Supplementary Note 1

Investigation of minor allele frequency cut-off for rare variants

We examined the burden of rare variants at different minor allele frequency (MAF) cut-offs, i.e., <1%, <0.1%, and <0.01%, using the discovery cohort with the gold standard examination. Only potentially high impact variants were focused on this analysis in order to select an optimal MAF cut-off for the downstream analyses. Thus, the burden analysis was done to test deletion and LoF variants impacting LoF intolerant genes with gnomAD upperbound of observation/expectation confidence intervals (o/e lof upper bound) <0.35 and also missense variants with missense badness, PolyPhen-2, and constraint (MPC) score>2. Similar to the main burden analysis, a logistic regression model was applied to model two morphological subtypes based on the gold standard examination with sex, platform and first 3 principal components as covariates. No significant results were reached by LoF and missense variants, possibly due to their limited statistical power, while the burden analysis of deletions show that using <1% as a MAF cut-off give the best p-value followed by <0.1% and <0.01% (Supplementary Figure 2). Therefore, for the subsequent analysis, the rare variants were defined as those with MAF <1%.

Population stratification and GRVS in European and non-European subsets.

We investigated the effectiveness of GRVS stratified by population, i.e., European subset and non-European subset. First, we examined whether there was population bias between the two subtypes in the discovery cohort. Logistic regression was performed to model morphological subtypes (based on gold standard examination) using different principal components (PC1-PC4). We found no association between the subtypes and any PCs (P>0.05). Visualizing different PCs also shows that samples are clustered together and well-mixed, except a few cases of those with non-European ancestry (n=8 out of 325 probands in the discovery cohort, Supplementary Figure 7a). In the discovery cohort, we reran GRVS analysis using only European subset. Due to the fact that majority of the samples are of European ancestry, the significant difference in GRVS between the two subtypes were retained, P=0.048 for gold standard examination (Supplementary Figure 7b) and $P=2.3\times10^{-5}$ for ADM (Supplementary Figure 7c). In addition, the GRVS in the replication cohort were calculated and compared separately for European and non-European subsets. We found a consistent result where GRVS is higher in ADM dysmorphic individuals than unaffected siblings ($P=8.2\times10^{-4}$) and ADM nondysmorphic individuals ($P=1.3\times10^{-3}$) in the replication cohort when limiting to those of European ancestry (Supplementary Figure 7d). Moreover, GRVS of non-European samples were compared, which also yielded a significant result ($P=1.5\times10^{-4}$ and $P=3.1\times10^{-4}$, for ADM dysmorphic vs unaffected siblings and ADM dysmorphic vs ADM nondysmorphic, respectively) despite being a much smaller subset (n=142).

ASD candidate variants:

We also identified 29 variants of unknown significance of interest in 26 probands that fell into three categories: 1) variants in known ASD/neurodevelopmental genes, but with unknown impact on gene function or disease, 2) variants in ASD/neurodevelopmental candidate genes with emerging evidence, or 3) tandem repeat expansions in previously reported ASD candidate loci[1]. Additional information regarding ASD candidate variants in category 1 are described below:

Paternally inherited loss-of-function (LoF) variant in *CIC* in subject 3-0328-000
Cic$^{+/-}$ mice exhibit mild hyperactivity compared to wild-type mice. Mice with conditional knockouts of *Cic* in forebrain show memory deficits, hyperactivity, altered cortical thickness, defects in neuronal maturation and maintenance, and altered dendritic branching. Conditional knockouts of *Cic* in hypothalamus and amygdala result in abnormal social interaction[2]. Five unrelated individuals with *de novo* LoF variants in *CIC* share similar clinical features, including intellectual disability, developmental delay, ASD, attention deficit and hyperactivity disorder, seizures, and brain abnormalities[2]. *De novo* LoF variants reported in Lu *et al*.[2] impact both *CIC* isoforms, whereas the paternally inherited LoF in our subject 3-0328-000 only affects the short isoform (*CIC-S*). The impact of this variant on *CIC* is unknown and there are no other reports of cases with only *CIC-S* impacted. However, the short isoform is expressed in mouse brain[3], The proband has complex ASD and intellectual disability; his phenotype is further described in Supplementary Data 7. The proband's father has no neuropsychiatric phenotype; he has post-secondary education and is employed as a nurse.

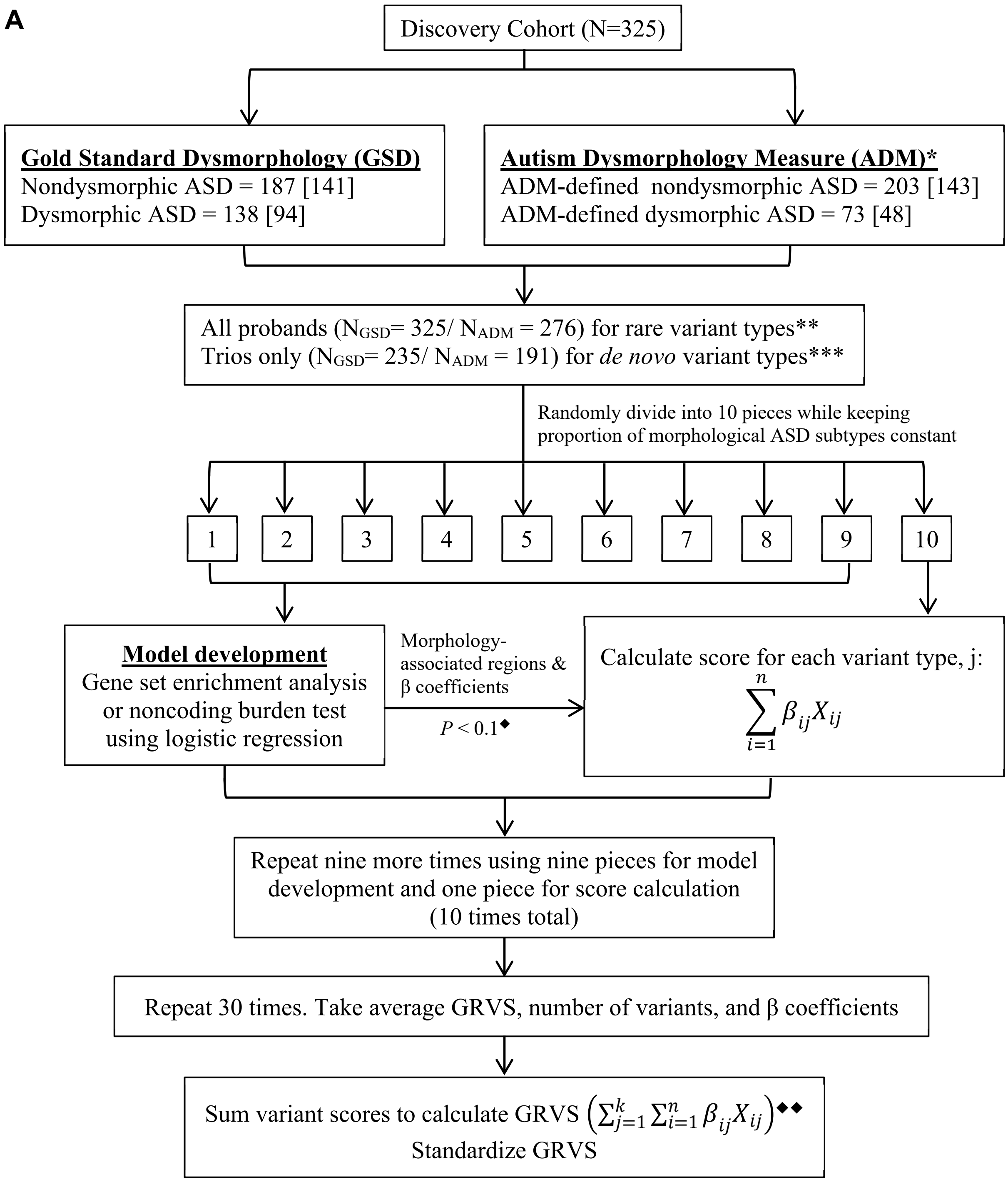
Paternal uniparental isodisomy involving homozygous missense variant in *FAT4* in subject 3-0095-000
A missense variant in *FAT4*, predicted to be damaging, was identified to be homozygous as a result of uniparental (paternal) isodisomy of chromosome 4. Homozygous LoF and missense variants of *FAT4* are associated with Van Maldergen Syndrome, which is characterized by intellectual disability, partially penetrant periventricular neuronal heterotopia, and craniofacial, skeletal, auditory and renal malformations[4]. Our subject has some features of Van Maldergem syndrome (severe infantile hypotonia, delayed closure of the anterior fontanel, hypospadias and a cerebellar abnormality). He also has a pathogenic *de novo* LoF variant in *WAC*: a gene associated with Desanto-Shinawi syndrome, of which ASD is a main feature[5]. Given the phenotypes associated with these two syndromes, we suggest that the *de novo WAC* variant is a main contributor to his ASD phenotype, although we cannot rule out the possible additional impact of the homozygous missense *FAT4* variant. The proband has complex ASD; his phenotype is further described in Supplementary Data 7.

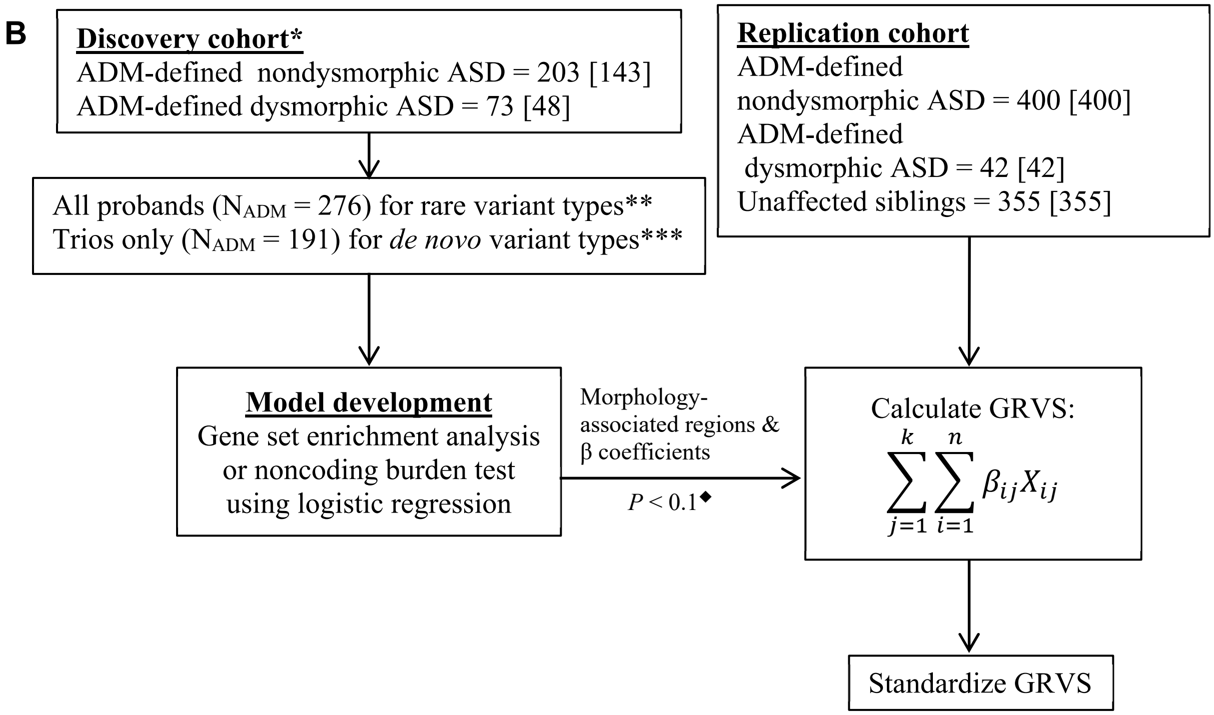
Paternally inherited "templated sequence insertion" involving *PHF21A* and *GLIS2* in subject 3-0439-000
PHF21A is within the chr11p11.2 deleted region associated with Potocki-Shaffer syndrome[6]. *De novo* LoF variants of this gene are associated with intellectual disability and craniofacial abnormalities, with ASD reported in 3 of the 10 reported cases[6-8]. Templated sequence insertions (TSIs) are characterized by reverse transcription of an RNA intermediate, LINE-1-based insertion, target site duplication, cryptic polyadenylation signal, and polyadenylation[9]. In subject 3-0439-000, the last intron and exon of *GLIS2* are inserted in an inverted manner into *PHF21A*, along with a polyadenylation sequence and microduplication of 17bp (Supplementary Figure 8). The impact of this TSI on *PHF21A*

expression and function is unknown. The proband has equivocal ASD (see Supplementary Data 7). His father has two years of post-secondary education. Further clinical information was not available.

Inherited “templated sequence insertion” involving *FGFR2* and *NPM1* in subjects 3-0209-000 and 3-0728-000

FGFR2 is associated with several mutation-specific disorders (OMIM: 176943). Activating *FGRF2* variants are associated with several distinct craniosynostosis syndromes, some of which are associated with neurodevelopmental abnormalities. Subject 3-0209-000 has a paternally inherited TSI of *NPM1* cDNA into *FGFR2*. The insertion is inverted and is followed by a polyadenylation insertion and a microduplication. We found the same insertion as a maternally inherited TSI in subject 3-0728-000 (Supplementary Figure 9). It is unknown whether this variant affects the coding sequence of *FGFR2*, and whether this variant will be associated with a known disorder or a different disorder. Both probands have high functioning forms of ASD and neither has craniosynostosis. Subject 3-0209-000 has essential ASD. His father is employed as a welder and has no neuropsychiatric phenotype. 3-0728-000 has complex ASD, attention deficit disorder and an anxiety disorder. His mother is employed at a call centre and has mental health issues including an anxiety disorder. (See Supplementary Data 7 for additional phenotypic information on the probands).


A
Discovery Cohort (N=325)
Gold Standard Dysmorphology (GSD)
Nondysmorphic ASD = 187 [141]
Dysmorphic ASD = 138 [94]
Autism Dysmorphology Measure (ADM)*
ADM-defined nondysmorphic ASD = 203 [143]
ADM-defined dysmorphic ASD = 73 [48]
All probands (N_{GSD}= 325/ N_{ADM} = 276) for rare variant types**
Trios only (N_{GSD}= 235/ N_{ADM} = 191) for *de novo* variant types***
Randomly divide into 10 pieces while keeping proportion of morphological ASD subtypes constant
1
2
3
4
5
6
7
8
9
10
Model development
Gene set enrichment analysis or noncoding burden test using logistic regression
Morphology-associated regions & β coefficients
$P < 0.1$◆
Calculate score for each variant type, j:
$\sum_{i=1}^{n} \beta_{ij} X_{ij}$
Repeat nine more times using nine pieces for model development and one piece for score calculation (10 times total)
Repeat 30 times. Take average GRVS, number of variants, and β coefficients
Sum variant scores to calculate GRVS $\left(\sum_{j=1}^{k} \sum_{i=1}^{n} \beta_{ij} X_{ij}\right)$◆◆
Standardize GRVS


B
Discovery cohort*
ADM-defined nondysmorphic ASD = 203 [143]
ADM-defined dysmorphic ASD = 73 [48]
All probands (N_{ADM} = 276) for rare variant types**
Trios only (N_{ADM} = 191) for *de novo* variant types***
Replication cohort
ADM-defined nondysmorphic ASD = 400 [400]
ADM-defined dysmorphic ASD = 42 [42]
Unaffected siblings = 355 [355]
Model development
Gene set enrichment analysis or noncoding burden test using logistic regression
Morphology-associated regions & β coefficients
$P < 0.1$◆
Calculate GRVS:
$\sum_{j=1}^{k}\sum_{i=1}^{n}\beta_{ij}X_{ij}$
Standardize GRVS

Supplementary Figure 1: Flowchart for GRVS calculation for discovery (A) and replication cohorts (B). Square brackets indicate the number of trios in each subtype.

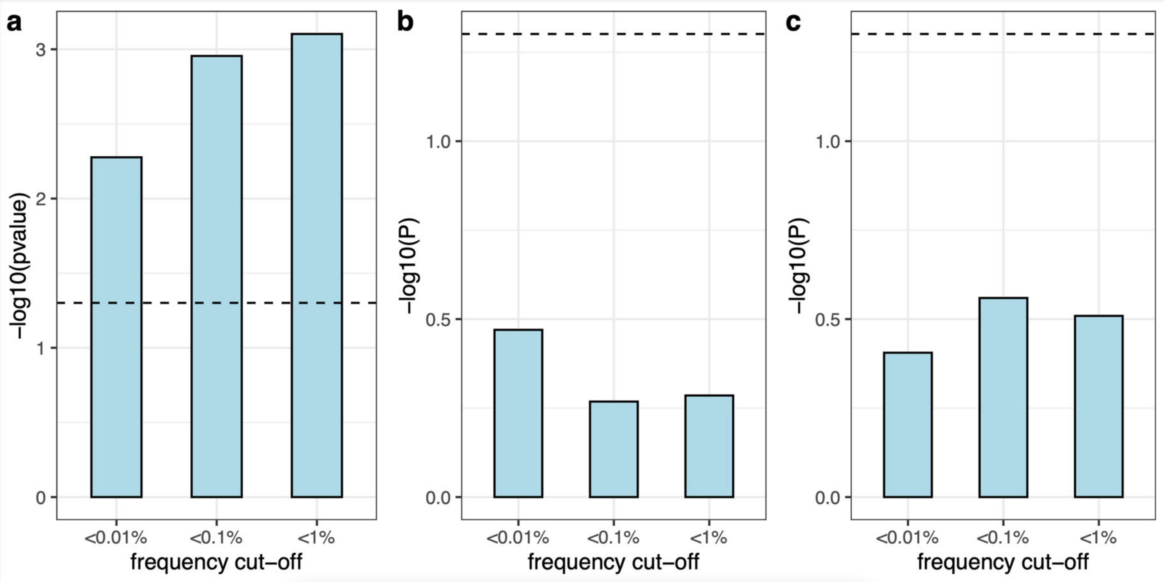
*Discovery cohort classified using ADM does not include false negative samples (i.e. complex ASD cases classified as ADM-defined nondysmorphic ASD), and only includes cases sequenced by Illumina.

**rare variant types that were analysed consisted of coding deletions >10kb, coding deletions ≤ 10kb, coding duplications >10kb, coding duplications ≤ 10kb, predicted loss-of-function variants, missense variants, predicted damaging missense variants, noncoding deletions >10kb, noncoding deletions ≤ 10kb, noncoding duplications >10kb, noncoding duplications ≤ 10kb, and noncoding SNVs and indels.

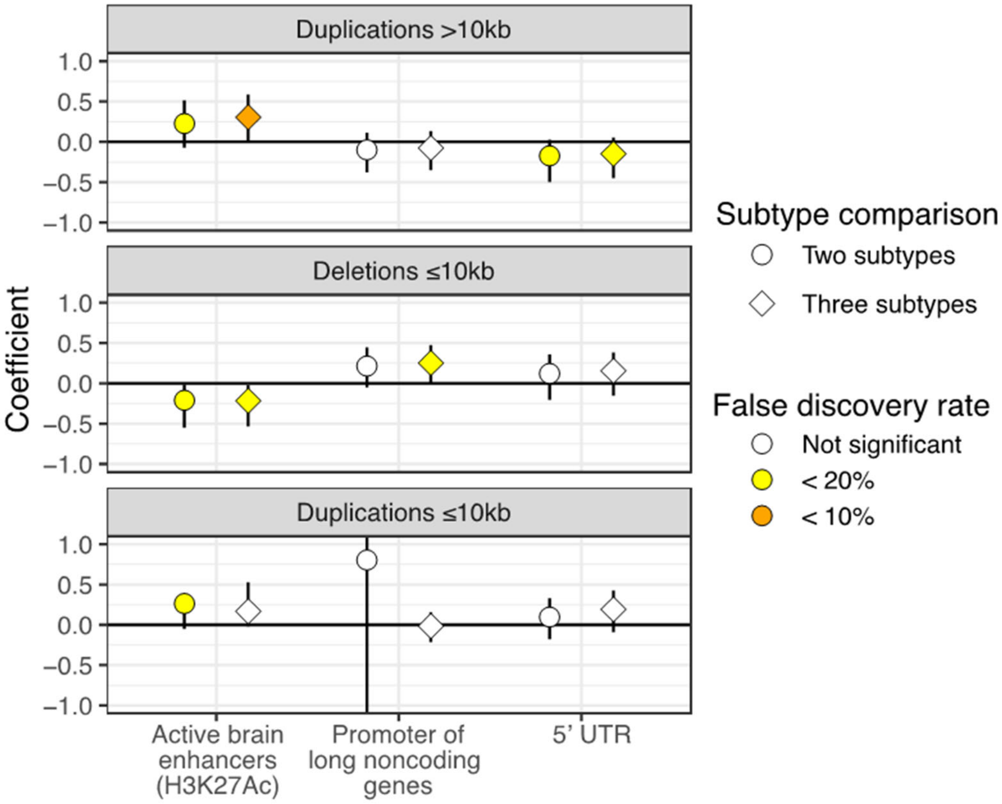
****de novo* variant types that were analysed consisted of predicted loss-of-function variants, missense variants, predicted damaging missense variants, and noncoding SNVs and indels.

◆Nagelkerke's R^2 was calculated at different *P* value thresholds (*P* < 1, 0.5, 0.1, 0.01, 0.005, and 0.001) to determine the optimal *P* value threshold.

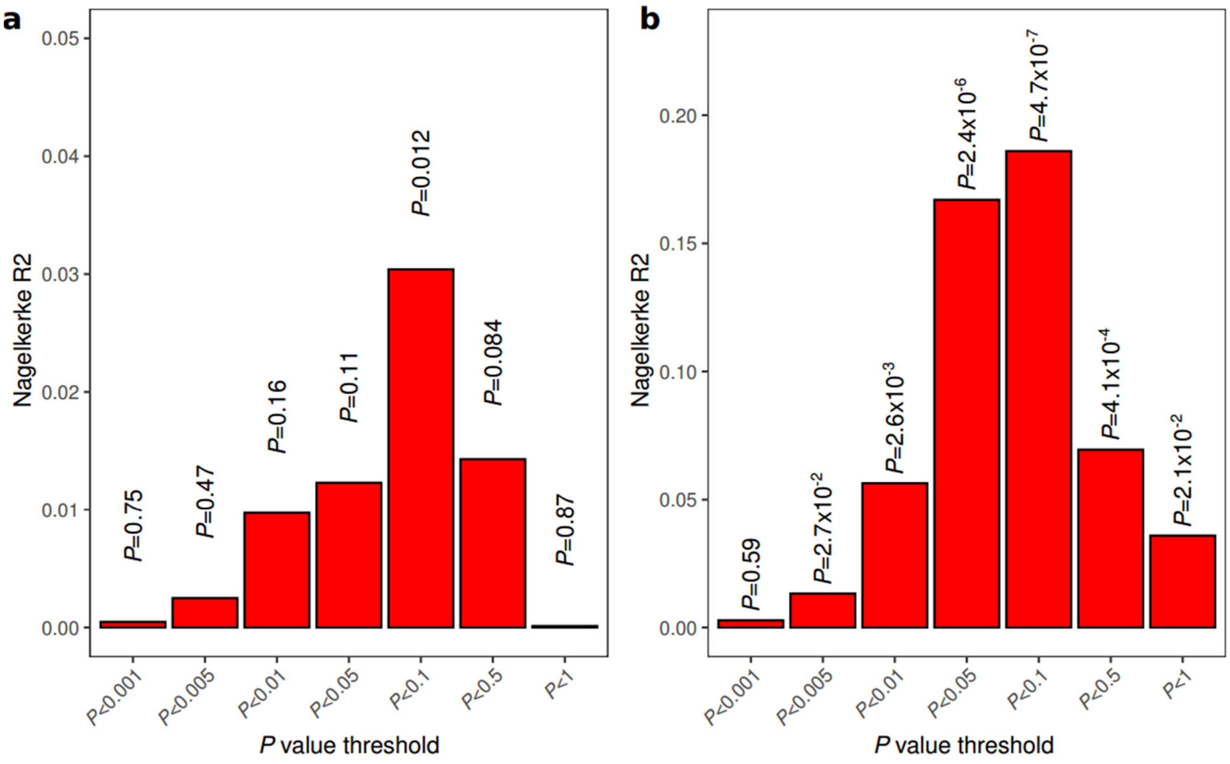
◆◆GRVS was calculated only probands with both parents sequenced because they had variant scores for both rare and *de novo* variants.



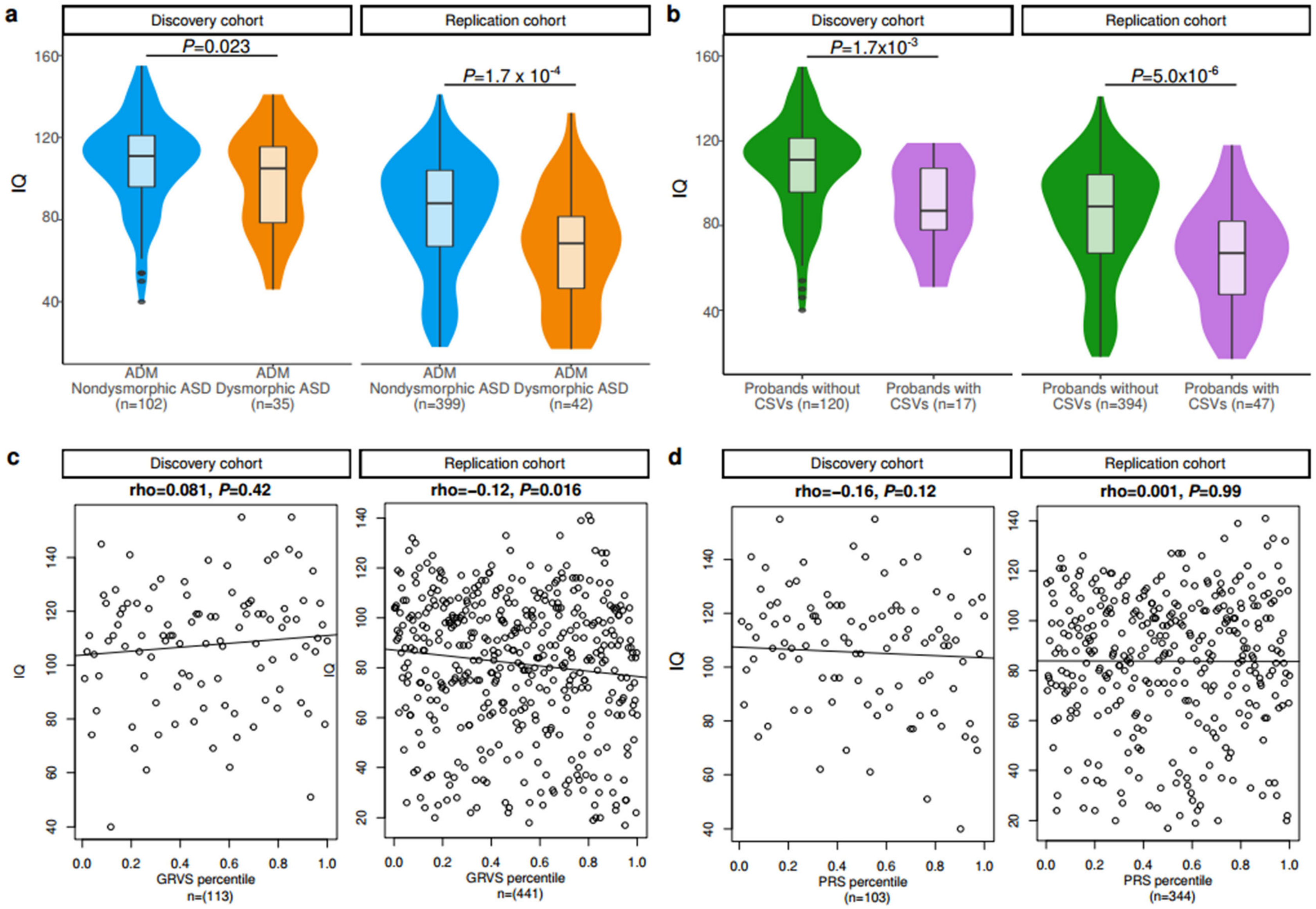
Supplementary Figure 2: Investigation of minimum allele frequency cut-off for rare variants.
The burden of rare variants was tested using different minor allele frequency cut-offs for **a)** deletion impacting LoF intolerant genes, **b)** LoF variants impacting LoF intolerant genes and c) missense variants with MPC score > 2. Bars indicate –log10 P values of the test and dotted lines indicate P=0.05 for a significant enrichment threshold.



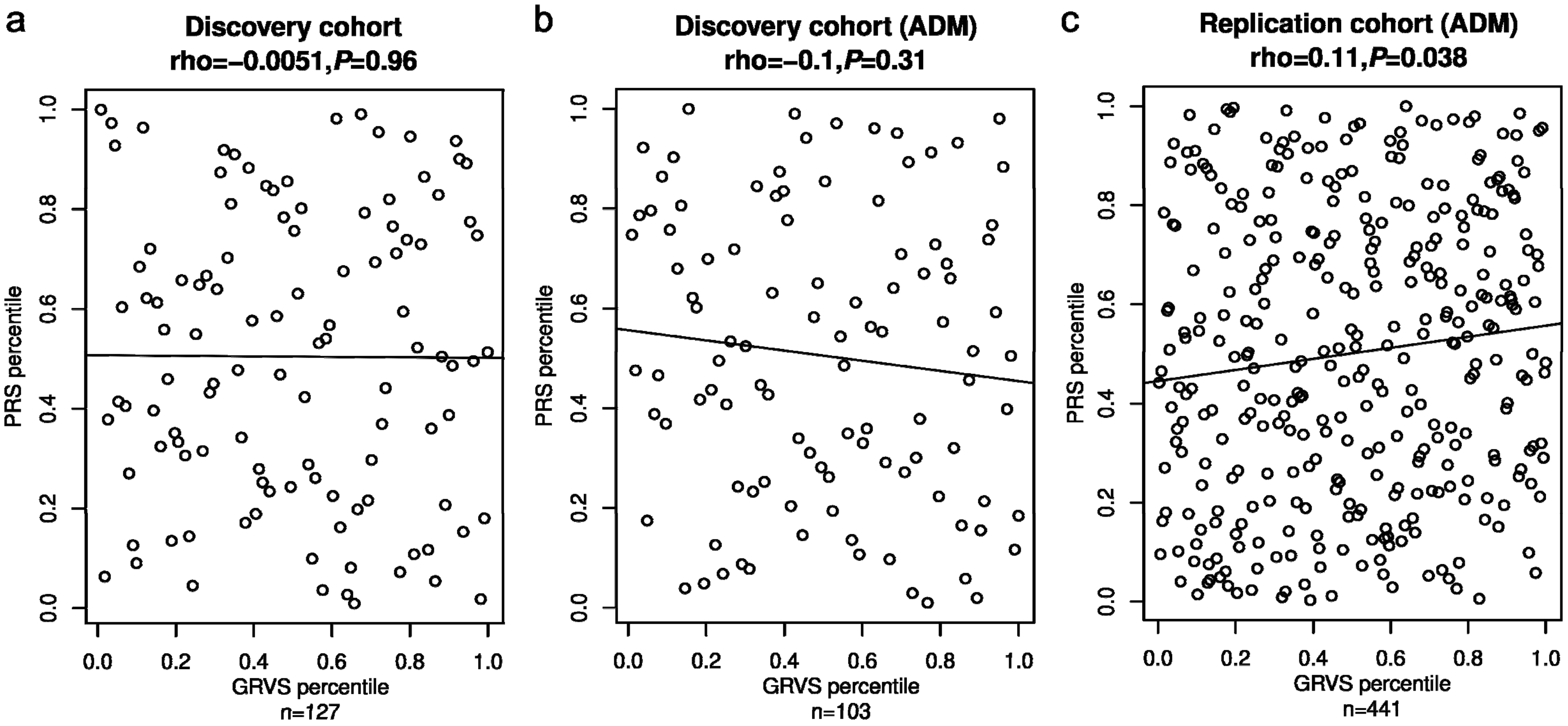
Supplementary Figure 3: Noncoding regions for which rare variants are significantly more prevalent in some subtypes of ASD.
Events and coefficients are as described in Figure 3. We show only noncoding regions for which duplications >10kb (top panel), deletions ≤10kb (middle panel), and duplication ≤10kb (bottom panel) are significantly more prevalent in different subtypes of ASD (n=325 samples). Symbol shapes indicate the subtype comparisons that were conducted for each combination of gene set and variant type. Two subtype comparison = nondysmorphic vs. dysmorphic ASD. Three subtype comparison = essential vs. equivocal vs. complex ASD. Coloured shapes indicate significant signals after multiple test correction by permutation-based FDR, where yellow, orange, and red, indicate permutation-based FDR < 20%, 10% and 5%, respectively. The data points (the centre) indicate estimated coefficient, while error bars indicate 95% confidence intervals of the estimated coefficient.



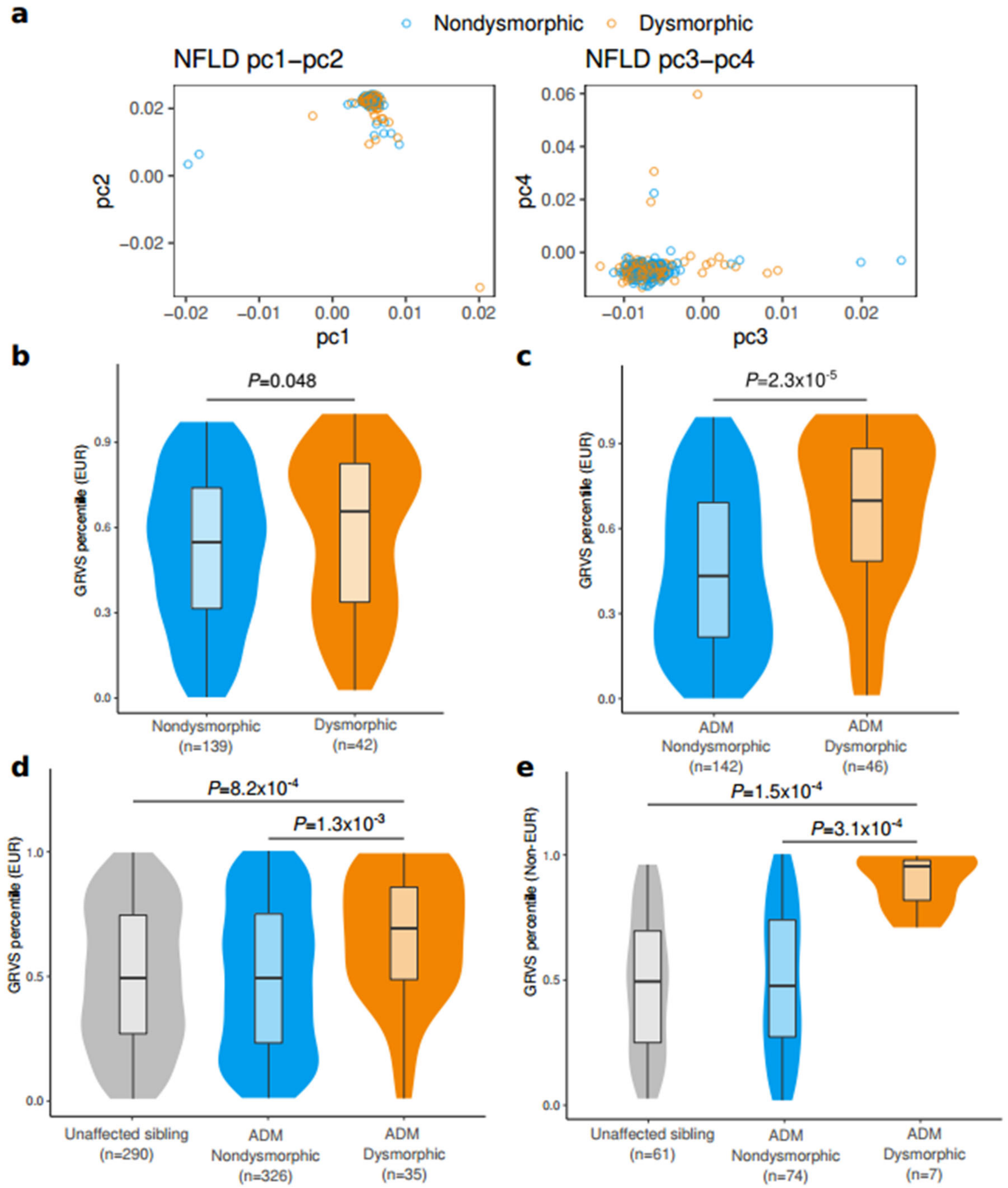
Supplementary Figure 4: Nagelkerke's R^2 to determine optimal *P* value threshold for GRVS. Shown is the distribution of Nagelkerke R^2 of GRVS in the discovery cohort using 10 × 30-fold cross validation on a) gold standard dysmorphology examinations, or b) Autism Dysmorphology Measure (ADM) at different *P* value thresholds, which were used to identify morphology-associated gene sets and noncoding regions for GRVS calculation. Based on both methods, the optimal *P* value threshold is *P*<0.1.



Supplementary Figure 5: Relationship between IQ, genetic variants and morphological ASD subtypes classified by the Autism Dysmorphology Measure (ADM). The left panels show results for the discovery cohort that was classified using the Autism Dysmorphology Measure, while the right panels show results for the replication cohort**.** IQ comparison between **a**) ASD subtypes classified by ADM, or **b**) probands with or without clinically significant variants (CSVs). Violin plots show the distribution of the samples' IQ; box plots contained within show the median and quartiles of IQ for each subtype, and the minima and maxima of box plots indicate 3× the interquartile range-deviated IQ from the median. P values denote the probability that the mean IQ of ADM-defined nondysmorphic ASD or probands without CSVs is not greater than ADM-defined dysmorphic ASD or probands with CSVs, respectively (one-sided, t-test). Correlation between IQ and **c**) GRVS or **d**) PRS percentiles. Each dot represents the PRS and GRVS percentile for a sample in the discovery cohort or replication cohort. The linear regression line indicates the linear correlation between IQ and GRVS or PRS percentiles. Correlation coefficient is quantified by two-sided Spearman's rho correlation. *P* values indicate the probability that the correlation is occurred due to chance.

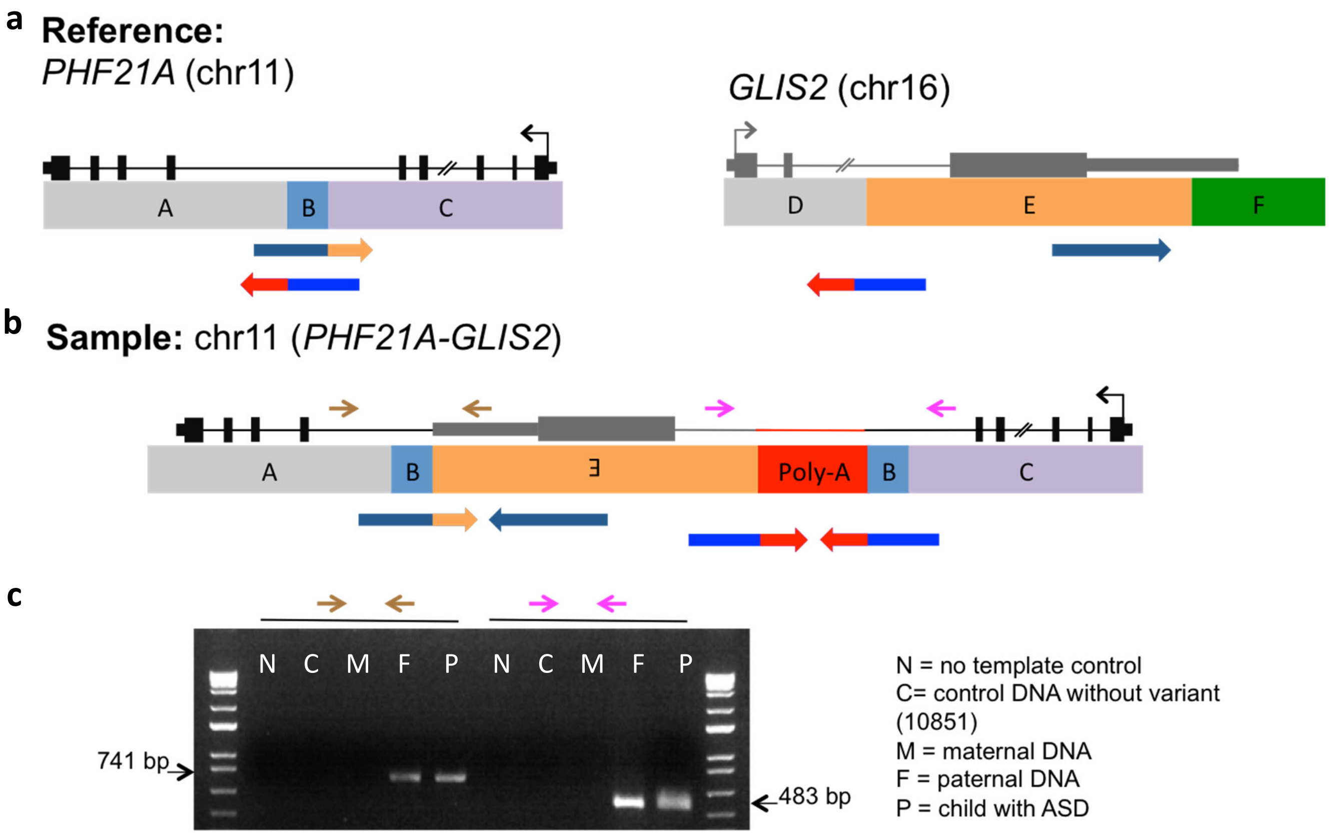


Supplementary Figure 6: Correlation between GRVS and PRS. Each dot represents the PRS and GRVS percentile for a sample in the discovery cohort using a) gold standard dysmorphology examinations, or b) the Autism Dysmorphology Measure (ADM), or c) in the replication cohort. The linear regression line indicates the linear correlation between GRVS and PRS percentiles. Correlation coefficient is quantified by Spearman's rho correlation. *P* values indicate the probability that the correlation is occurred due to chance.



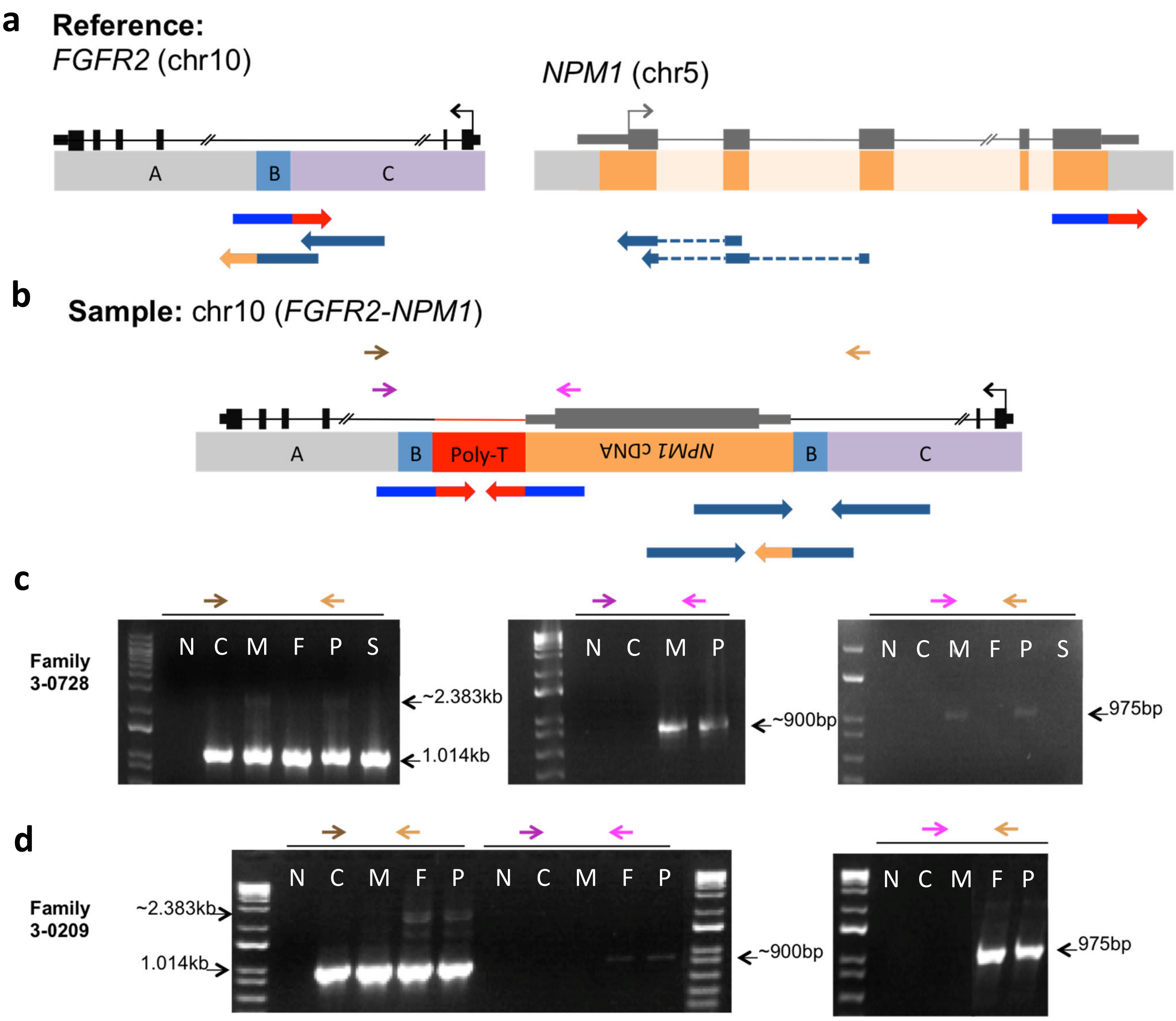
Supplementary Figure 7: Population (EUR and non-EUR) and GRVS analysis.
a) Scatter plot of principle components (PC1 vs PC2 and PC3 vs PC4) of population stratification by common variants (n=325 samples). Data points are samples colour coded by morphological subtypes (blue=nondysmorphic, orange=dysmorphic). In the discovery cohort, GRVS of samples of European ancestry were compared between two morphological subtypes classified by **b)** gold standard dysmorphology examinations, and **c)** Autism Dysmorphology Measure (ADM). In the replication cohort, unaffected siblings and two morphological subtypes were compared in **d)** European subset and **e)** non-European

subset. Box plots show the median and quartiles of IQ for each subtype, and the minima and maxima of box plots indicate 3× the interquartile range-deviated scores from the median. *P* values were calculated using one-sided Wilcoxon ranked-sum test assuming higher GRVS in dysmorphic subtype.



Supplementary Figure 8: Illustration of paternally inherited templated sequence insertion of last intron and exon of *GLIS2* into intron 14 of *PHF21A* in subject 3-0439-000.

a) Alignment of WGS reads (blue arrows) to *PHF21A* and *GLIS2* reference sequence (black and grey genes, respectively, and blocks A-F). Mate pairs are depicted by the same hue of blue. Split reads are depicted by the red and orange-coloured section on the blue reads that align to poly-A and section E, respectively. **b**) In the sample's genomic sequence, section E was duplicated and inserted into intron 1 of *PHF21A* along with a non-reference poly-A insertion (red block and line) and microduplication of section B. As a result, most of the last intron and exon of *GLIS2* (grey) were inserted into *PHF21A* (black genes). The sample's genes and genomic sequence is shown through black, grey and/or red lines and boxes, blocks A-C, E, and a poly-A block. Brown and magenta arrows depict the location of primers for PCR validation. **c**) PCR validation of variant allele in family 3-0439 at 5' and 3' end (brown and magenta arrows, respectively). Source data are provided as a Source Data file.



N= no template control
C= control DNA without variant (10851)
M = maternal DNA
F = paternal DNA

P = child with ASD
S = unaffected sibling

Supplementary Figure 9: Illustration of insertion of *NPM1* cDNA into *FGFR2* in two subjects.

a) Alignment of WGS reads (blue arrows) to *FGFR2* and *NPM1* reference sequence (black and grey genes, respectively). Mate pairs are depicted by the same hue of blue. Split reads are depicted by the red and orange-coloured section on the blue reads that align to poly-A and *NPM1* cDNA, respectively. Dotted lines indicate that the WGS read perfectly aligned to *NPM1* exons. **b**) In the sample's genomic sequence, most of *NPM1* cDNA (grey) was inserted into *FGFR2* intron (black) in an inverted manner along with a non-reference poly-A insertion (red block and line) and microduplication of section B. The sample's genes and genomic sequence is shown with black, grey, and red lines and boxes, blocks A-C, *NMP1* cDNA, and a poly-A block. Brown and magenta arrows depict the location of primers for

PCR validation. PCR validation of variant allele in families **c**) 3-0728 and d) 3-0209. Source data are provided as a Source Data file. Brown primer pairs were used to amplify DNA of each sample. Purple and magenta primer pairs were used to conduct nested PCR on the 5' end of PCR products of brown primer pairs. Magenta and tan primer pairs were used to conduct nested PCR the 3' end of PCR products of brown primer pairs.

References

1. Trost, B. *et al.* Genome-wide detection of tandem DNA repeats that are expanded in autism. *Nature* (2020).
2. Lu, H.C. *et al.* Disruption of the ATXN1-CIC complex causes a spectrum of neurobehavioral phenotypes in mice and humans. *Nat Genet* **49**, 527-536 (2017).
3. Lam, Y.C. *et al.* ATAXIN-1 interacts with the repressor Capicua in its native complex to cause SCA1 neuropathology. *Cell* **127**, 1335-47 (2006).
4. Cappello, S. *et al.* Mutations in genes encoding the cadherin receptor-ligand pair DCHS1 and FAT4 disrupt cerebral cortical development. *Nat Genet* **45**, 1300-8 (2013).
5. Varvagiannis, K., de Vries, B.B.A. & Vissers, L. WAC-Related Intellectual Disability. in *GeneReviews((R))* (eds. Adam, M.P. *et al.*) (Seattle (WA), 1993).
6. Kim, H.G. *et al.* Translocations disrupting PHF21A in the Potocki-Shaffer-syndrome region are associated with intellectual disability and craniofacial anomalies. *Am J Hum Genet* **91**, 56-72 (2012).
7. Kim, H.G. *et al.* Disruption of PHF21A causes syndromic intellectual disability with craniofacial anomalies, epilepsy, hypotonia, and neurobehavioral problems including autism. *Mol Autism* **10**, 35 (2019).
8. Hamanaka, K. *et al.* De novo truncating variants in PHF21A cause intellectual disability and craniofacial anomalies. *Eur J Hum Genet* **27**, 378-383 (2019).
9. Onozawa, M., Goldberg, L. & Aplan, P.D. Landscape of insertion polymorphisms in the human genome. *Genome Biol Evol* **7**, 960-8 (2015).