D CT LV1 is an effect that encompasses both the normative effect (e.g., similar non-zero modules (M1, M16) shared by TD and Type II autism) and an atypical effect specific to Type I autism (e.g., non-zero modules that are unique to Type I (M3, M5, M6, M9, M12), or which appear in Type I but in the opposite directional (M7) from TD and Type II). Qualitatively, similar non-zero modules appear present for both TD and Good or Type II subtypes in both models. In contrast, somewhat different non-zero modules appear that are unique to the Poor group in ELO CT LV2 compared to Type I in A3D CT LV1.

Overlap with genomic cortical patterning effects in post-mortem cortical tissue

A caveat to the analyses showing large-scale associations (e.g., cortical patterning effects) between imaging phenotypes and gene expression is that these data deal with gene expression patterns measured peripherally in blood leukocytes. However, a recent paper on broad transcriptomic dysregulation across cortical regions in post-mortem cortical tissue reveals a potentially similar kind of patterning effect for how gene expression is dysregulated across the cortex in autism [42]. Thus, we took the opportunity here to assess whether there is indeed overlap between genes that show an atypical transcriptomic regional identity (ARI) effect in post-mortem cortical tissue and the genes that pop up in our imaging-gene expression association analyses (i.e. genes from non-zero modules). Here we find that fMRI non-zero module genes from both the ELO and A3D model significantly overlap with ARI genes identified in post-mortem cortical tissue [42] (Fig. 6A). With regard to structural cortical patterning of surface area and cortical thickness, we also find significant overlap between the ARI gene set and genes from PLS effects that are interpreted as showing normative patterning effects in TD and ELO Good or A3D Type II subtypes. In contrast, PLS LVs interpreted as specific atypical effects in ELO Poor or A3D Type I subtypes do not show strong overlap with the ARI gene set (Fig. 6B; Supplementary Table 8). The A3D CT LV1 effect, characterized by a mixture of shared normative effects between TD and Type II versus specific atypical effects in Type I, was an exception, as the enrichment with ARI genes was trending on significance after FDR correction (*FDR q* = 0.055; Supplementary Table 8). Thus, these analyses are primary validation for the idea that imaging-blood gene expression associations revealed here and in our prior work can pick up on similar genes that when measured in cortical tissue, show atypical gene expression
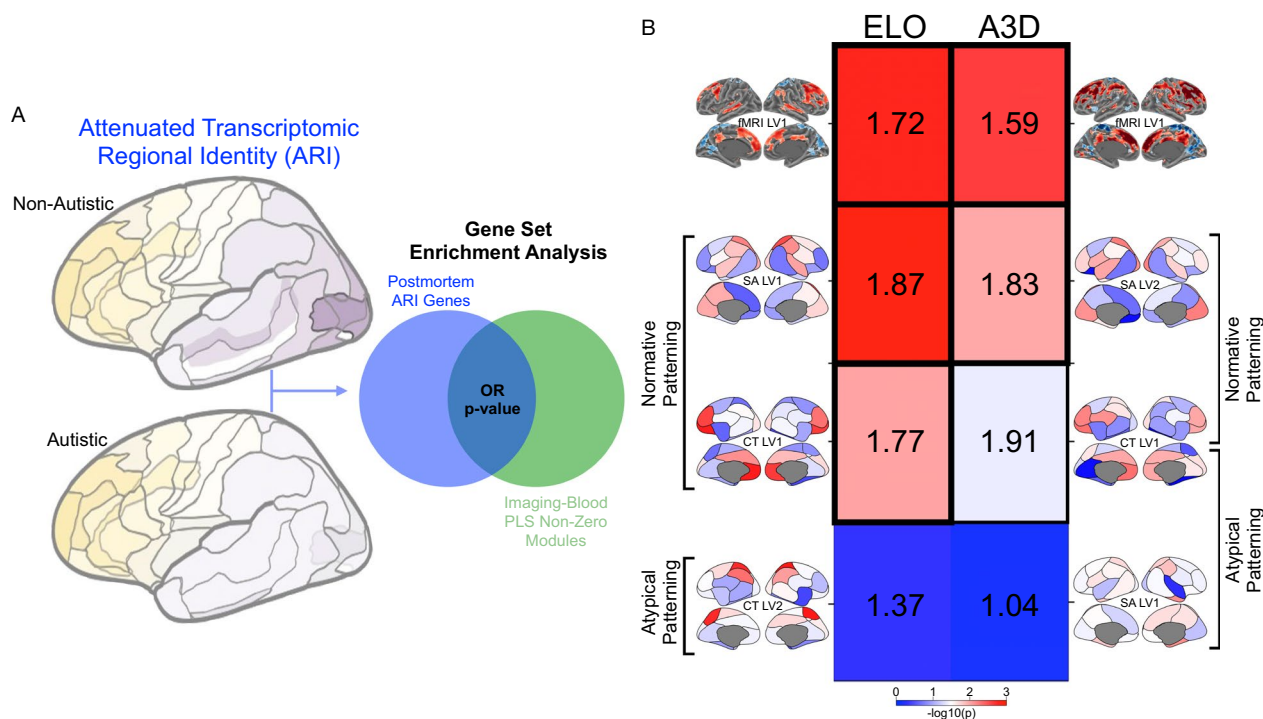
Fig. 6 Enrichment analysis showing overlap between attenuated transcriptomic regional identity (ARI) genes detected in post-mortem cortical tissue and non-zero modules detected in imaging-gene expression PLS analyses. Panel **A** depicts a schematic of how these enrichment analyses were conducted. A gene set known as attenuated transcriptomic regional identity (ARI) genes were extracted from a prior study on post-mortem cortical tissue in autistic patients [42]. These ARI genes are genes that have substantial between-region gene expression differences in non-autistic brains, but much more attenuated regional identity differences in gene expression in autism. ARI genes are depicted in blue in the Venn diagram. In green, the Venn diagram shows our set of genes isolated from PLS imaging-gene expression association analyses (i.e. genes from non-zero modules). The degree of overlap is then tested with enrichment odds ratios and hypergeometric *p* values from gene set enrichment analyses. Panel **B** shows a heatmap of the enrichment odds ratios (numbers in the center of each cell) and where color in each cell indicates the -log10(p-value) for each hypergeometric test. Cells with a thick black outline indicate enrichment tests that pass at FDR q < 0.05, while cells with smaller thin black outline pass at FDR q < 0.1. The columns in the heatmap represent comparisons when the gene list derives from the ELO (left) or A3D (right) model. Each row of the heatmap indicates a different gene list extracted from PLS analyses, with the top row indicating genes from non-zero modules in PLS analyses that examine fMRI responses to speech. The remaining rows depict comparisons where the PLS gene lists come from associations with surface area (SA) or cortical thickness (CT) and we have annotated which rows can be interpreted as effects indicative of 'normative patterning' effects (effects driven by TD correlations and which are shared in Good ELO or Type II A3D subtypes) versus atypical patterning effects (e.g., effects driven specifically by the ELO Poor or Type I A3D subtypes)

patterning dysregulation in autism. In particular, when our analyses of blood gene expression identify normative patterning effects shared in TD and ELO Good or A3D Type II subtypes, these effects are quite similar to those appearing in the ARI gene list from post-mortem cortical tissue.

Discussion

This work offers a key proof-of-concept for the A3D model approach to stratifying autism into phenotypically, biologically, and developmentally-meaningful subtypes. The A3D model suggests that a first-level split in the autism population should be between subtypes differentially characterized by profound disability (Type I) versus difference (Type II) and that such a distinction is evident

from early in life and continually emerges throughout later development. Non-core LIMA features form the basis for differentiating Type I versus Type II autisms under A3D. An unsupervised data-driven modeling approach can identify these types in an automated fashion with high degree of accuracy in generalizing to new data. An important aspect of our unsupervised subtyping approach was the ability to validate data-driven discovery in independent data and empirically estimate generalization accuracy. Further validation of the meaningfulness of these phenotypically-defined subtypes was also determined within independent datasets that allow tests for developmental and biological significance. These aspects of our approach to validation are commonly missing across the literature using unsupervised data-driven

Mandelli *et al. Molecular Autism*   (2024) 15:41

Page 14 of 16

approaches for subtyping [15]. Given these key validations, we have immediately translated our unsupervised data-driven discoveries into a supervised classification model that can be reproducibly used by the field in future work. Our stratification prediction model is available as a free web application that will allow researchers to insert their own MSEL and VABS data and immediately get subtype labels as output (https://landiit.shinyapps.io/Autisms3D/).

In addition to providing a robust, reproducible, and highly generalizable prediction model for detecting Type I versus Type II autisms, we have also provided replicable evidence in two independent longitudinal datasets demonstrating that identification of such subtypes at early points in development is predictive of later differences in developmental trajectories over the first decade of life. This key feature of the results brings together the 'developmental' component of A3D by showing that the subtypes can be identified early and that such early generated subtype labels are indicative of different developmental paths as they grow older. The current results are somewhat similar to past work stratifying autism using only the VABS [14]. At early ages, VABS allows for a stratification into 3 subtypes that differ developmentally [14] in ways that are somewhat similar to those described in this work. However, with only information provided from the VABS, at early ages some autistic individuals may be classified into an intermediate subtype whereby later life outcomes are less certain [14]. By adding information from the MSEL to measure the full set of LIMA features, this intermediate subtype identified from the VABS alone may now be resolved as either Type I or II under the A3D framework.

In this work, we also tested the A3D model to determine whether the subtype labels are sensitive to differences at the level of neuroimaging phenotypes and relationships to gene expression. The A3D model highlights several distinctions that are very similar in nature to the previous ELO stratification model [18–20]. The similar biological sensitivity of A3D and ELO is important, since the two models do differ substantially in terms of labels, with A3D potentially being much more conservative. Some subtle and qualitative differences did emerge between A3D and ELO with respect to how neuroimaging phenotypes are related to gene expression. First, regarding fMRI-gene expression associations, it was clear that the A3D model generally produces associations in Type I that are a bit stronger than those seen in the ELO Poor subtype. Second, regarding surface area-gene expression associations, A3D revealed two kinds of associations—one being a normative effect shared between TD and Type II, and a separate atypical effect apparent only in Type I. This can be contrasted to only

a normative genomic patterning effect on surface area in the ELO model. Finally, whereas CT-gene expression associations in the ELO model comprised two orthogonal effects—one normative and shared between TD and Good subtype, while another is atypical specifically in the Poor subtype—the A3D model identifies only one latent variable pair that captures both the CT normative patterning effect in TD and Type II, but also capture atypical patterning in different co-expression modules for Type I. These types of subtle and qualitative differences may be important for distinguishing A3D from the ELO model. Given that A3D tends to identify somewhat larger effects alongside unique imaging-gene expression relationships for Type I and/or relationships with opposing directionality in Type I compared to Type II and TD, it could be inferred that A3D may be better than ELO at highlighting biology in the most profoundly affected autistic individuals compared to the ELO model. However, we also emphasize the large degree of similarities between results from A3D and ELO. This overlap illustrates that both ELO and A3D are detecting differential biology that is otherwise hidden from case–control modeling.

Limitations
There are some limitations and caveats to underscore. First, the neuroimaging results are derived from relatively small sample sizes. Thus, future work should attempt to replicate and extend these findings in larger samples. Facilitating such future work will be the utilization of our app (https://landiit.shinyapps.io/Autisms3D/) to fully reproduce the subtype labels in new datasets. Second, treatment history data was largely not available for most of the data utilized and therefore it is unclear to what extent early intervention treatment might impact the subtype labels. Future work examining the subtypes in the context of early intervention research will be fundamentally important for this question. This future work on treatment will enable a much larger view behind whether the subtypes are important across the BOT objectives and will be an important test regarding whether the subtypes are indeed indicative of truly distinctive latent entities. Third, while the A3D model is based on all four LIMA features, fMRI speech-related responses tested in this work are primarily relevant to one of the LIMA features (i.e. language). Future work could extend the investigation of the neural underpinnings to other domains (e.g., resting state, motor, other cognitive domains). Fourth, the current work requires both MSEL and VABS to identify subtypes. However, in many contexts, researchers may characterize LIMA features with other measures. Future work is necessary to extend and validate the A3D model with a range of other psychometric

Mandelli *et al. Molecular Autism*   (2024) 15:41

Page 15 of 16

instruments in order to allow more flexibility beyond the MSEL and VABS.

Conclusions

To conclude, we demonstrate that the autism Type I and Type II subtypes, as described by A3D theoretical framework, are present in early development and that their early developmental trajectories are markedly different in language, motor, intellectual and adaptive skills. The A3D model may also get us closer to a more precision understanding of the neurobiology behind different types of autistic individuals.

Abbreviations

LIMA	Language, intellectual, motor, and adaptive functioning
MSEL	Mullen Scales of Early Learning
VABS	Vineland Adaptive Behavior Scales
NDA	National Institute of Mental Health Data Archive
UCSD	University of California, San Diego
UCSD ACE	University of California, San Diego Autism Center of Excellence
SC	Social-communication
RRB	Restricted repetitive behaviors
BOT	Biology, outcomes, treatment
AUTISMS-3D	Disability versus difference over development
A3D	AUTISMS-3D
DSM-IV	Diagnostic and Statistical Manual IV edition
PDD-NOS	Pervasive developmental disorder not otherwise specified
IQ	Intellectual quotient
ELO	Early language outcomes
reval	Stability-based relative clustering validation
VR	MSEL Visual Reception subscale
FM	MSEL Fine Motor subscale
RL	MSEL Receptive Language subscale
EL	MSEL Expressive Language subscale
ADOS-2	Autism Diagnostic Observation Schedule
fMRI	Functional magnetic resonance imaging
MRI	Magnetic resonance imaging
TD	Typically-developing
LD/DD	Language or developmental delay
DQ	Developmental quotient
UMAP	Uniform Manifold Approximation and Projection
IR-FSPGR	Inversion recovery fast-spoiled prepared gradient recalled
SA	Surface area
CT	Cortical thickness
GCLUST	Genetically sensitive brain parcellation of SA and CT
GLM	General linear model
ROI	Region of interest
AIC	Akaike Information Criterion
PLS	Partial least squares
RNA	Ribonucleic acid
RIN	RNA Integrity Number
WGCNA	Weighted gene coexpression network analysis
TOM	Topological overlap matrix
LV	Latent variable
BSR	Bootstrap ratios
ARI	Attenuated transcriptomic regional identity
OR	Odds ratio
FDR	False discovery rate
GUID	NDA global unique identifiers
LH	Left hemisphere
RH	Right hemisphere

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13229-024-00613-5.

> Supplementary Material 1

Data availability
All data utilized in this work can be found in the National Institute of Mental Health Data Archive (NDA; https://nda.nih.gov). NDA global unique identifiers (GUID) and collection IDs are provided in Supplementary Table 2. UCSD ACE data are available on NDA through collection IDs 9, 2466, 2968, 2290 and 2115.

Code availability
Analysis code to reproduce the analyses and figures is available on GitHub (https://github.com/IIT-LAND/a3d_msel_vabs). The *reval* Python library can be found on GitHub (https://github.com/IIT-LAND/reval_clustering) and the documentation can be found at https://reval.readthedocs.io. The Shiny app that allows users to input their own MSEL and VABS data and get subtype label predictions can be found at https://landiit.shinyapps.io/Autisms3D/.

Declarations

Ethics approval and consent to participate
This study was approved by the Institutional Review Board at University of California, San Diego. Informed written consent and assent were obtained from all parents and participants.

Competing interests
MVL is an Associate Editor of Molecular Autism and is one of the handling editors for the Collection in this journal entitled 'Neuroimaging in Autism Spectrum Disorders'. All other authors have no competing interests to declare.

Author details
[1]Laboratory for Autism and Neurodevelopmental Disorders, Center for Neuroscience and Cognitive Systems, Istituto Italiano di Tecnologia, Rovereto, Italy. [2]Center for Mind/Brain Sciences, University of Trento, Rovereto, Italy. [3]Department of Psychiatry, University of California, San Diego, La Jolla, CA, USA. [4]Education, and Clinical Center, VISN 22 Mental Illness Research, VA San Diego Healthcare System, San Diego, CA, USA. [5]Autism Center of Excellence, Department of Neurosciences, University of California, San Diego, La Jolla, CA, USA.

Mandelli *et al. Molecular Autism* (2024) 15:41

Page 16 of 16

References

1. Lombardo MV, Mandelli V. Rethinking our concepts and assumptions about autism. Front Psychiatry. 2022;3(13): 903489.
2. Happé F, Frith U. Annual research review: looking back to look forward—changes in the concept of autism and implications for future research. J Child Psychol Psychiatry. 2020;61(3):218–32.
3. Lombardo MV, Lai MC, Baron-Cohen S. Big data approaches to decomposing heterogeneity across the autism spectrum. Mol Psychiatry. 2019;24:1435.
4. Maenner MJ, Warren Z, Williams AR, Amoakohene E, Bakian AV, Bilder DA, et al. Prevalence and characteristics of autism spectrum disorder among children aged 8 years - autism and developmental disabilities monitoring network, 11 sites, United States, 2020. MMWR Surveill Summ. 2023;72(2):1–14.
5. Rødgaard EM, Jensen K, Vergnes JN, Soulières I, Mottron L. Temporal changes in effect sizes of studies comparing individuals with and without autism: a meta-analysis. JAMA Psychiat. 2019;76(11):1124–32.
6. Pellicano E, den Houting J. Annual research review: shifting from "normal science" to neurodiversity in autism science. J Child Psychol Psychiatry. 2021;63:381.
7. Singer A, Lutz A, Escher J, Halladay A. A full semantic toolbox is essential for autism research and practice to thrive. Autism Res. 2022;16:497.
8. Mottron L. A radical change in our autism research strategy is needed: back to prototypes. Autism Res. 2021;14(10):2213–20.
9. Lord C, Charman T, Havdahl A, Carbone P, Anagnostou E, Boyd B, et al. The lancet commission on the future of care and clinical research in autism. Lancet. 2021;S0140–6736(21):01541–5.
10. Lai MC, Lombardo MV, Chakrabarti B, Baron-Cohen S. Subgrouping the autism "spectrum": reflections on DSM-5. PLoS Biol. 2013;11(4): e1001544.
11. Constantino JN, Todd RD. Autistic traits in the general population: a twin study. Arch Gen Psychiatry. 2003;60(5):524–30.
12. Constantino JN, Charman T, Jones EJH. Clinical and translational implications of an emerging developmental substructure for autism. Annu Rev Clin Psychol. 2021;17(1):365–89.
13. Rosen NE, Lord C, Volkmar FR. The diagnosis of autism: from kanner to DSM-III to DSM-5 and beyond. J Autism Dev Disord. 2021;51(12):4253–70.
14. Mandelli V, Landi I, Busuoli EM, Courchesne E, Pierce K, Lombardo MV. Prognostic early snapshot stratification of autism based on adaptive functioning. Nat Mental Health. 2023;1(5):327–36.
15. van Agelink Rentergem JA, Deserno MK, Geurts HM. Validation strategies for subtypes in psychiatry: a systematic review of research on autism spectrum disorder. Clin Psychol Rev. 2021;87:102033.
16. Lord C, Petkova E, Hus V, Gan W, Lu F, Martin DM, et al. A multisite study of the clinical diagnosis of different autism spectrum disorders. Arch Gen Psychiatry. 2012;69(3):306–13.
17. Landi I, Mandelli V, Lombardo MV. Reval: a python package to determine best clustering solutions with stability-based relative clustering validation. Patterns (N Y). 2021;2(4):100228.
18. Lombardo MV, Pierce K, Eyler LT, Carter Barnes C, Ahrens-Barbeau C, Solso S, et al. Different functional neural substrates for good and poor language outcome in autism. Neuron. 2015;86(2):567–77.
19. Lombardo MV, Pramparo T, Gazestani V, Warrier V, Bethlehem RAI, Carter Barnes C, et al. Large-scale associations between the leukocyte transcriptome and BOLD responses to speech differ in autism early language outcome subtypes. Nat Neurosci. 2018;21(12):1680–8.
20. Lombardo MV, Eyler L, Pramparo T, Gazestani VH, Hagler DJ, Chen CH, et al. Atypical genomic cortical patterning in autism with poor early language outcome. Sci Adv. 2021;7(36):eahl663.
21. Xiao Y, Wen TH, Kupis L, Eyler LT, Goel D, Vaux K, et al. Neural responses to affective speech, including motherese, map onto clinical and social eye tracking profiles in toddlers with ASD. Nat Hum Behav. 2022;6:443.
22. Mullen E. Mullen scales of early learning. Circle Pine: American Guidance Service; 1995.
23. Sparrow S, Balla D, Cicchetti DV, Doll EA. Vineland-II scales of adaptive behavior. Circle Pines: American Guidance Service; 2005.
24. Sparrow SS, Cicchetti DV, Saulnier C. Vineland-3: Vineland adaptive behavior scales. San Antonio: Pearson; 2016.
25. Pierce K, Carter C, Weinfeld M, Desmond J, Hazin R, Bjork R, et al. Detecting, studying, and treating autism early: the one-year well-baby check-up approach. J Pediatrics. 2011;159(3):458–65.
26. Lord C, Risi S, Lambrecht L, Cook EH, Leventhal BL, DiLavore PC, 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. 2000;30:205–23.
27. Lange T, Roth V, Braun ML, Buhmann JM. Stability-based validation of clustering solutions. Neural Comput. 2004;16(6):1299–323.
28. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction [Internet]. arXiv; 2020 [cited 2022 Jun 13]. Available from: http://arxiv.org/abs/1802.03426
29. Liu Y, Hayes DN, Nobel A, Marron JS. Statistical significance of clustering for high-dimension, low-sample size data. J Am Stat Assoc. 2008;103(483):1281–93.
30. Chen CH, Gutierrez ED, Thompson W, Panizzon MS, Jernigan TL, Eyler LT, et al. Hierarchical genetic organization of human cortical surface area. Science. 2012;335(6076):1634–6.
31. Chen CH, Fiecas M, Gutiérrez ED, Panizzon MS, Eyler LT, Vuoksimaa E, et al. Genetic topography of brain morphology. Proc Natl Acad Sci USA. 2013;110(42):17089–94.
32. Genovese CR, Lazar NA, Nichols T. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. Neuroimage. 2002;15(4):870–8.
33. Burnham KP, Anderson DR. Multimodel inference: understanding AIC and BIC in model selection. Sociol Methods Res. 2004;33(2):261–304.
34. Krishnan A, Williams LJ, McIntosh AR, Abdi H. Partial least squares (PLS) methods for neuroimaging: a tutorial and review. Neuroimage. 2011;56(2):455–75.
35. Schroeder A, Mueller O, Stocker S, Salowsky R, Leiber M, Gassmann M, et al. The RIN: an RNA integrity number for assigning integrity values to RNA measurements. BMC Mol Biol. 2006;31(7):3.
36. Du P, Kibbe WA, Lin SM. lumi: a pipeline for processing Illumina microarray. Bioinformatics. 2008;24(13):1547–8.
37. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 2004;5(10):R80.
38. Pramparo T, Lombardo MV, Campbell K, Barnes CC, Marinero S, Solso S, et al. Cell cycle networks link gene expression dysregulation, mutation, and brain maldevelopment in autistic toddlers. Mol Syst Biol. 2015;11(12):841.
39. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinf. 2008;29(9):559.
40. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol. 2005. https://doi.org/10.2202/1544-6115.1128.
41. Oldham MC, Horvath S, Geschwind DH. Conservation and evolution of gene coexpression networks in human and chimpanzee brains. Proc Natl Acad Sci USA. 2006;103(47):17973–8.
42. Gandal MJ, Haney JR, Wamsley B, Yap CX, Parhami S, Emani PS, et al. Broad transcriptomic dysregulation occurs across the cerebral cortex in ASD. Nature. 2022;611(7936):532–9.
43. Catani M, Jones DK, Ffytche DH. Perisylvian language networks of the human brain. Ann Neurol. 2005;57(1):8–16.