The Contributions of Rare Inherited and Polygenic Risk to ASD in Multiplex Families

Timothy S Chang¹,¹⁰, Matilde Cirnigliaro²,¹⁰, Stephanie A Arteaga³,¹⁰, Laura Pérez-Cano⁴, Elizabeth K Ruzzo²,⁵, Aaron Gordon³, Lucy Bicks³, Jae-Yoon Jung⁶,⁷, Jennifer K Lowe²,⁵, Dennis P Wall⁶,⁷, Daniel H Geschwind¹,²,⁴,⁵,⁸,⁹*

¹ Movement Disorders Program, Department of Neurology, David Geffen School of Medicine, University of California, Los Angeles, CA, USA
² Department of Psychiatry and Biobehavioral Sciences, Semel Institute, David Geffen School of Medicine, UCLA, Los Angeles, CA, USA
³ Program in Neurogenetics, Department of Neurology, David Geffen School of Medicine, University of California, Los Angeles, CA, USA
⁴ Department of Neurology, David Geffen School of Medicine, UCLA, Los Angeles, CA, USA
⁵ Center for Autism Research and Treatment, Semel Institute, David Geffen School of Medicine, UCLA, Los Angeles, CA, USA
⁶ Department of Pediatrics, Division of Systems Medicine, Stanford University, Stanford, CA, USA
⁷ Department of Biomedical Data Science, Stanford University, Stanford, CA, USA
⁸ Department of Human Genetics, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA, USA
⁹ Lead Contact
¹⁰ These authors contributed equally

* Correspondence: Daniel H Geschwind, dhg@mednet.ucla.edu

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.
Abstract

Autism Spectrum Disorder (ASD) has a complex genetic architecture involving contributions from *de novo* and inherited variation. Few studies have been designed to address the role of rare inherited variation, or its interaction with polygenic risk in ASD. Here, we performed whole genome sequencing of the largest cohort of multiplex families to date, consisting of 4,551 individuals in 1,004 families having 2 or more affected children with ASD. Using this study design, we identify seven novel risk genes supported primarily by rare inherited variation, finding support for a total of 74 genes in our cohort and a total of 152 genes after combining with other studies. Probands demonstrated an increased burden of mutations in 2 or more known risk genes (KARGs) — in three families both probands inherited protein truncating variants in two KARGs. We also find that polygenic risk is over transmitted from unaffected parents to affected children with rare inherited variants, consistent with combinatorial effects in the offspring, which may explain the reduced penetrance of these rare variants in parents. We also observe that in addition to social dysfunction, language delay is associated with ASD polygenic risk over-transmission. These results are consistent with an additive complex genetic risk architecture of ASD involving rare and common variation and further suggest that language delay is a core biological feature of ASD.
Introduction

The last decade has seen enormous progress in identifying genes imparting risk for autism spectrum disorder (ASD) (O’Roak et al. 2012; Iossifov et al. 2014; O’Roak et al. 2014; De Rubeis et al. 2014; Krumm et al. 2015; Sanders et al. 2015; Turner et al. 2016; Stessman et al. 2017; Ruzzo et al. 2019; Satterstrom et al. 2020), identifying over 100 genes whose risk is due primarily to rare de novo copy number variants or protein truncating variants (Yuen et al. 2017; Werling et al. 2018; Feliciano et al. 2019; Satterstrom et al. 2020; Sanders et al. 2015; Turner et al. 2017; Dong et al. 2014; Iossifov et al. 2014; O’Roak et al. 2012; Neale et al. 2012; Sebat et al. 2007; Marshall et al. 2008; Bucan et al. 2009). The majority of these risk genes appear to act pleiotropically, contributing to a broad range of neurodevelopmental disorders in addition to ASD, ranging from epileptic encephalopathy, attention-deficit/hyperactivity disorder, intellectual disability, and schizophrenia in some cases (Epi4K Consortium et al. 2013; Satterstrom et al. 2019; Stessman et al. 2017; Satterstrom et al. 2020; Singh et al. 2017; Cross-Disorder Group of the Psychiatric Genomics Consortium. 2019). These findings have delivered on the original promise of genetic investigations by galvanizing functional genomic and neurobiological studies that have begun to yield mechanistic insight into ASD pathophysiology (de la Torre-Ubieta et al. 2016; Geschwind and State 2015; Abrahams and Geschwind 2008).

However, ASD is also highly heritable (Gaugler et al. 2014; Colvert et al. 2015; Sandin et al. 2017; Bai et al. 2019) and therefore is expected to have a substantial contribution from common (Gaugler et al. 2014; Weiner et al. 2017) and rare variation transmitted from parents to their affected offspring. Indeed, at least 50% of genetic risk is predicted to be due to common variation, 15-20% is due to de novo variation and other Mendelian forms, and the remaining genetic risk is yet to be determined (Grove et al. 2019; Gaugler et al. 2014; Iossifov et al. 2014; De Rubeis et al. 2014; Satterstrom et al. 2020). Even with substantial sized cohorts (greater than
25,000), genome wide association studies (GWAS) in ASD have to date only identified five loci reaching genome wide significance (Grove et al. 2019). This is consistent with high genetic heterogeneity, as well as contribution from other rare forms of inherited variation.

Another potential contribution to the preponderance of genes supported by de novo variation, and less success in detecting inherited variation, may be study designs with a heavy focus on families with simplex ASD, having only a single proband and no family history, so that probands are depleted for inherited genetic risk for ASD (Ronemus et al. 2014; Leppa et al. 2016; Yuen et al. 2015; Feliciano et al. 2019). Indeed, our recent study of families having 2 or more affected children, demonstrated a different genetic architecture in these multiplex ASD families (Lajonchere and AGRE Consortium 2010; Ruzzo et al. 2019) with a significant signal from rare, inherited variants and a less significant contribution from de novo variation (Ruzzo et al. 2019). Studying multiplex families allowed us to identify 16 new ASD genes supported predominantly by inherited rare variants in addition to de novo variants (Ruzzo et al. 2019).

In this current study, we performed whole genome sequencing on the largest collection of ASD multiplex families to date, including 4,551 individuals, comprised of 1,836 probands and 418 unaffected siblings from 1,004 families to further explore the contribution of rare and common inherited variation to ASD genetic risk architecture. We identified 74 ASD risk genes harboring rare protein disrupting or damaging missense variants, seven of which were not identified in previous studies and thus are novel risk genes. We also leverage WGS to assess the contribution of polygenic risk (PGR) in those with inherited and de novo variants to further frame the genetic architecture of ASD. We show that while PGR is not over-transmitted to ASD probands harboring de novo variants, PGR is significantly over-transmitted to those harboring inherited protein coding rare variants and those with a more severe ASD phenotype. These data
provide new evidence supporting a complex view of ASD genetic architecture that includes a mix of common and rare genetic variation, whereby PGR contributes to ASD risk in combination with rare inherited protein truncating variants. The contribution of PGR to autism risk may also explain why parents harboring rare protein disrupting variants are asymptomatic, since they lack the polygenic contribution from common variants that is co-transmitted along with rare risk variants to their affected children.
Results

We performed whole genome sequencing at a mean >30x coverage in 4,551 individuals from 1,004 multiplex ASD families (Table S1). Study individuals were from the Hartwell Autism Research and Technology Initiative (iHART) cohort within the Autism Genetic Resource Exchange (AGRE) (Lajonchere and AGRE Consortium 2010), and included an initial smaller published subset of this cohort (493 multiplex ASD families, (Ruzzo et al. 2019)). This analysis of multiplex ASD families included 1,836 affected children and 418 unaffected children where both biological parents were sequenced (fully-phaseable) (Figure 1) and is more than double in size compared to the previous study (Ruzzo et al. 2019). Cases were diagnosed according to widely used protocols that have been well validated (Lord, Rutter, and Le Couteur 1994; Lord et al. 2000), and details of clinical ascertainment in AGRE have been previously published (Leppa et al. 2016; Geschwind et al. 2001).

Of the 1,004 multiplex ASD families, 714 families have 2 fully-phaseable affected children (71.1%), 109 families have 3 fully-phaseable affected children (10.8%), 8 families have 4 fully-phaseable affected children (0.8%), 4 families have 5 fully-phaseable affected children (0.4%), and 368 families have at least 1 fully-phaseable unaffected child (36.7%). Among the fully-phaseable children, 1620 are males and 634 are females. We also utilized 76 families with monozygotic concordant pairs to inform quality control (Ruzzo et al. 2019).

Overall genetic architecture and transmission patterns in multiplex families

Whole genome sequencing yielded an average coverage of 37.48X with 80.6% of bases covered at ≥30X and 99.2% of bases covered at ≥10X (Figures S1A-S1E). Quality metrics were within the expected range for whole genome sequencing (WGS) (J. Wang et al. 2015) including
a transition to transversion ratio of 2.155, a heterozygous to non-reference homozygous ratio of 1.6, and dbSNP concordant rate of 99.92%. Our WGS processing pipeline retained high quality variants, for example by removing variant calls for samples with a genotype quality (GQ)<25, depth (DP)<10 and alternate allele depth to depth (ALT AD/DP)<0.2 (Methods, Figure S2). In addition, we applied a validated machine learning approach based on random forests, Artifact Removal by Classifier (ARC) (Ruzzo et al. 2019), to distinguish true rare de novo variants (RDNVs) from likely sequencing, mapping errors or artifacts from lymphoblastoid culture (Methods) (Conrad et al. 2011; 1000 Genomes Project Consortium et al. 2015).

After ARC, we find an overall rate of 53 rare de novo variants per subject, with no differences between cases and controls (Methods, Wilcoxon rank sum test p=0.20), consistent with published literature (Besenbacher et al., 2016; Conrad et al., 2011; Kong et al., 2012; Michaelson et al., 2012; Turner et al., 2016; Yuen et al., 2017). As previously, we find a robust association of paternal age and de novo association rate (Figure S3, Methods, 1.02 RDNVs per year of paternal age, p<2.2 x 10^{-16}) (Deciphering Developmental Disorders Study 2017; Francioli et al. 2015; Goldmann et al. 2016; Michaelson et al. 2012). The mean number of rare inherited variants (MAF<0.001, see Methods for additional quality criteria), private inherited variants, and rare de novo variants (protein truncating, missense or synonymous) per subject was 145.5, 32.4, and 0.46 respectively. Consistent with the hypothesized and previously observed depletion of rare de novo ASD risk in multiplex families (Leppa et al. 2016; Ruzzo et al. 2019), we observed no enrichment of rare de novo PTVs (logistic regression, p = 0.79) and missense variants (logistic regression, p = 0.46) in affected children compared to unaffected children (Figure 2A). We obtained similar results when comparing the rates of these variant classes in highly constrained genes (pLI≥0.9, pLI≥0.995) between affected and unaffected children (logistic
regression, \( \text{pLI} \geq 0.9 \) missense \( p = 0.63 \), \( \text{pLI} \geq 0.9 \) PTV \( p = 0.69 \), \( \text{pLI} \geq 0.995 \) missense \( p = 0.07 \),
\( \text{pLI} \geq 0.995 \) PTV \( p = 0.99 \) (Figures 2B and 2C). We next assessed the rate of rare inherited variants in affected and unaffected children. We found no excess of rare inherited protein truncating variants (PTVs) (Methods, logistic regression, \( p = 0.60 \)) and missense variants (logistic regression, Polyphen-2 mis3 (Adzhubei et al., 2010) \( p = 0.62 \); MPC \( \geq 1 \) (Samocha et al. 2017) \( p=0.77 \)) in affected children (Figure 2D). The same held true when limiting the analysis to rare inherited PTV and missense variants in highly constrained genes (logistic regression, \( \text{pLI} \geq 0.9 \) missense mis3 \( p = 0.78 \), \( \text{pLI} \geq 0.9 \) missense MPC\( \geq 1 \) \( p=0.4 \), \( \text{pLI} \geq 0.9 \) PTV \( p = 0.12 \), \( \text{pLI} \geq 0.995 \) missense mis3 \( p = 0.96 \), \( \text{pLI} \geq 0.995 \) missense MPC\( \geq 1 \) \( p = 0.22 \), \( \text{pLI} \geq 0.995 \) PTV \( p = 0.89 \) (Figures 2E and 2F). Similarly, we observed no difference in the rates of private inherited variants (overall, PTV, and missense variant rates in all genes and in highly constrained genes) between affected and unaffected children (Figures 2G-2I).

Non-coding variation in multiplex families

We extended our investigation of the distinct genetic architecture of ASD multiplex families to non-coding regions of the genome. Previous category-wide association studies (CWAS) carried out in simplex families have suggested a biologically plausible association of promoter regions with ASD supported by de novo variation (Werling et al. 2018; An et al. 2018). Specifically, it has been observed that this non-coding association signal is driven by evolutionary conservation and that it is stronger in promoter regions of conserved loci more distal to the transcription start site (TSS) (An et al. 2018). Therefore, we first compared non-coding variant burden between affected and unaffected children in iHART for variants mapping to 3kb (2kb upstream and 1kb downstream of the TSS) promoters of all genes. We found no significant enrichment of promoter
rare inherited (logistic regression, p = 0.81), private inherited (logistic regression, p = 0.14), and rare de novo (logistic regression, p = 0.5) variants in cases (Figures S4A, S4B, and S4C). These results held true when focusing the rare and private inherited non-coding variant testing to promoters of ASD risk genes (Methods, Figures S4D and S4E). We extended the promoter definition from 2kb to 10kb upstream of the transcription start site with the aim of capturing distal promoter regions. The extension of promoter regions to 11kb (including 1kb downstream of TSS) did not alter the previous results (logistic regression; rare inherited, p = 0.87; private inherited, p = 0.11; rare de novo, p = 0.55) (Figures S4A, S4B, and S4C).

The strongest enrichment of de novo conserved promoter variants in ASD simplex cases in prior work has been found at transcription factor binding sites (TFBSs) (An et al. 2018). Hence, we filtered variants in the 11kb promoter regions of all genes by keeping variants in regions defined by three different genome-wide maps of TFBSs: (1) cis-regulatory modules (CRMs), which are clusters of predicted TFBSs (Gheorghe et al. 2019), as well as (2) cistromes and (3) cismotifs, which are experimentally and technically reproducible TF binding regions (Vorontsov et al. 2018). Again, we found no significant excess of TFBS rare inherited (logistic regression; CRMs, p = 0.99; cistromes, p = 0.91; cismotifs, p = 0.9), private inherited (logistic regression; CRMs, p = 0.21; cistromes, p = 0.13; cismotifs, p = 0.15), and rare de novo (logistic regression; CRMs, p = 0.98; cistromes, p = 0.58; cismotifs, p = 0.76) variants in iHART affected children compared to unaffected children (Figures S4F, S4G, and S4H).

Finally, we mapped non-coding variants within the 11kb promoters of all genes and non-coding variants across the whole genome to the brain-specific regulatory regions under sequence constraint within the human lineage, identified through the recent co-localization approaches (Xu et al. 2020). Xu et al. previously showed these brain-specific regulatory regions were enriched
for non-coding *de novo* variants in affected compared to unaffected children from ASD simplex families (Xu et al. 2020). In our study, we found no significant burden of brain-specific constrained functional non-coding rare inherited (logistic regression; 11kb promoter intersected set, p = 0.93; genome intersected set, p = 0.82), private inherited (logistic regression; 11kb promoter intersected set, p = 0.27; genome intersected set, p = 0.15), and rare *de novo* (logistic regression; 11kb promoter intersected set, p = 0.54; genome intersected set, p = 0.18) variants in iHART cases (Figures S4I, S4J, and S4K). We were not able to replicate in iHART known signals for non-coding *de novo* variant sets (Table S2), consistent with findings from our previous study (Ruzzo et al. 2019). This could be the result of the *de novo* depletion observed in this multiplex cohort and decreased power for non-coding association testing.

Previous power calculation for non-coding variation across ranges of relative risks, category sizes, and sample sizes assessed with the CWAS framework in simplex families confirmed how different the frequency and effect size of non-coding variant categories are from the well-defined and functionally characterized coding variants (Werling et al. 2018). It was estimated that over 8,000 simplex families would be needed to have sufficient power to discover a *de novo* non-coding class equivalent to *de novo* missense variation (Werling et al. 2018). From the filtering strategies we performed, we narrowed iHART non-coding variants to putatively functional sets more likely to be associated with ASD. Across all tested sets we found very small effect sizes for non-coding variation (Table S2), which results in low power at our current sample size and is consistent with previous non-coding investigation focusing on *de novo* variants (Werling et al. 2018; An et al. 2018) (Table S3). In addition, taking advantage of the multiplex family study design, we assessed the range of effect sizes for non-coding rare inherited variants, confirming estimates similar to those for the non-coding *de novo* variants (Werling et
al. 2018) (Table S2). Based on the observed effect sizes from our study, we computed power by simulation (Table S3) and predicted that a sample size of at least 8,000 fully-phaseable children would provide sufficient power for the investigation of private inherited variants in 11kb promoters of all genes, whereas 9,000 children would be needed to test association of non-coding private inherited variants in functional regions such as cistromes, cismotifs, and brain-specific co-localized regions (Figure S4L-N).

Seven New ASD Risk Genes

We do not observe a global excess for damaging rare de novo variants in affected children in these multiplex families, consistent with a lower contribution and distinct genetic architecture compared with simplex families. However, we still do identify pathogenic de novo variants in previously established ASD-risk genes in some affected family members. We therefore used the Transmitted And De novo Association (TADA) test (He et al. 2013), a Bayesian framework that combines inherited and de novo genetic variants, to identify ASD risk genes in this cohort. Qualifying variants included rare de novo or transmitted PTVs and de novo missense variants predicted to damage the encoded protein (Adzhubei et al., 2010) (Methods, Figure 3A, and Table S4). We combined this extended multiplex cohort with 1,660 affected children and their fully-phaseable parents with qualifying variants from previous ASD genetic analyses (Sanders et al. 2015; Iossifov et al. 2014; De Rubeis et al. 2014; Pinto et al. 2014) (Methods, Figure 3A, Figure S5A). We identified 74 genes significantly associated with ASD (FDR<0.1) (Figure 3B). Of these 74 genes, 46% were supported only by de novo variants. The remaining genes (54%) were supported by inherited variants, half of which had more than 50% of genetic variants with an inherited component (Figure 3B). There were nine genes identified in this study, but not our
previous iHART study (Ruzzo et al. 2019). Two genes (MED13L, KMT2A) were previously identified in ASD studies and harbored de novo variants (Simons Foundation Autism Research Initiative 2020; Satterstrom et al. 2020) as was also shown in our study. Seven genes are novel, having not been identified in previous studies (PLEKHA8, PRR25, FBXL13, VPS54, SLFN5, SNCAIP, TGM1) (Figure 3C). Consistent with the distinct architecture involving more inherited variation in multiplex ASD, five of these genes were predominantly supported by inherited protein truncating variants (PRR25, FBXL13, VPS54, SLFN5, TGM1).

We next assessed the shared biological functions among the 74 ASD risk genes from our TADA analysis. These genes formed a significant indirect protein-protein interaction (PPI) network (Methods, p=0.0079) involving 39 genes as significant seed genes, meaning they were more connected to other genes in the network than expected by chance (Figure 3D). These genes enriched for gene ontology terms including synaptic transmission, learning, social behavior and long-term synaptic depression (Figure 3E).

Previous work has demonstrated that ASD risk genes are most highly expressed as a group during brain development (Satterstrom et al. 2020; Ruzzo et al. 2019; Sanders et al. 2015; De Rubeis et al. 2014; Iossifov et al. 2014; Krumm et al. 2015). We subsequently used human brain fetal single cell transcriptomics (Polioudakis et al. 2019) to determine in which cell types the 74 ASD risk genes were expressed. We observed enriched expression of ASD risk genes in excitatory neurons (Expression Weighted Cell Type Enrichment (EWCE) migrating excitatory neuron FDR=0.0015, maturing excitatory neuron FDR=0.00016, Methods, Figures 4A,4B) and interneurons (EWCE interneurons CGE FDR=0.027, interneuron MGE FDR=0.04, Methods, Figures 4A,4B) as expected based on previous results (Ruzzo et al. 2019; Satterstrom et al. 2019).
We next evaluated how these 74 ASD risk genes varied during development using data from two relevant systems: long-term maturation in 3-dimensional induced pluripotent stem cell (IPSC)-derived human forebrain organoid model, human cortical spheroids (hCS), which have been shown to largely recapitulate human fetal brain development up to 600 days postnatally in vitro (Gordon et al. 2021) and the BrainSpan dataset, which represents an in vivo reference for cortical development (Stein et al. 2014; Kang et al. 2011; M. Li et al. 2018). Given the developmental function of ASD risk genes identified through de novo variants (Ronemus et al. 2014; Satterstrom et al. 2020; Iossifov et al. 2014), genes with predominantly de novo variant support (<50% inherited component support) were mostly highly expressed early in development in both the hCS model and BrainSpan (Figure 4C), consistent with earlier findings (Parikshak, Gandal, and Geschwind 2015). Interestingly, the average expression of genes with predominantly de novo variant support had higher expression compared to the average expression of genes with predominantly inherited variant support during this period as well.

We also clustered the ASD risk genes using the hCS trajectory data, which identified three clusters. ASD risk genes with greater than a 50% inherited component were represented in all clusters. Cluster 1 included genes which gradually rose during development and were enriched for genes involved in learning and memory (Figure 4D). Clusters 2 and 3 included genes that were elevated early in development and were enriched for genes with histone transferase activity (Figure 4D). Overall, these results show ASD risk genes with predominant de novo support have highest expression early in development and decline over time, while a subset of ASD risk genes with both de novo or inherited support (Cluster 1) that are involved in cognition and learning rise more gradually across fetal neuronal development.
Known ASD Risk Genes in Multiplex Families support Oligogenic Burden

We next collated a list of all known ASD risk genes to determine their contribution to risk in multiplex families by identifying unique ASD risk genes from our study and recent large ASD genetic studies (Satterstrom et al. 2020; Sanders et al. 2015) with an FDR < 0.1, yielding a total of 152 genes (Table S5). Not surprisingly, affected children had significantly more rare inherited protein truncating variants in known ASD risk genes (KARGs) compared to unaffected children (p=0.0196, Figure 5A). In contrast to observations in simplex families (O’Roak et al. 2012; Iossifov et al. 2014; O’Roak et al. 2014; De Rubeis et al. 2014; Krumm et al. 2015; Sanders et al. 2015; Turner et al. 2016; Stessman et al. 2017; Satterstrom et al. 2020), rare de novo protein truncating variants in known ASD risk genes were not increased in affected versus unaffected children (p=0.366, Figure 5B).

Given the lower female to male ratio of ASD and multiple threshold theory of a female protective effect (Maenner et al. 2020; Fombonne 2009; Werling and Geschwind 2013), we examined rare inherited protein truncating variants in known ASD risk genes by sex. There was no difference in mean number of rare inherited protein truncating variants in known ASD risk genes in affected females versus affected males (logistic regression, p=0.94, Figure 5C, Methods), or affected versus unaffected children stratified by sex (Figure 5C). Both paternal and maternal inheritance of variants has been associated with ASD risk (Brandler et al. 2018; Jacquemont et al. 2014). We observe no difference between the two modes of inheritance in this cohort (Figure 5D), although we acknowledge that larger population scale studies may be needed to see more subtle differences.

We next examined whether there were affected children who inherited more than one PTV in a known ASD risk gene (KARG), referred to as oligogenic transmission (Turner et al.
Indeed, we found that there was an excess of affected children with one or two protein truncating variants in KARGs (Mann-Whitney test, \(p=0.022\), Figure 5E, Methods). But, we did not observe this effect when considering mutation burden across all genes, consistent with the recognized mutation intolerance of KARGs (\(p=0.486\), Figure S5B) (Sanders et al. 2015; Samocha et al. 2017). We also evaluated known risk genes for ASD and developmental delay (DD) (Deciphering Developmental Disorders Study 2017), which totaled 208 genes (Table S5). Similar to observations in known ASD risk genes, the ASD and DD risk genes showed an increase in the number of inherited protein truncating variants (\(p=0.027\), Figure 5A) and an excess of one or two protein truncating variants (\(p=0.017\), Figure 5E) in affected children. There was no significant difference of known risk genes for ASD and developmental delay by sex (Figure 5C) or inheritance type (Figure 5D).

To further investigate these observations, we identified six affected children carrying two rare inherited PTVs in known ASD risk genes: specifically, these children are sibling pairs from three different families (Figure 6). In family A, the two affected siblings inherit rare PTVs in both CAPN12 and FBXL13 from the father, while the unaffected child has inherited only the FBXL13 variant. In family B, both affected children inherited a PTV in MFRP from the father and a PTV in SHANK3 from the mother. Interestingly, in family C, of the three affected dizygotic twins, the two females inherited PTVs in both RANBP17 and WDFY3 from the father, while the male did not. Intriguingly, the two female affected twins had their first steps and phrases much later than their male affected twin who, in contrast to the two affected females, did not show severely impaired social behavior and social interaction skills. Of note, RNA-seq analysis on 120 human fetal brains highlighted RANBP17 as a high-confidence ASD risk gene exhibiting higher expression in females (FDR < 0.1) (O’Brien et al. 2019).
Finally, we investigated the effect of rare inherited and rare de novo variants in these 152 known ASD risk genes on critical phenotypic measures describing affected children’s cognitive, social, and motor development. We observed significantly lower verbal IQ scores in all groups of affected children including carriers of rare inherited KARG variants (p=0.001), de novo carriers (p=0.02), and non-carriers (p<2x10^-16, Figure 7A). Interestingly, we found significantly higher than population average non-verbal IQ scores (Raven population average = 100) in the affected non-carriers (p=1x10^-7, Figure 7B), consistent with IQ lowering effects observed with rare variants in other studies (Satterstrom et al. 2020; Weiner et al. 2017). Both affected non-carriers and rare inherited carriers showed significantly delayed first steps (affected non-carriers p=1x10^-9, rare inherited carrier p=0.03, Figures 7C), first words (affected non-carriers p<2x10^-16, rare inherited carrier p<2x10^-16, Figures 7D), and first phrases (affected non-carriers p<2x10^-16, rare inherited carrier p<2x10^-16, Figures 7E). We observed the largest effect of rare inherited KARG variants on language development (Figures 7D and 7E), specifically on the “first phrase” milestone achievement, with a median delay of 30 months compared to unaffected children from the same cohort (Figure 7E). We observed a similar significant trend for rare de novo carriers, with a median delay of 22 months (Figure 7E). As expected, all groups of affected children showed significantly higher SRS total raw scores compared to unaffected children from the same cohort (Figure 7F). We did not find significant differences among the three affected groups for any of the tested measures (Kruskal-Wallis test, Methods).

Overall, these findings show that contributory rare protein truncating variants from known ASD risk genes in affected children are inherited from unaffected parents. Furthermore, some affected children likely have an oligogenic burden. Lastly, ASD phenotypes, particularly non-verbal IQ and language development, vary by the form of known ASD risk gene inheritance.
Polygenic Risk is Over Transmitted in Affected Children with Inherited Variants and Associated with ASD Phenotype

Given the pleiotropy associated with de novo variants and complex patterns of inheritance including oligogenic inheritance of rare PTVs, we next assessed the degree to which polygenic risk (PGR) contributes to ASD susceptibility in those harboring either de novo or inherited rare variants in the 74 TADA genes identified in this cohort. This is especially important, given the observation of a contribution from rare inherited variants from unaffected parents. We used the polygenic transmission disequilibrium test (pTDT) (Weiner et al. 2017), an extension of the TDT (Spielman, McGinnis, and Ewens 1993) to polygenic risk scores, which compares the mean of the offspring polygenic risk score (PRS) distribution to its parentally derived expected value. We confirmed overall, that affected children have an over transmission of ASD polygenic risk (two-sided, one-sample t-test, p=0.005, Figure 7G, Methods) compared with unaffected siblings, to whom no over-transmission of PRS is observed (p=0.08, Figure 7G).

To compare the contributions of PRS in the different variant classes, we stratified both affected and unaffected children into those with rare de novo PTV and missense variants in ASD risk genes and those with rare inherited PTVs (Figure 7G). Among affected children with an inherited protein truncating variant from statistically supported ASD risk genes identified in this study, we observed a significant over-transmission of PGR (p=0.05). This additive effect of ASD PGR to ASD rare inherited variation was not observed for PRS for schizophrenia (SCZ), bipolar disorder (BD), and educational attainment (EA) (Figure S6A, Methods). In contrast, affected children harboring a rare de novo PTV and/or missense variant (p=1, Figure 7G) and unaffected children with an inherited PTV variant do not exhibit an over-transmission of PRS (p=0.2,
Figure 7G). We observed an over-transmission of ASD common risk in both affected (p=0.02) and unaffected (p=0.03) non-carriers (Figure 7G).

To investigate the effect of common risk for ASD on phenotypes related to cognitive and social impairment, and motor and language development, we performed pTDT analysis for affected non-carriers divided into typical and atypical groups based on critical cutoffs for each measure of interest (Methods). Interestingly, we observed the significant over-transmission of ASD PRS in children showing an atypical language development (age of first word and phrase greater than 24 and 33 months, respectively, suggesting developmental concerns before or at 36 months) (first word p=0.01, first phrase p=0.03, Methods, Figure 7H) and not in those with a typical language development (age of first word and phrase before or at 24 and 33 months, respectively, Methods) (first word p=0.4, first phrase p=0.4, Methods, Figure 7H). We did not find the same effect when testing PRS for SCZ, BD, and EA (Figure S6B, Methods). These findings suggest that the influence of common risk on child language development is specific to ASD polygenic risk.

In addition, we stratified the affected non-carriers by their social behavior and social interaction skills using T-score cut-off ranges from the Social Responsiveness Scale (SRS). As expected, the social impairment of these affected children falls within a spectrum: despite their diagnosis. Some children show neurotypical social skills; in contrast, others show deficiencies in reciprocal social behavior that would be categorized as mild or severe ASD (Figure 7I, Methods). Interestingly, we found a significant over-transmission of ASD PRS only in children with severe social impairment (p=0.05, Figure 7I, Methods) and not in those with mild social impairment (p=0.4, Figure 7I). We did not observe any difference in the contribution of common risk to social behavior among these groups when testing PRS for SCZ, BD, and EA (Figure S6C,
Methods). These results confirm social behavior impairment as a core diagnostic domain for ASD and suggest its association with ASD PRS. Finally, we did not note any suggestive link between ASD PGR over-transmission and cognitive impairment (Figures S6D and S6E, Methods) or motor development (Figure 5E, Methods).

Taken together, these findings indicate that polygenic ASD risk is over transmitted to affected children, especially those with inherited protein truncating variants in ASD risk genes, and far less so to those harboring rare de novo variants. Furthermore, over transmission of common ASD risk is observed in ASD children with more severe language and social impairment, connecting language impairment to common genetic risk for ASD. This extends our understanding of ASD risk transmission to demonstrate co-inheritance of both rare and common risk in ASD probands.

**Discussion**

Most genetic studies on ASD focus on simplex families, in which only one child is affected and optimally isolates strongly acting, rare de novo variants (O’Roak et al. 2012; Iossifov et al. 2014; O’Roak et al. 2014; De Rubeis et al. 2014; Krumm et al. 2015; Sanders et al. 2015; Turner et al. 2016; Stessman et al. 2017; Satterstrom et al. 2020). These large sequencing efforts on simplex cohorts have revolutionized ASD research, elucidating ASD genetic architecture and facilitating the identification of high-confidence risk genes for the disorder. However, our previous study on the first release of the iHART cohort, which was the largest cohort composed of multiplex families with more than one affected child, detected for the first time a strong additional risk contribution from rare inherited variants which was not detected in simplex families (Ruzzo et al. 2019). This previous study suggested a different genetic architecture for multiplex families.
which subsequently led to the discovery of new ASD risk genes whose association signal was
derived from rare inherited variation and that impacted distinct biological processes from those
hit by rare de novo variation (Ruzzo et al. 2019). WGS analysis on the extended iHART/AGRE
cohort, which is two times the size, provided the opportunity to (1) validate and extend the list of
ASD risk genes supported by rare inherited variation, (2) confirm the distinct ASD genetic
architecture of multiplex families in a sufficiently large cohort, and (3) thoroughly assess the
additive effects of both common and rare variation for ASD and its phenotypic spectrum.

We identified 7 novel ASD risk genes, 5 of which are impacted by a significantly higher
proportion of rare inherited PTVs versus rare de novo variants. Our extended cohort analysis also
validated two recently identified ASD risk genes, KMT2A and MED13L, and confirm that their
association signal is mainly supported by rare de novo variation (Satterstrom et al. 2020). The
overall yield of new risk genes defined by the presence of rare variants in this study likely
reflects diminishing returns of novel ASD risk genes with increasing sample sizes, as has been
projected for exome sequencing studies (Coe et al. 2019). We also did not identify an increased
number of non-coding rare inherited, private inherited, or rare de novo variants in affected
children, despite focusing on putatively functional elements of the non-coding genome such as
brain-specific co-localized regions and TFBSs (Xu et al. 2020; Gheorghe et al. 2019; Vorontsov
et al. 2018). Non-coding ASD genetic risk in simplex families has been identified from large
DNase I hypersensitivity regions (Turner et al. 2016), TFBSs within the distal promoters of
conserved loci (An et al. 2018), and non-coding variants predicted to be deleterious from deep-
learning (Zhou et al. 2019). Our analysis agrees with previous analyses suggesting that sample
sizes of roughly 8000 fully-phaseable children would be necessary to be able identify non-coding
ASD genetic risk with 80% power (Werling et al. 2018).
Of the 74 genes associated with ASD (FDR < 0.1) from the current study, 39 converge in a significant indirect protein-protein interaction network. These significant seed genes include known ASD risk genes (*CHD8, GRIN2B, SHANK3*) (Ruzzo et al. 2019; Satterstrom et al. 2020) and newly identified risk genes (*FBXL13, SNCAIP*). Common genetic variation in FBXL13 was associated with sleep-wake cycle preferences (Veatch et al. 2017; Hu et al. 2016, 89; Lane et al. 2016), which is disrupted in ASD. *SCNAIP*, also known as Synphilin-1, has been extensively studied by promoting neurodegeneration in Parkinson’s disease when interacting with alpha-synuclein (Chung et al. 2001; Engelender et al. 1999, 1; Krüger 2004; Dashtipour et al. 2017). Parkinsonism and movement disorders including repetitive behaviors are recognized features of ASD (Qiu et al. 2010; Estes et al. 2011; Abbott et al. 2018; Starkstein et al. 2015; Bell, Wittkowski, and Hare 2019).

The predominant contribution of rare inherited variants to ASD risk in multiplex families is confirmed by our focused and comprehensive analysis on known ASD risk genes that had been previously identified in cohorts with predominantly simplex families (Satterstrom et al. 2020; Turner et al. 2017). Interestingly, affected children in multiplex families have a significant excess of rare inherited, but not rare *de novo* protein truncating variants in known ASD risk genes. We found that, in addition to ASD resulting from rare inherited variation, inherited multiplex ASD risk may in some cases involve transmission of more than one rare PTV to affected children. Affected children are more likely to carry one or two rare inherited PTVs in known ASD risk genes compared to unaffected children. A similar oligogenic burden in affected children was identified for *de novo* variants in simplex families (likely gene disrupting, severe missense variants, in fetal brain transcription factor binding sites, 3’ UTR and deletions that disrupt an exon) (Turner et al. 2017). We identified three pairs of siblings with two rare inherited
PTVs in known ASD risk genes. In one family with two affected females and one affected male, the females had two inherited PTVs in known ASD risk genes (*WDFY3*, *RANBP17*) while the male, who had milder ASD symptoms, did not share these same variants. *RANBP17*, which was mutated in affected females of this family, showed a female-biased expression in human fetal brain (O’Brien et al. 2019), but not in human adult brain (Oliva et al. 2020). This specific case and sex specific expression of *RANBP17* supports the multiple threshold theory in females with ASD (Maenner et al. 2020; Fombonne 2009; Satterstrom et al. 2020; Werling and Geschwind 2013), which posits females require a greater burden of genetic risk to develop ASD compared to males.

Previous analyses investigating additivity between common polygenic and rare variation in simplex families showed that ASD common polygenic risk is over transmitted in affected children carrying a strongly acting *de novo* variant (PTVs and deletions) (Weiner et al. 2017). In this cohort, we showed that PRS is significantly over-transmitted to affected children harboring inherited protein coding rare variants in ASD risk genes. Because we observed a small, but significant over-transmission of ASD common risk in both affected and unaffected children who did not carry rare inherited ASD risk genes, our findings suggest the over-transmission of PRS in ASD acts in combination with rare inherited ASD genetic risk towards the development of ASD. Thus, PRS combines with rare variants to influence risk for ASD (Takahashi et al. 2020; Serdarevic et al. 2020) and may partially explain why parents with protein coding rare variants in ASD risk genes are asymptomatic, as they lack the polygenic ASD risk which was transmitted to their affected children.

Another important observation is that ASD PRS is not only over-transmitted to affected children with severe social impairment, but also those with atypical language development at or
before 36 months. In contrast, PRS for SCZ, BD, and EA did not show a similar effect on language development and social behavior. Other recent work (in preprint form) found a positive association between ASD polygenic risk and earlier age of first word in affected children from SSC and SPARK (Warrier et al. 2021). However, this signal was no longer significant after accounting for presence of high-impact de novo variants and full-scale IQ (Warrier et al. 2021).

Interestingly, results from another recent preprint, relevant to our SRS finding, show a positive association between polygenic risk for ASD and severity of social communication deficits, but not of social behavior impairment measured by SRS (Antaki et al. 2021). We did not observe any suggestive effect of ASD common risk on cognitive impairment or motor development, consistent with observations from other cohorts highlighting the importance of rare variation vis a vis the manifestation of these phenotypes (Weiner et al. 2017; Antaki et al. 2021).

The novel link between language delay and over-inheritance of PRS for ASD shown here supports the relevance of assessing accompanying language impairment in the context of ASD diagnosis. Moreover, these results suggest that indeed language impairment is a core feature of ASD and should not be a secondary consideration with regards to diagnosis, as it is in the current version of DSM-V. Since this is the first time an association between ASD PRS and language delay has been observed, it will be important to replicate it in additional cohorts. Nevertheless, this work highlights the value of multiplex family cohorts in ASD research where the effects of rare and common inherited variation are more easily observed and can be associated with neurobiologically relevant phenotypes, such as language or motor function.
Acknowledgments

We thank the subjects and families from the Autism Genetic Resource Exchange (AGRE) who participated in this study, in addition to the Hartwell Foundation, and New York Genome Center. We thank Stephanie N. Kravitz, Cheyenne L. Schloffman, David Keller, Min Sun, Tor Solli-Nowlan, Sasha Sharma, Marlena Duda, Greg Madden McInnes, Ravina Jain, Valenti ’Moncunill, Josep M. Mercader, Montserrat Puiggros, Anika Gupta, and David Torrents for technical support. We are grateful to all of the families at the participating SSC sites as well as the principal investigators (A. Beaudet, R. Bernier, J. Constantino, E. Cook, E. Fombonne, D. Geschwind, R. Goin-Kochel, E. Hanson, D. Grice, A. Klin, D. Ledbetter, C. Lord, C. Martin, D. Martin, R. Maxim, J. Miles, O. Ousley, K. Pelphrey, B. Peterson, J. Piggot, C. Saulnier, M. State, W. Stone, J. Sutcliffe, C. Walsh, Z. Warren, and E. Wijsman).

Funding/Support

This work was supported by grants from The Hartwell Foundation and the NIH (R01MH100027, R01MH064547, S10OD011939, P50HD055784, UM1HG008901, P50DC018006, K08AG065519 R01MH064547, S10OD011939) and from the Stanford Precision Health and Integrated Diagnostics Center and from the Stanford Bio-X Center. We are grateful to The Hartwell Foundation for supporting the genetic sequencing and analyses of iHART. We thank the Hartwell Foundation for the opportunity to contribute to their mission of benefitting children, and for the financial support that enabled the whole-genome sequencing and analyses. We are grateful to the Simons Foundation for additional support for genome sequencing. We thank the New York Genome Center for conducting sequencing and initial quality control. We thank Amazon Web Services for their grant support for the computational infrastructure and
storage for the iHART database. The Autism Genetic Resource Exchange (AGRE) is a program of Autism Speaks and was supported by grant NIMH U24 MH081810. The Simons Simplex Cohort data described in this study can be obtained through application to SFARI Base (https://base.sfari.org).

**Author Contributions**


**Declaration of Interests**

The authors declare no competing interests
Figure Legends

Figure 1. Overview diagram of study analyses. Whole genome sequencing was performed on 1,004 multiplex ASD variants. Single nucleotide variants (SNV) and small insertions/deletions (Indels) were called using the Genome Analysis Toolkit (GATK). Variant, sample and genotype quality control was performed. Rare de novo, rare inherited, and private inherited variants were included for downstream analyses. We identified ASD risk genes using the Transmission and de novo association (TADA) analysis. Risk genes were characterized using integrative genomics. We also investigated the genetic architecture of ASD phenotypes. We identified genes supported by rare inherited risk by evaluating the phenotype of children with multiple inherited risk genes. We tested if common ASD polygenic risk was over transmitted to affected versus unaffected children stratified by rare de novo and inherited variants.

Figure 2. (A-C) Rate of rare de novo variants in all genes, genes with pLI≥0.9, and genes with pLI>0.995 stratified by variant type. (D-F) Rate of rare inherited variants in all genes, genes with pLI>0.9, and genes with pLI≥0.995 stratified by variant type. (G-I) Rate of private inherited variants in all genes, genes with pLI≥0.9, and genes with pLI>0.995 stratified by variant type. mis3 = Polyphen-2 missense 3, MPC≥1=missense deleterious metric

Figure 3. 74 ASD risk genes identified by Transmission and de novo association (TADA) analysis. (A) Overview of iHART subjects and additional datasets included in the TADA analysis. (B) 74 ASD risk genes identified in the iHART TADA analysis (FDR<0.1) are displayed as the -log10(q-value). The nine newly identified ASD risk genes from our expanded iHART cohort are bold. The dashed horizontal line marks the FDR=0.1 threshold. Bars are
colored by the proportion of inherited protein truncating variants (PTVs) for each gene (inherited PTVs/(inherited PTVs + de novo PTVs + de novo Mis3 + de novo small deletions). (C) The nine newly identified ASD risk genes from our expanded iHART cohort. Bars are colored according to (B). (D) Protein-protein interaction network formed by the 74 ASD risk genes using Disease Association Protein-Protein Link Evaluator (DAPPLE). Genes are colored by the proportion of inherited PTVs. The gene name for two of the seven newly identified ASD risk genes are in orange. (E) Gene Ontology enrichment terms for 39 significant seed genes from the protein-protein interaction network.

Figure 4. Characterization of TADA ASD risk genes during development. (A) Mean expression and percentage of cells expressing TADA ASD risk genes from single cell data of the human fetal brain cortex (Polioudakis et al. 2019). * Indicates cell type specific TADA ASD risk genes (Methods). (B) Cell types enriched in TADA ASD risk genes tested using Expression Weighted Cell Type Enrichment (EWCE). * FDR<0.05. (C) TADA ASD risk genes variation during development using a long-term maturation human cortical spheroid (hCS) and BrainSpan. ASD risk genes are stratified by those with predominantly inherited or de novo support. (D) Clustering of the ASD risk genes from hCS trajectory data identified 3 clusters. Days post conception, TADA class (predominant inherited or de novo) and scaled expression are shown. The average expression of each cluster and significant gene ontology biological terms are shown. vRG=ventral radial glia, PgS = cycling progenitor (S phase), PgG2M= cycling progenitor (G2M phase), Per=pericyte, oRG=outer radial glia, OPC=oligodendrocyte precursor, Mic=microglia, IP=intermediate progenitor, InMGE=Interneuron medial ganglionic eminence, InCGE=interneuron caudal ganglionic eminence, ExN=excitatory neuron, ExM−U=maturing
excitatory neuron upper enriched, ExM=maturing excitatory neuron, ExDp2=excitatory neuron deep layer 2, ExDp1=excitatory neuron deep layer 1, End=endothelial, ExMig=migrating excitatory neuron, ExDp=excitatory neuron deep

Figure 5. Known ASD risk genes variant rates, and known ASD and developmental disorder risk genes variant rate. Mean number of protein truncating variants in affected versus unaffected children with inherited (A) and de novo (B) variants. Mean number of inherited protein truncating variants in affected versus unaffected children stratified by sex (C) and inheritance type (D). Fraction of subjects and number of genes with protein truncating variants from known ASD risk genes or known ASD and developmental disorder risk genes (E).

Figure 6. iHART family pedigrees where affected children inherited more than one protein truncating ASD risk gene.

Figure 7. Phenotypic effect of rare and common risk for ASD in iHART affected children. (A-F) Phenotypic distributions in iHART affected children, stratified by those carrying rare inherited (RI) or rare de novo (RDN) variants in known ASD risk genes (KARG) and those not carrying such variants. (A) Verbal IQ distribution for affected children. Standard scores from PPVT-3 (Peabody Picture Vocabulary Test, 3rd edition) were used and compared to a standard hypothesized median value of 100. Affected non-carrier: n = 767, mu = 89; affected RI carrier: n = 70, mu = 88; affected RDN carrier: n = 10, mu = 84.5. (B) Non-verbal IQ distribution for affected children. Standard scores from the Raven test were used and compared to a standard hypothesized median value of 100. Affected non-carrier: n = 987, mu = 104; affected RI carrier:
n = 98, \( \mu = 100 \); affected RDN carrier: \( n = 12, \mu = 108.5 \). (C) Age of walking distribution for affected children whose ADI-R (Autism Diagnostic Interview-Revised) data was available. A standard hypothesized median value of 12 months, observed in the normative cohort of the WHO Multicentre Growth Reference Study (MGRS) (Onis 2006), was used for comparison. Affected non-carrier: \( n = 1180, \mu = 12 \); affected RI carrier: \( n = 120, \mu = 13 \); affected RDN carrier: \( n = 21, \mu = 13 \). (D) Age of first word for affected children whose ADI-R data was available. Affected non-carrier: \( n = 1223, \mu = 26 \); affected RI carrier: \( n = 126, \mu = 30 \); affected RDN carrier: \( n = 21, \mu = 18 \); unaffected non-carrier: \( n = 295, \mu = 12 \). (E) Age of first phrase for affected children whose ADI-R data was available. Affected non-carrier: \( n = 1220, \mu = 42 \); affected RI carrier: \( n = 126, \mu = 48 \); affected RDN carrier: \( n = 21, \mu = 40 \); unaffected non-carrier: \( n = 283, \mu = 18 \). (F) SRS (Social Responsiveness Scale) total raw score distribution for affected and unaffected iHART children. Affected non-carrier: \( n = 709, \mu = 104 \); affected RI carrier: \( n = 73, \mu = 98 \); affected RDN carrier: \( n = 8, \mu = 113.5 \); unaffected non-carrier: \( n = 153, \mu = 19 \). Non-carrier distributions are depicted in black, RI carrier ones in blue, and RDN ones in red. Standard hypothesized median values used for each phenotypic comparison are reported as green vertical lines. For panels D to F, these median values were derived from the corresponding phenotypic distributions for iHART unaffected non-carriers (shown in green) for whom phenotypic data was available. Each group of affected children was tested for difference among its phenotypic median value and the standard hypothesized one by two-sided one-sample sign test. The three tests performed for each phenotypic measure were corrected by multiple hypothesis testing using the Benjamini and Hochberg (BH) procedure as FDR (false discovery rate)-controlling method. An asterisk is used to denote an FDR < 0.05. Median values together with their 95% confidence intervals are represented by horizontal point ranges over the
distributions. (G-I) ASD polygenic risk transmission stratified by ASD risk gene carrier type and association with language and social impairment. (G) pTDT in affected and unaffected children, tested for all subjects (All - affected: n = 1231; unaffected: n = 288) and stratified by those carrying rare inherited (TADA RI carrier - affected: n = 70; unaffected: n = 8) or rare de novo (TADA RDN carrier - affected: n = 12) variants in the TADA ASD risk genes and those without such variants (non-carriers - affected: n = 1149; unaffected: n = 279). (H) pTDT in affected non-carriers of variants in the TADA ASD risk genes stratified by those with atypical (age of first word: n = 407; phrase: n = 564) language development (i.e., whose ages of first word and phrase indicate that developmental concerns are present at or before 36 months, according to corresponding ADI-R, Autism Diagnostic Interview-Revised, scoring system for the two items) and those with typical (age of first word: n = 432; phrase: n = 275) language development. (I) pTDT in affected non-carriers of variants in the TADA ASD risk genes stratified by the degree of social impairment based on SRS (Social Responsiveness Scale) T-score cut-off ranges. T-scores (based on child’s sex) equal to 59 or less were considered as normative (n = 122), those 60 to 75 were considered mild ASD (n = 70), and those equal to or greater than 76 were considered severe ASD (n = 275) (Lyall et al. 2014). Polygenic transmission disequilibrium is shown on the y axis of each panel as point ranges of the standard deviation on the mid-parent distribution mean values together with their 95% confidence intervals. The probability of each pTDT deviation distribution mean being equal to zero was tested by two-sided one-sample t-test. An asterisk is used to denote a p-value < 0.05.

Supplemental Figure Legends
Figure S1. WGS coverage statistics for 4,551 iHART/AGRE samples. (A) For the 4,551 iHART samples with WGS data, the percent of genomic bases covered at 1X, 10X, 20X, 30X, and 40X bases. On average, 99.23 ± 0.35% of bases were covered at a depth of ≥ 10X. The average fold coverage per sample across the cohort and no differences in the categories of (B) ASD affection status, (C) sex, or (D) family member type – where family member type was categorized as either Child (proband, sibling, MZ or DZ twin), Father, Mother, or Other (e.g., cousin). (E) The percent of exonic and genomic bases covered at ≥ 10x across all family members present in the 865 fully-phaseable iHART families. Exonic regions are those that are annotated as protein-coding exons (> 75MB) in Gencode V19. All non-N bases in the reference genome (>2.8Gb) were considered genomic regions.

Figure S2. Whole genome sequencing pipeline overview. An overview of the variant, sample, and genotype quality control conducted on the raw joint called variants to obtain the rare de novo, rare inherited, and private inherited variants used in downstream analyses. Abbreviations: VCF (Variant Call Format), QC (Quality Control), VQSR (Variant Quality Score Recalibration), IBD (Identity by Descent), DP (Read Depth), GQ (Genotype quality), AD (Allele Depth), HNP (Healthy Non-phaseable), PSP (Progressive Supranuclear Palsy), GIAB (Genome in a Bottle Consortium).

Figure S3. Rare de novo variants (RDNVs) in iHART samples before and after Artifact Removal by Classifier (ARC). (A) Distribution of raw de novo variant ARC scores from 2,100 fully-phaseable non-monozygotic (MZ) twin samples. Red represents samples derived from lymphoblastic cell lines (LCL) and green represents samples derived from whole blood (WB).
(B) In LCL (pink) and WB (blue) fully-phaseable (non-MZ twin) samples, the number of rare de novo variants identified per sample before ARC (N = 2,100 samples) and the number of rare de novo variants identified in LCL (pink) and WB (blue) fully-phaseable (non-MZ twin) samples after ARC (variants with an ARC score < 0.4 are filtered out) and after excluding ARC outlier samples (samples with > 90% DNs removed by ARC) (N = 1,443). After ARC, there is a less significant difference in the rate of rare de novo variants based on the biological sequencing source (LCL mean = 52.6 and WB mean = 53.6; LCL median = 51 and WB median = 53). The Wilcoxon rank sum test was used to evaluate the difference in DN rates between the biological sequencing source (LCL versus WB). (C) Histograms displaying the number of rare de novo coding variants per fully-phaseable sample. Before ARC, coding RDNVs come from 2,100 fully-phaseable non-MZ twin samples; after ARC (which filters out variants with an ARC score < 0.4), variants are derived from samples excluding ARC outliers (samples with > 90% DNs removed by ARC) and MZ twins (n = 1,443). (D) The correlation between paternal age and the rate of rare de novo variants before and after ARC. This analysis considers 1,677 fully-phaseable ASD children (excluding MZ twins) for which paternal age was known before ARC and 1,157 fully-phaseable ASD children (excluding MZ twins and ARC outliers) for which paternal age was known after ARC. The linear regression line is in red. On the left, the graph shows the number of raw rare de novo variants (SNVs and indels) per child by paternal age (years) at the time of the participant’s birth. On the right, the graph shows the number of rare de novo variants (SNVs and indels) per child identified after running ARC by paternal age (years) at the time of the participant’s birth.
Figure S4. (A-K) Number of non-coding variants per fully-phaseable child in affected (salmon) and unaffected (turquoise) children. Mean ± standard error variant count per child are shown. (A-D) For promoter regions, rare inherited non-coding variants in 1,836 affected versus 418 unaffected children (A), private inherited non-coding variants in 1,836 affected versus 418 unaffected children (B), and rare de novo non-coding variants in 1,164 affected versus 279 unaffected children (C). (D-E) For promoter regions of ASD risk genes, rare inherited non-coding variants in affected versus unaffected children (D), and private inherited non-coding variants affected versus unaffected children (E). (F-H) For transcription factor binding sites in promoter regions, rare inherited non-coding variants in affected versus unaffected children (F), private inherited non-coding variants affected versus unaffected children (G), and rare de novo non-coding variants in affected versus unaffected children (H). (I-K) For brain-specific co-localized regions, rare inherited non-coding variants in affected versus unaffected children (I), private inherited non-coding variants affected versus unaffected children (J), and rare de novo non-coding variants in affected versus unaffected children (K). (L-N) Curves showing power at increasing sample sizes for the quantitative burden testing through logistic regression of rare inherited (L), private inherited (M), and rare de novo (N) non-coding variants. Each curve corresponds to a tested variant set (see color-coded legend), with individual points representing power values specific to the “variant count per child” predictor, computed based on 1,000 simulations, for the observed ASD odds ratio (OR) and at specific sample sizes. 95% confidence intervals for these power values are depicted as ribbons. The dashed grey horizontal line marks the 0.8 standard threshold used for power while the solid grey vertical line shows the current sample size for the specific variant set (2,254 fully-phaseable children for rare and private inherited non-coding variants and 1,443 children for rare de novo non-coding variants).
Figure S5. (A) Detailed overview of iHART subjects and additional datasets included in the TADA analysis. Initial dataset included 2,254 children with fully-phaseable parents from 1,004 ASD multiplex families. Subjects were removed if they were unaffected, were considered ARC outliers or were a monozygotic twin. Subsets of the 1660 affected children were included in TADA for de novo or inherited analysis. Additional datasets used for TADA mega-analysis are shown. (B) Fraction of subjects and number of genes with protein truncating variants from all genes.

Figure S6. Polygenic risk for schizophrenia (SCZ), bipolar disorder (BD), and educational attainment (EA) in iHART and ASD pTDT stratified by cognitive and motor impairment. (A) pTDT for SCZ, BD, and EA in affected and unaffected children, tested all together (All - affected: n = 1231; unaffected: n = 288) and stratified by those carrying rare inherited (TADA RI carrier - affected: n = 70; unaffected: n = 8) or rare de novo (TADA RDN carrier - affected: n = 12) variants in the TADA ASD risk genes and those without such variants (non-carriers - affected: n = 1149; unaffected: n = 279). (B) pTDT for SCZ, BD, and EA in affected non-carriers of variants in the TADA ASD risk genes stratified by those with atypical (age of first word: n = 407; phrase: n = 564) language development (i.e., whose ages of first word and phrase indicate that developmental concerns are present at or before 36 months, according to corresponding ADI-R, Autism Diagnostic Interview-Revised, scoring system for the two items) and those with typical (age of first word: n = 432; phrase: n = 275) language development. (C) pTDT for SCZ, BD, and EA in affected non-carriers of variants in the TADA ASD risk genes
stratified by the degree of social impairment based on SRS (Social Responsiveness Scale) T-score cut-off ranges. T-scores (based on child’s sex) equal to 59 or less were considered as normative (n = 122), those 60 to 75 were considered mild ASD (n = 70), and those equal to or greater than 76 were considered severe ASD (n = 275) (Lyall et al. 2014). (D-E) pTDT for ASD in affected non-carriers of variants in the TADA ASD risk genes stratified by those with (IQ < 70) or without (IQ >= 70) cognitive impairment, based on their verbal IQ scores from PPVT-3 (Peabody Picture Vocabulary Test, 3rd edition) (D) and non-verbal IQ ones from the Raven test (E). Verbal IQ >= 70: n = 430; Verbal IQ < 70: n = 106; Non-verbal IQ >= 70: n = 682; Non-verbal IQ < 70: n = 39. (F) pTDT for ASD in affected non-carriers of variants in the TADA ASD risk genes stratified by those with typical (age of walking in months < 16, n = 669) or atypical (age of walking in months >= 16, n = 135) motor development. The age of 16 months chosen for stratification is the 97th percentile estimate for milestone achievement age in the normative cohort of the WHO Multicentre Growth Reference Study (MGRS) (Onis 2006). Polygenic transmission disequilibrium is shown on the y axis of each panel as point ranges of the standard deviation on the mid-parent distribution mean values together with their 95% confidence intervals. The probability of each pTDT deviation distribution mean being equal to zero was tested by two-sided one-sample t-test. An asterisk is used to denote a p-value < 0.05. More details provided in the Methods section.
Tables

Table S1. Sample demographics, family relationships and whole genome sequencing metrics

Table S2. Results from non-coding variant quantitative burden testing through logistic regression. For each non-coding variant set tested, total number of variants, significance of the logistic regression model built, and significance and ASD odds ratio (OR) for the three model predictors (variant count per child, child sex, and sample source of DNA used for whole-genome sequencing) are shown. Male sex and lymphoblastoid cell line (LCL) sample source of DNA were used as reference values for the two categorical variables. ASD OR for the variant count per child predictor represents the change in ASD odds for one unit increase of variant count per child.

Table S3. Results from power analysis for non-coding variant quantitative burden testing through logistic regression. For each non-coding variant set tested, power values for quantitative burden testing through logistic regression and specific to the variant count per child predictor are reported. Power was computed with a posthoc approach (for the observed ASD OR and at the current sample size) using three different simulation strategies (simr powerSim() function, “sample” custom function, and “random” custom function). Power was estimated over 1,000 simulations. A description of the simulation strategies can be found in the related methods’ section. 95% confidence intervals for the power values are shown within parentheses.

Table S4. Detailed TADA results including qualifying variants, cohorts included, results, and results by variant class.
Table S5. Comparison table of TADA ASD risk genes, Known ASD risk gene, and development disorder risk gene list. A “1” indicates the inclusion of a gene within the gene list, while a “0” represents the absence of a gene.
Methods

LEAD CONTACT AND MATERIALS AVAILABILITY

The Hartwell Foundation's Autism Research and Technology Initiative (iHART) will provide the whole genome sequencing data generated during this study upon request and approval of the data use agreement available at http://www.ihart.org. Lead Contact, Daniel Geschwind (dhg@mednet.ucla.edu) can be contacted for any further information or requests for resources and reagents.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

ASD multiplex family samples

As a result of this study being limited to previously-existing coded data and specimens, the UCLA and Stanford IRBs considered it “Not human subjects research” and thus exempt from review. Subjects originated from the Autism Genetic Resource Exchange (AGRE) (Lajonchere, 2010) and were carefully selected from families including at least two members with ASD. Sequencing was not performed on patients with known genetic causes of ASD or syndromes overlapping with ASD-features. Table S1 contains a complete list of sequenced samples.

Of subjects selected from the Autism Genetic Resource Exchange (AGRE), 4,551 individuals from 1,004 ASD families passed quality control. Table S1 provides information about these 4,551 samples, including sex, ethnicity, phenotype, and familial relationship. Our variant analysis included 4,531 individuals (2,136 affected and 2,395 unaffected); 20 children for whom both biological parents were not sequenced were excluded. If not otherwise specified, our
analyses included a subset of 2,254 children (1,836 affected and 418 unaffected) whose parents
had both undergone sequencing.

The Rutgers University Cell and DNA Repository (RUCDR; Piscataway, NJ) provided our
purified DNA. Whenever possible, whole blood DNA was used; for many samples, however,
only lymphoblastoid cell line (LCL) DNA was extracted at the time of recruitment.

**Control cohorts**

We refer to several control cohorts throughout the manuscript that we have used to assess variant
frequencies in non-ASD samples. These cohorts are described below.

PUBLICLY AVAILABLE DATABASES

Annotations from publicly available databases were obtained from ANNOVAR and include: the
1000 genomes project (1000g2015aug_all) (Auton et al., 2015), gnomAD (version 2.0.2)
(Karczewski et al., 2019), the Exome Aggregation Consortium (exac03nonpsych) (Lek et al.,
2016), the NHLBI Exome Sequencing Project (esp6500siv2_all)
(http://evs.gs.washington.edu/EVS/), and allele frequency estimates from whole-genome
sequenced unrelated, healthy subjects (http://www.completegenomics.com/public-data/69-
genomes/, cg46).

**UCLA INTERNAL CONTROLS**
The “UCLA internal controls” referred to throughout this manuscript are comprised a 379 unrelated, whole-genome sequenced samples with *Progressive Supranuclear Palsy* (PSP). ASD and PSP have no known etiological overlap or comorbidity.

*Healthy Non-Phaseable (HNP) samples*

In this manuscript, “HNPs” refers to 1,890 healthy iHART samples for which no biological parents have been sequenced (non-phaseable). These samples likely harbor genetic variants associated with ASD, and so provide allele frequency estimates that are generally more permissive when filtering for inherited risk variants.

**METHOD DETAILS**

**Whole-genome sequencing and data processing**

Whole-genome sequencing was conducted on DNA samples through New York Genome Center (NYCC). Following DNA sample quality and quantity assessment, genotyping was performed using Illumina Infinium Human Exome-12 v1.2 or Infinium Human Core Exome microarrays (San Diego, CA). The genome-wide genotyping data was subsequently used to confirm sample identity and conduct sex checks in in PLINK v1.07 (Purcell et al., 2007). VerifyIDintensity (VII) (Jun et al., 2012) was used to assess contamination; samples with two or more modes of contamination greater than 3% were excluded from sequencing. Samples that passed identity and quality checks were then sequenced at NYGC using the Illumina TruSeq Nano library kits and Illumina's HiSeq X (San Diego, CA) (Ruzzo et al. 2019).
iHART WGS data were processed through the same bioinformatics pipeline, adapted from GATK’s best practices (DePristo et al., 2011; Van der Auwera et al., 2013). The bioinformatics processing steps are outlined in our previous manuscript (Ruzzo et al. 2019), with the only difference being that GATK tools were updated to version 3.4 and Picard MarkDuplicates tool to version 1.83.

**Quality control assessment**

WGS data underwent standard quality control checks to ensure the accuracy of sequencing and variants, as well as sample identity. Checks were conducted on samples for relatedness, duplicates, contamination, sequencing coverage, variant quality (using GATK’s VariantEval module, data not shown), concordance between genotyping chip and WGS data, and concordance between self-declared sex and observed biological sex. A total of 4,551 individuals from 1,004 ASD families, from the Autism Genetic Resource Exchange (AGRE) passed quality control (Table S1).

*Whole-genome sequence coverage*

We calculated genome-wide per-base sequencing coverage for each sample using SAMtools v1.2 (Li et al., 2009). Custom scripts were used to calculate (i) average coverage and (ii) percentage of the genome (excluding gap regions, downloaded from the UCSC table browser) covered at 1X, 10X, 20X, 30X, and 40X. On average across all samples, 99.2% of bases were covered at a depth of ≥10X (Figure S1A-E).

**Variant inheritance classifications**
Variants were classified into one of eight inheritance types: (i) de novo, (ii) paternally inherited, (iii) maternally inherited, (iv) missing, (v) uncertain, (vi) unknown phase, (vii) newly homozygous, or (viii) newly hemizygous. Definitions for newly homozygous, newly hemizygous, unknown phase, and uncertain inheritance types are available in our previous manuscript (Ruzzo et al. 2019). All VQSR failed and multi-allelic variants were excluded before reaching this classification step. Variants were also required to have a read depth of ≥10X, genotype quality of ≥25, and alternative allele reads/total reads ratio of ≥0.2.

**Defining rare inherited and private variants**

Rare inherited variants were defined as SNVs and indels with an allele frequency (AF) ≤0.1% among publicly available databases (1000g2015aug_all, ExACv3.0, cg46, gnomAD), internal controls, and iHART HNPs. These variants were also restricted to those not flagged as low-confidence in the Genome in a Bottle Consortium (GIAB; Zook et al., 2014) and missing in less than or equal to 25% of controls.

Private variants were defined as SNVs and indels observed in only one iHART family (AF~0.05%) and missing in less than or equal to 25% of iHART HNPs. In addition, private variants were required to meet the following criteria: (i) not observed among control cohorts (AF=0), (ii) missing in less than or equal to 25% of PSP controls, (iii) not flagged as low-confidence in the GIAB Consortium. We only analyzed private variants among the 2,254 fully-phaseable children in our study.

**Non-coding analyses**
Definition of non-coding variants

In this manuscript, we defined non-coding SNVs and indels as variants that do not occur within a VEP annotated coding transcript. These include 17 of the 35 VEP consequences:

“3_prime_UTR_variant”, “5_prime_UTR_variant”, “downstream_gene_variant”,
“upstream_gene_variant”, “feature_elongation”, “feature_truncation”, “intergenic_variant”,
“intron_variant”, “mature_miRNA_variant”, “non_coding_transcript_exon_variant”,
"non_coding_transcript_variant”, “regulatory_region_ablation”,
“regulatory_region_amplification”, “regulatory_region_variant”, “TF_binding_site_variant”,
“TFBS_ablation”, “TFBS_amplification” (Ruzzo et al. 2019). To ensure strict filtering of non-coding variants, only the first most damaging consequence associated with a single variant were taken into account if there were multiple annotations available. We only considered non-coding variants that had not been flagged as low-confidence by the GIAB consortium.

Samples included for non-coding analyses

iHART non-coding rare and private inherited analyses included variants from 2,254 fully-phaseable children (N\textsubscript{aff}=1,836, N\textsubscript{unaff}=418). iHART non-coding RDNVs were restricted to 1,442 non-ARC outlier samples (N\textsubscript{aff}=1,164, N\textsubscript{unaff}=279) and variants identified as high confidence by ARC.

Description of non-coding variant sets tested

The non-coding rare inherited, private inherited, and rare \textit{de novo} variant sets described above were further refined before quantitative burden testing, with the aim of increasing power to detect any association with ASD by narrowing variants to those more likely to be functional
and/or contributing to the disorder according to previous studies. The first refined sets of non-coding variants used (Figure 4A-C) consist of those mapping to promoters, defined as both 2kb and 10kb upstream and 1kb downstream (3kb and 11kb promoters, respectively) from the transcription start site (TSS), by referencing the longest transcript for each gene (ties in transcript length were resolved by selecting the lower Ensembl Transcript ID). Rare inherited and private inherited variants mapping to 3kb promoters where further filtered (Figure S4D-E) by keeping those in promoters of genes belonging to ASD risk gene sets identified and defined in this study, the 74 TADA ASD risk genes (with FDR < 0.1 in TADA-mega analysis) and the 152 known ASD risk genes. We also specifically focused on variants mapping to transcription factor binding sites (TFBS) located within 11kb promoters (Figure S4F-H). As single global maps of genomic regions bound by transcription factors (TFs) to use for variant filtering we chose (1) cis-regulatory modules (CRMs) from the whole UniBind database (Gheorghe et al. 2019), and highly reliable (category A) (2) cistromes and (3) cismotifs, as defined in (Vorontsov et al. 2018). Finally, we took advantage of a new co-localization approach for interpreting the functionality of non-coding regions (Xu et al. 2020), which evaluates both their sequence constraint within the human lineage and their tissue- or cell type- specific regulatory role, by keeping variants in both 11kb promoters and across the whole genome that were mapping to the 10-tissue catalogue of brain-specific co-localized regions (http://www.funlda.com/colocalization/region) (Figure S4I-K).

Non-coding variant quantitative burden testing

Quantitative burden of non-coding variants in each refined set was compared between affected and unaffected children (Figure S4A-K) through both regression and other statistical tests (one-
sample proportion test and Poisson rate ratio test) on total variant counts in the two groups. In addition, the relationship between phenotypic group (affected versus unaffected) and sex (female versus male) / sample source of DNA used for whole-genome sequencing (whole blood versus LCL) / carrying a non-coding rare de novo variant (carrier versus non-carrier) was explored by performing independence tests. All these statistical tests (p-values < 0.05 were considered significant) returned consistent and reproducible results. However, regression was the preferred approach for quantitative burden testing, since it gave us the advantage to correct for two important covariates, child sex and sample source of DNA used for whole-genome sequencing. Non-coding variants belonging to each refined set were counted for each child and used to perform regression, together with child phenotypic group, sex, and sample source of DNA. First, we modeled the response variable “variant count per child” by including the predictor variables “child phenotype”, “child sex”, and “sample source of DNA” through Poisson / quasi-Poisson / negative binomial regression models (R stats glm() function and MASS glm.nb() function were used). The negative binomial models were the best model to correct for over-dispersion of the count data. For rare de novo variants and both rare inherited and private inherited variants mapping to 3kb promoters of genes belonging to ASD risk gene sets (Figure S4D-E), we checked for potential zero-inflation of the count data and additionally built zero-inflated Poisson and zero-inflated negative binomial regression models (R pscl zeroinfl() function was used). Even for these variant sets, the negative binomial models were the best to predict zero counts. Alternatively, we modeled the response variable “child phenotype” by including the predictor variables “variant count per child”, “child sex”, and “sample source of DNA” through logistic regression. For all models built, we used (1) the residual deviance to perform a goodness of fit test for the overall model and (2) the difference between the residual deviance for the model with
predictors and for the null model to test the significance of the overall model. These tests confirmed that the logistic regression models were the only models with a good fit and significance for the overall model. For this reason, although all approaches for quantitative burden testing described above returned consistent results, in this paper we chose to report those from logistic regression analyses (Table S2) for which we also computed power (Table S3). ASD odds ratio (OR) values were obtained by exponentiating the estimated regression coefficients.

Power analysis for quantitative burden testing through logistic regression

Power for quantitative burden testing through logistic regression was computed for each tested variant set by simulation. First, we used a posthoc approach, computing power for the observed ASD ORs and at the current sample sizes for all three predictor variables “variant count per child”, “child sex”, and “sample source of DNA”. We adopted three different simulation strategies to generate simulated data to use as new input to rerun logistic regression and estimate power, based on (1) the R simr powerSim() function (Green and MacLeod 2016) and on two custom functions named (2) “sample” and (3) “random”, respectively. The first two strategies are use bootstrapping, which resamples from the existing dataset, while the “random” strategy uses random variable generation and regression coefficients from the logistic model for the three predictor variables to compute the simulated response variable “child phenotype”. Quality of simulations was assessed by visualizing both simulated variables and resulting logistic regression coefficients. Power was estimated as the proportion of significant regression results (p < 0.05) over 1,000 simulations, together with 95% confidence intervals for this value. The three simulation strategies returned very consistent results (Table S3): in this paper we chose to report those obtained with the “sample” custom function strategy (Table S3 and Figures S4L-S4N).
addition, for each tested variant set we built power curves (through strategies 2 and 3, returning reproducible results) specific to the variant count per child predictor, for the observed ASD OR and at specific increasing sample sizes (private inherited: 2000-9000; rare de novo: 1,000-8,000). Again, power was estimated over 1,000 simulations. Curves were fitted by local regression for visualization.

The posthoc power analysis confirmed that the observed ASD ORs for the non-coding variant sets tested are too small to be detected at the current sample sizes, consistently with results from previous analysis of ASD simplex families (Werling et al. 2018) (Table S3). The power curves built for increasing sample sizes showed that a sample size of at least 8,000 fully-phaseable children would give us sufficient power to detect the observed ASD OR for private inherited variants mapping to 11kb promoter regions (at least 9,000 fully-phaseable children for the following private inherited sets: 3kb promoters, cistromes, cismotifs, and brain co-localized regions across the whole genome), whereas a sample size of at least 7,000 fully-phaseable children would allow us to detect the observed ASD OR for rare de novo variants mapping to brain co-localized regions across the whole genome (Figures S4M-N and Table S3 (as reference for ASD ORs and p values for each set)). Notably, the increasing sample sizes tested would not give us sufficient power to detect the observed ASD OR for any of the sets of rare inherited variants (Figure S4L and Table S3).

**Artifact Removal by Classifier (ARC)**

The Artifact Removal by Classifier (ARC) supervised model was designed to separate true rare de novo variants from LCL-specific genetic aberrations or other kinds of artifacts, such as
sequencing and mapping errors (https://github.com/walllab/iHART-ARC). Details on the training, features, and performance evaluation of ARC are provided in our previous manuscript (Ruzzo et al. 2019).

**ARC outlier samples**

When applied to all 2,254 children (partially or fully-phaseable), we found a subset of outlier samples (fully-phaseable n=657) that had an ARC score less than 0.4 for >90% of their raw *de novo* variant calls. We excluded all ARC outlier samples from downstream analyses that involved *de novo* variants unless otherwise mentioned.

**De novo variant rate vs. paternal age**

We analyzed the correlation between paternal age and *de novo* variant rate in 1,157 fully-phaseable, affected iHART children (excluding MZ twins, ARC outliers, and 7 samples without paternal age information). Details on the linear model used in this analysis are available in our previous study (Ruzzo et al. 2019). We observed a robust signal after running ARC (*P*<2.2 x 10^{-16}), but not prior to application of ARC (Figure S3D). According to our study, the RDNV rate increased by 1.02 per year of paternal age (95% CI = 1.02-1.03), consistent with previously published studies (Deciphering Developmental Disorders Study 2017; Francioli et al. 2015; Goldmann et al. 2016; Michaelson et al. 2012).

**Rates for rare *de novo* variants**

In our analysis, we only considered *de novo* variants in our 2,254 fully-phaseable children. Rare *de novo* variants were defined as: (i) absent in all controls, (ii) given an ARC score ≥0.4, (iii)
missing in less than or equal to 25% of controls, (iv) no flagged as low-confidence by the GIAB consortium, (v) not present in an ARC outlier sample (n=657). Due to the use of LCL MZ twins (n=154) as the ARC training set, all MZ twin \textit{de novo} variants were excluded from \textit{de novo} rate calculations. Consequently, all \textit{de novo} variant rate calculations were performed using 1,443 fully-phaseable non-MZ twins and non-ARC outliers (N_{aff}=1,164; N_{unaff}=279).

Our results suggest a mean genome-wide rate of 52.8 RDNVs per child (Figure S3B), which is consistent with previously reported rates (mean=64.4; range 54.8-81) (Besenbacher et al., 2016; Conrad et al., 2011; Kong et al., 2012; Michaelson et al., 2012; Turner et al., 2016; Yuen et al., 2017). We used the Wilcoxon rank sum test to determine if the number of \textit{de novo} variants was significantly different in affected versus unaffected children.

**Comparison of rates of variants in affected and unaffected children**

For rare inherited, rare \textit{de novo} and private inherited variants, we compared the rates (number of variants per child) for missense, synonymous and protein truncating variants in affected versus unaffected children. We used the logistic regression model:

\[
\log \left[ \frac{P(\text{Phenotype} = \text{Affected})}{1 - P(\text{Phenotype} = \text{Affected})} \right] = \alpha + \beta_1(\text{Variants\_per\_child}) + \beta_2(\text{Sex}) + \beta_3(\text{DNA\_source})
\]

We performed this analysis with all genes and those with pLI>0.9 and 0.995.

**TADA mega-analysis**

*Samples and qualifying variants*
In order to combine evidence found in ASD cases from rare de novo (DN) or transmitted (inherited) PTVs as well as de novo Mis3 variants predicted to damage the encoded protein, we used the Transmitted and De novo Association (TADA) test (He et al., 2013). Our TADA analysis was conducted on 1,660 genetically non-identical (one MZ twin retained) ASD cases from 864 iHART families with at least one ASD case and both biological parents sequenced. We treated these 1,660 affected children and their biological parents as independent trios in the TADA analysis. To enhance our ability to identify novel ASD-risk genes, we merged qualifying variants found in ASD cases from our current (iHART) cohort with a previous TADA mega-analysis (Sanders et al., 2015), which included both variants discovered from the Autism Sequencing Consortium (ASC) and Simons Simplex Collection (SSC), as well as small de novo CNV deletions (SmallDel) identified from SSC and Autism Genome Project (AGP) probands (Table S4).

Qualifying variants in our cohort included rare de novo and transmitted PTVs as well as de novo Mis3 variants present in the 1,660 affected samples, not flagged as low-confidence by the GIAB consortium (Zook et al., 2014). As in the previous TADA mega-analysis (Sanders et al., 2015), transmitted PTVs were required to have an AF≤0.1% among public databases (1000g, ESP6500, ExACv3.0, cg46, gnomAD), internal controls, and iHART HNPs. De novo variants were required to be absent (AF=0) across all public databases, internal controls, and iHART HNPs. ARC was run on all non-MZ twins to obtain high confidence de novo variants. De novo variants shared in MZ-twins were also considered qualifying variants and not filtered on their ARC score. For the TADA samples identified as ARC outliers (n=421), their de novo variants were excluded but their transmitted PTVs were retained as qualifying variants. We annotated the de novo Mis3
variants in the iHART cohort using the “probably damaging” PolyPhen-2 (Adzhubei et al., 2010) v2.2.2r395 HumDiv predictions from the Whole Human Exome Sequence Space (WHESS dataset). In cases where several qualifying variants in a gene were found in a sample, only the most damaging variant was retained.

For the TADA analysis, we then calculated the counts for all the qualifying variants from the different cohorts into a gene-by-variant-type matrix, which included variant counts for a total of 18,472 gencodeV19 genes with HGNC approved gene names. Specifically, DN PTV/Mis3 variants counts are identified in 5,275 ASD cases collected from iHART (this manuscript), SSC (Iossifov et al., 2014), and ASC (De Rubeis et al., 2014), while counts for the transmitted and non-transmitted PTVs are compiled from 4,563 ASD cases and 8,359 controls collected from the ASC (De Rubeis et al., 2014) and iHART (this manuscript) (Table S4). In addition, we calculated the counts of DN SmallDel in 4,687 ASD cases from the SSC (Sanders et al., 2015) and the AGP (Pinto et al., 2014) (Table S4).

We note that 119 of the 424 AGRE samples included as “cases” in the original TADA analysis (De Rubeis et al., 2014) with only transmitted PTVs accounted for, were able to have their DN PTV and Mis3 variants accounted for in iHART. This was due to iHART sequencing these samples (or their corresponding monozygotic twin) and their biological parents. The acquisition of these AGRE samples and the method used to count their DN PTV/Mis3 variants and transmitted PTVs are outlined in our previous manuscript (Ruzzo et al. 2019).

*TADA parameters*
The TADA parameters used in this study, match those used in our previous TADA mega-analysis (Ruzzo et al. 2019).

**Novel gene discovery**

For a gene to be considered novel, we applied strict parameters - it had to have an FDR <0.1 in our TADA mega-analysis and lack genome-wide statistical support across all previous studies (Sanders et al. 2015; Satterstrom et al. 2019; De Rubeis et al. 2014) with statistical rigor.

**Genes with large inherited PTVs contribution**

Given our signal for rare inherited variants, we sought to highlight genes for which a large contribution of the TADA ASD-risk association signal is derived from inherited PTVs. Conservatively, we considered only variants where the inheritance was known (de novo vs. inherited). Therefore, we adjusted the total number of TADA-qualifying variants to ignore PTVs from cases because some of the TADA-mega analysis qualifying variants originate from case-control studies (not iHART) where inheritance is unknown. We applied two methods to identify genes with a large contribution from inherited PTVs. The total number of qualifying variants (N) in each TADA gene was defined as $N_{DN,PTV} + N_{DN,SmallDel} + N_{DN,Mis3} + N_{Inherited,PTV}$. We defined the percentage of inherited component of a gene as $(N_{Inherited,PTV}/N) \times 100$.

**Fetal Single Cell Expression of iHART TADA analysis**

Expression of iHART genes were examined in previously published single cell human developing neocortex, single cell fetal dataset (Polioudakis et al. 2019). The mean expression and percent expressed were calculated within each cell type and hierarchically clustered using
**Developmental Trajectory Analyses**

Development trajectory RNAseq data was used from previous studies including the long term maturation human cortical spheroid (hCS) model (Gordon et al. 2021) and the BrainSpan dataset (Stein et al. 2014; Kang et al. 2011; M. Li et al. 2018). Briefly, hCS samples included six human induced pluripotent stem cell (hiPSC) lines derived from five different individuals that were cultured for up to 694 days in vitro. In total, 62 samples were collected for RNA sequencing (from four individuals, five hiPSC lines) at 13 timepoints. RNAseq libraries were prepared using Truseq stranded RNA RiboZero Gold (Illumina) and were sequenced using 100-bp paired end reads on an Illumina HiSeq 4000. Reads were then mapped to hg38 with Gencode v.25 annotations using STAR (v.2.5.2b) (Dobin et al. 2013). Gene expression levels were quantified using RSEM (v.1.3.0) (B. Li and Dewey 2011). Genes with low levels of expression (less than ten reads in more than 20% of the samples) were removed from the analysis. Outliers were then
removed using standardized sample network connectivity (Z scores smaller than −3) (Oldham, Langfelder, and Horvath 2012). To quantify the technical variation in the RNA sequencing, we calculated the first five PCs of the Picard sequencing metrics

(http://broadinstitute.github.io/picard/; v.2.5.0). These PCs, referred to as seqPC1–seqPC5, were then included in the linear model.

To help control for variability between the individuals racial background, we used the GATK (v.3.3) haplotype caller to call single nucleotide polymorphism (SNPs) from the aligned reads (McKenna et al. 2010). We filtered for sites with missing genotypes (>5%), rare minor allele frequency (<0.05) and out of Hardy–Weinberg equilibrium (<1×10−6) (Purcell et al. 2007). Genetic ancestry was inferred by running multidimensional scaling (MDS) on these high-quality SNPs together with HapMap3.3 (hg38). The first two MDS values, referred to as ancestryPC1/2, were then included in our linear model. For principal component analysis (PCA), as well as to visualize single gene trajectories, gene expression was normalized using CQN (without quantile normalization, sqn = FALSE) (v.1.28.0) and ancestryPC1-2 and SeqPC1-5 were regressed out before batch correction using Combat (Leek et al. 2012) from the sva package (v.3.30.0) in R. Single gene trajectories trends lines were fitted using the loess method (Cleveland and Devlin 1988) from the ggplot2 package (Wickham 2009, 2) in R.

In the BrainSpan RNA sequencing data (M. Li et al. 2018) to quantify gene expression at each developmental stage, the cortical samples were aligned to hg38 using Gencode v.25 annotations via STAR (Dobin et al. 2013). Gene expression was then quantified using the union exon model in featureCounts (Liao, Smyth, and Shi 2014). We removed low quality samples in which the
RNA integrity number (RIN) was lower than 8, there were less than 25% coding bases or ribosomal bases made up more than 25% of total bases (as called by Picard tools). Genes with low levels of expression (less than ten mapped reads in more than 80% of the samples) in a given developmental stage were removed. We retained 196 samples from 24 individuals (9 female and 15 male).

We clustered the scaled normalized expression of hCS genes using hierarchical clustering on the Euclidean distance between genes. The most representative genes of a clustered were determined by each genes correlation with the cluster eigengene. Genes were clustered by their expression in the hCS using hierarchal clustering on the Euclidean distance between the genes. Cluster eigengenes were calculated using the module Eigengenes function from the WGCNA package (Langfelder and Horvath 2008). The gene in each cluster were correlated to the cluster module eigengene and the top five genes were annotated on the heatmap. The average expression per time point was calculated and trajectories trends lines were fitted using the loess method (Cleveland and Devlin 1988) from the ggplot2 package (Wickham 2009, 2). GO terms enrichment was performed using the enrichGO function from the clusterProfiler package (Yu et al. 2012) (v.3.12.0). Enrichment was performed on biological process and molecular function GO terms. All genes expressed in the hCS were used as background.

**Known ASD Risk Genes (KARGs) Analysis**

A comprehensive list of 152 known ASD-risk genes (KARG) was generated by merging unique genes, with HGNC approved gene names, identified at an FDR threshold of <0.1 in the iHART mega-analysis and previous TADA mega-analyses (Ruzzo et al., 2019; Sanders et al., 2015;
The same rare qualifying variants as TADA (DN/transmitted PTV and rare DN Mis3) were considered. Transmitted PTVs in ASD-risk genes were identified in the 2,254 fully-phaseable children (N_{aff}=1,836, N_{unaff}=418), while DN PTV or Mis3 qualifying variants in ASD-risk genes were restricted to those identified in the 1,443 non-ARC outlier samples (N_{aff}=1,164, N_{unaff}=279). We required transmitted PTVs to be in non-GIAB regions and have an AF≤0.1% in public databases (1000g, ESP6500, ExACv3.0, cg46, gnomAD), internal controls, and iHART HNP samples. DN variants identified in the iHART cohort were required to have an ARC score ≥0.4 and be in non-GIAB regions as well be absent in all public databases, internal controls, and HNP samples (AF = 0).

The list of risk genes for ASD and developmental disorders (DDD) was generated by merging unique genes, with HGNC approved gene names from known ASD risk genes and the Deciphering Development Disorders study (Deciphering Developmental Disorders Study 2017) with a FDR threshold of <0.1. This gene set totaled 208 genes (Table S5).

### Polygenic Transmission Disequilibrium Test (pTDT) Analysis

#### Polygenic Risk Scoring

A polygenic risk score (PRS) was calculated to provide a quantitative estimate for an offspring’s genetic predisposition for ASD, SCZ, BD, and EA in iHART. To generate PRS, 3,255 samples of European ancestry, identified by multidimensional scaling (MDS)-derived hard race calls, were used (N_{aff}=1,492, N_{unaff}=1,763). For each disorder/phenotype tested, we obtained summary statistics from a GWAS with similar ancestry: for ASD we used the ASD iPSYCH GWAS (iHART overlaps with the PGC (Psychiatric Genomics Consortium) ASD GWAS; N_{aff}=8,605, N_{unaff}=19,526) (Gandal et al. 2018); for SCZ we chose the SCZ CLOZUK+PGC meta-analysis
GWAS (Naff= 40,675, Nunaff=64,643) (Pardiñas et al. 2018); for BD we used the BD PGC GWAS (Naff= 20,129, Nunaff=21,524) (Ruderfer et al. 2018); for EA (number of years of schooling completed) we chose the SSGAC (Social Science Genetic Association Consortium) EA meta-analysis GWAS (N= 293,723) (Okbay et al. 2016). Variants of interest were those in each GWAS with MAF≤0.05, mind≤0.1, geno=0, hwe≤1e-6. Candidate variant weights were then generated by running LDpred (Vilhjálmsson et al. 2015). PRS models corresponding to increasing fractions of contributing causal SNPs (0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1, Inf) were computed. Then, we evaluated the produced PRS models for each disorder/trait for their prediction of ASD affectation, using sex and the first two genotype principal components, PC1 and PC2, as model covariates. Sex is a confound of ASD affectation status. We removed (regressed out) PC1 and PC2 for the final evaluation of each PRS association with ASD affectation. In order to choose the best model for each disorder/trait, we examined both its Nagelkerke’s R2 and its regression coefficient for ASD affectation. Consistency and convergence for both measures across different p-value thresholds were observed at the p=1 one (assuming all SNPs are causal) that we chose uniformly for all phenotypes of interest. We observe a significant difference in ASD PRS between the tested iHART affected and unaffected subjects (p=0.01, beta=0.01, difference (vs. PRS null model) in Nagelkerke’s R2=0.003).

\textit{pTDT}

We used the polygenic transmission disequilibrium test (pTDT) (Weiner et al. 2017) to determine whether an offspring’s genetic predisposition for ASD is consistent with what is expected based on parental transmission of polygenic risk for ASD. We first calculated the mid-
parent polygenic risk score (PRS<sub>MP</sub>) for the 1,519 children of European ancestry, using the following equation:

\[
P_{RS_{MP}} = \frac{PR_{S_{mother}} + PR_{S_{father}}}{2}
\]

The PRS<sub>MP</sub> was then used to compute the pTDT deviation for each child. The pTDT deviation is defined as follows:

\[
p_{TDT}^{deviation} = \frac{PR_{S_C} - PR_{S_{MP}}}{SD(PR_{S_{MP}})}
\]

where PRS<sub>C</sub> is the polygenic risk score for the affected or unaffected child, and SD(PR<sub>S<sub>MP</sub></sub>) is the standard deviation of the sample-specific mid-parent PRS. In order to determine if the pTDT deviation was significantly different from zero, the two-sided, one-sample t-test was conducted using the pTDT test statistic (t<sub>pTDT</sub>):

\[
t_{pTDT} = \frac{mean(pTDT \text{ deviation})}{SD(pTDT \text{ deviation})\sqrt{n}}
\]

where n is the number of families included in the pTDT study. The same strategy was used for testing over-transmission of polygenic risk for SCZ, BD, and EA in iHART.

*pTDT Variant Analysis*

Carriers for variants in the 74 genes identified at an FDR threshold of <0.1 in the current iHART mega-analysis were considered in the pTDT analysis. The same rare qualifying variants as TADA (DN/transmitted PTV and rare DN Mis3) were considered. Transmitted PTVs in ASD-risk genes were identified in the 1,519 European children with both biological parents sequenced (fully-phaseable according to PRS calculation) who do not identify as a parent in the cohort as
well (\(N_{\text{aff}}=1,231, N_{\text{unaff}}=288\)), while DN PTV or Mis3 qualifying variants in ASD-risk genes were restricted to those identified in the 1,066 non-ARC outlier samples (\(N_{\text{aff}}=867, N_{\text{unaff}}=199\)). We required transmitted PTVs to be in non-GIAB regions and have an AF\(\leq0.1\%\) in public databases (1000g, ESP6500, ExACv3.0, cg46, gnomAD), internal controls, and iHART HNP samples. DN variants identified in the iHART cohort were required to have an ARC score \(\geq0.4\) and be in non-GIAB regions as well be absent in all public databases, internal controls, and HNP samples (AF = 0).

**Phenotypic comparisons among children in iHART**

We used phenotypic data available for iHART children to (1) investigate the phenotypic effect of rare inherited (RI) and rare *de novo* (RDN) variants in the 152 known ASD risk genes (KARG) and to (2) identify potential links between polygenic risk (PGR) over-transmission and phenotype severity through stratified pTDT based on children that do not carry any rare variant in the 74 TADA genes.

For the former analysis, we retrieved data available for the 2,254 fully-phaseable children whereas, for the latter one, we used those with phenotype available for the 1,519 fully-phaseable children of European ancestry.

We tested phenotypic measures descriptive of general intelligence, motor development, language development, and social behavior. The selected measures were verbal and non-verbal IQ scores, age of walking (AOW), ages of first word and phrase, and scores representing social behavior and social interaction skills. We used the standard scores from PPVT-3 (Peabody Picture
Vocabulary Test, 3rd edition) (Dunn and Dunn 1997) and the Raven test (Raven and Raven 2003) as measures of verbal and non-verbal IQ, respectively. Age of walking and ages of first word and phrases in months were retrieved from ADI-R (Autism Diagnostic Interview-Revised, 2003 version with 93 items) (Rutter, Le Couteur, and Lord 2013). We made use of total raw scores and T-scores (with scoring by child’s sex) from SRS (Social Responsiveness Scale) (Constantino and Gruber 2005) as measures of deficiencies in social behavior.

For each measure, we used (1) the corresponding standard hypothesized value to which we compared the phenotypic distribution median value of each group of affected children (stratified by those carrying RI or RDN variants in KARG and those without such variants) and (2) cut-off value(s) for stratified pTDT analysis of affected TADA ASD risk gene non-carriers. IQ scores are normally distributed with a mean of 100 and a standard deviation of 15. We used the median value of 100 as the standard hypothesized value and a cut-off value of 70 (two standard deviations below the mean) to stratify children by those with (IQ < 70) or without (IQ >= 70) cognitive impairment.

To represent normal variation in age of walking, we used the window of achievement for the “walking alone” milestone observed in the normative cohort of the WHO Multicentre Growth Reference Study (MGRS) (Onis 2006). Specifically, the 50th percentile estimate of 12 months derived from this window was used as the standard hypothesized value, whereas the 97th percentile estimate of 16 months was used to divide children in those with atypical (AOW >= 16) or typical (AOW < 16) motor development.
Unaffected children from iHART were assessed for their language development through the AGRE (Lajonchere and AGRE Consortium 2010) custom seven-item language questionnaire. For both language measures, ages of first word and phrase, we used the available data to build “control non-carrier” distributions representing normal variation in milestone achievement among the unaffected siblings. We used the median values of 12 months and 18 months as reference hypothesized values for age of first word and age of first phrase, respectively. Both ADI-R items, age of first word and age of first phrase, are scored in section D of the diagnostic algorithm, with a score of 1 indicating that the age of milestone achievement suggests that developmental concerns are present at or before 36 months and a score of 0 denoting that the age of milestone achievement falls within the range of normal variation. We used this scoring system to stratify children by those with atypical (score = 1) or typical (score = 0) language development.

Total raw scores from SRS were available for both affected and unaffected children in iHART. We used the scores available for the unaffected non-carriers to build a reference distribution representing normal variation in social behavior and social interaction skills. We used the median score of 19 derived from this distribution as reference hypothesized value for testing. We computed corresponding T-scores using means, standard deviations, and sample sizes from the child’s sex-based school-age norm tables reported in the SRS manual (Constantino and Gruber 2012). We used the following SRS T-score cut-off ranges to divide the affected non-carriers into groups showing different degrees of social impairment: T-scores equal to 59 or less were considered as normative, those among 60 and 75 as representative of mild ASD, and those equal to 76 or more as characteristic of severe ASD (Lyall et al. 2014).
When available, we compared redundant measures for the same phenotype derived from different assessment tools to check general quality and concordance of the phenotypic data (not shown).

When data were collected using different version forms for the same assessment tool, we only used those coming from the most frequently used version form to ensure that form version did not drive potential phenotypic differences. An exception was made for SRS, for which we combined the compatible preschool and school-age forms.

When data for the same child were collected more than once, in more than one visit, we kept only those from the most recent visit for further analysis. Again, a different criterium was used for SRS. We removed data from assessments flagged as invalid. When valid SRS data for the same child were collected in more than one visit, (1) if both a parent and a teacher were available as corresponding respondents, the teacher-based assessment was preferred as more reliable, whereas (2) if the same respondent contributed to longitudinal assessments, the earliest visit data was used for downstream analysis (Lowe et al. 2015). In addition, since SRS has not been validated in non-verbal children, we used the ADI-R item age of first word to identify children that had not reached or regained this milestone and we removed their entries from the SRS dataset (Lowe et al. 2015). Finally, we computed sex-based T-scores only for affected non-carriers with ages 4 to 18 years (school-age sample).

Many children in the iHART cohort were still not able to walk or were non-verbal at the time of testing. We wanted to still keep these informative observations for the analysis despite not
having a numeric value for those cases. Therefore, we converted the codes for “milestone not reached” and “milestone lost and not regained yet” into the age of those children at the time of testing plus one month, optimistically assuming for them a potential milestone achievement soon after the latest assessment. We verified that this approach did not remarkably alter the distributions observed for the three items through comparison with those built from the smaller dataset in which the code cases had been simply removed (not shown). Similarly, with the aim of not losing relevant information coming from our Raven data, we converted the codes for “below age-specific test norms” and “above age-specific test norms” non-verbal IQ scores into the cohort lowest minus one point and highest plus one point scores, respectively. There were two cases of affected “double carrier” of both a RDN and a RI variant in KARG and a case of affected “double carrier” of variants in the TADA ASD risk genes. We decided to group these children as RDN carriers, assuming a larger effect size for this rare genetic variation type.

We excluded some specific groups of children from these phenotypic comparison analyses: affected children with confirmed motor development impairment (cerebral palsy diagnosis), affected children with seizures, unaffected children with learning disabilities, unaffected children with language disabilities, unaffected children with previous and no longer confirmed ASD diagnosis, and unaffected children carrying qualifying rare variants in KARG and in the TADA ASD risk genes. These children were removed with the idea of minimizing the potential confounding effect of comorbid phenotypes and ASD-associated genetic variation on the measures tested. Even though we had IQ scores available for very few unaffected children we decided to remove them from the analyses, keeping them only on affected children. We only used data from unaffected children to build the ADI-R item and SRS total raw score “control
non-carrier” distributions mentioned above.

Since the phenotypic measure distributions were not normal and asymmetric, we tested potential differences between their median values and standard hypothesized values through two-sided one-sample sign tests. The three tests performed for each phenotypic measure were considered as simultaneous multiple tests and we used the Benjamini and Hochberg (BH) procedure as FDR (false discovery rate)-controlling method. An adjusted p-value < 0.05 was considered significant. For the phenotypic measure-stratified pTDT analysis, we compared the pTDT deviation distribution mean values against the expected value of 0 through two-sided one-sample t-tests. We defined a nominal p-value < 0.05 as significant.

QUANTIFICATION AND STATISTICAL ANALYSIS
Statistics were calculated using R (3.5.1) unless otherwise notes. The significance of PPI networks were evaluated using DAPPLE metrics based on 1,000 permutations (within the DAPPLE parameter); P values > 0.05 were considered significant.

TADA
The sample sizes for the TADA-mega analysis are provided in Table S4. After performing the Benjamini-Hochberg correction on the TADA results, q-values (False Discovery Rate (FDR)) <0.1 were considered significantly associated with ASD. We find 74 genome-wide significant genes when we apply the field standard FDR<0.1.

DATA AND CODE AVAILABILITY
Following request and approval of the data use agreement found at http://www.ihart.org, the whole-genome sequencing data generated during this study can be downloaded from the Autism Research and Technology Initiative (iHART) of the Hartwell Foundation. Autism Speaks and AGRE must approve the release of whole-genome sequencing data generated during this study.

We provide the source code and a full tutorial for ARC (Artifact Removal by Classifier), a random forest supervised model developed to distinguish rare, \textit{de novo} variants from LCL-specific genetic aberrations or other types of artifacts such as sequencing and mapping errors, at https://github.com/walllab/iHART-ARC.
References


Purcell, Shaun, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A. R. Ferreira, David Bender, Julian Mall, et al. 2007. “PLINK: A Tool Set for Whole-Genome


1,004 multiplex ASD families

Fig 1

30x whole-genome sequencing

Joint Called VCFs
Individuals: 4,610
Cases: 2,184
Fully Phaseable Families: 1,004
Variants: 81,175,025

Variant QC
Exclude VQSR failed and multi-allelic variants

Sample QC
Exclude duplicates, sex mis-match, IBD fail, non-phaseable children

Genotype QC
Filter for DP, GQ, AD and non-GIAB regions
Individuals: 2,254
Cases: 1,836
Fully Phaseable Families: 865
Variants: 60,844,051

Rare De Novo  Rare Inherited  Private Inherited

Identify Risk Genes

Investigate Genetic Architecture and Phenotype Associations

Known ASD Risk Genes
Pathway and Cell-type Enrichment
Identify Genes Supported by Rare Inherited Risk
Integrate Common with Rare Risk

-log10(q)

Rare De Novo  Rare Inherited

Expected Mid-parent

Common Polygenic Transmission (pTDT)
A Rare De Novo

B Rare De Novo in pLI ≥ 0.9 Genes

C Rare De Novo in pLI ≥ 0.995 Genes

D Rare Inherited

E Rare Inherited in pLI ≥ 0.9 Genes

F Rare Inherited in pLI ≥ 0.995 Genes

G Private Inherited

H Private Inherited in pLI ≥ 0.9 Genes

I Private Inherited in pLI ≥ 0.995 Genes
A 2254 children with fully phaseable parents (1004 ASD multiplex families)

418 unaffected
101 ARC outlier
75 monozygotic twin

TADA Mega-analysis

B Proportion of inherited PTVs

-\log_{10}(q-value)

Proportion of Inherited PTVs

No Inherited Component (46%)
Inherited Component <50% (27%)
Inherited Component >50% (27%)

C

-\log_{10}(q-value)

Proportion of Inherited PTVs

D

-\log_{10}(q-value)

Proportion of Inherited PTVs

E

Biological Pathway

learning or memory
modulation chemical synaptic transmission
social behavior
long-term synaptic depression
vocalization behavior

Additional Datasets

De novo PTV/Mis3
2,591 ASD in SSC (Iossifov et al. 2014)
1,445 ASD in ASC (De Rubeis et al. 2014)

De novo SmallDel
2,591 ASD in SSC (Sanders et al. 2015)
2,096 ASD in AGP (Pinto et al. 2014)

Transmitted and non-transmitted PTV
3,046 ASD + 6,842 Control in ASC (De Rubeis et al. 2014)
iHART family A
CAPN12
FBXL13

iHART family B
MFRP
SHANK3

iHART family C
RANBP17
WDFY3

- Not sequenced individual
Joint Called VCFs

- Samples: 4,610
- Cases: 2,184
- Variants: 81,175,025

**Variant QC1**
- Exclude VQSR failed variants
- Exclude multi-allelic variants

**Sample QC2**
- Exclude duplicate samples
- Exclude imputed sex and reported sex mismatch
- Exclude chip-based IBD fail samples
- Exclude non-phaseable children

**Genotype QC**
- DP ≥ 10
- GQ ≥ 25
- AD ≥ 0
- ALT AD/DP ≥ 0.2

**Variant/Sample QC3**
- iHART HNP percent missing ≤ 25%
- PSP percent missing ≤ 25%
- Not flagged in GIAB
- Fully-phaseable children

**Rare De Novo**
- ARC pass children
- ARC ≥ 0.4
- AF = 0 in 1000 Genomes Project, ESP6500, ExAC, PSP Controls, HNPs, and gnomAD
- Variants: 669 (coding)
  - Samples: 1,443 (1,164 Affected, 279 Unaffected)
  - 537 in Affected (0.46 mean variants/child)
  - 132 in Unaffected (0.47 mean variants/child)

**Rare Inherited**
- AF ≤ 0.001 in 1000 Genomes Project, ESP6500, ExAC, PSP Controls, HNPs, and gnomAD
- Variants: 327,896 (coding)
  - Samples: 2,254 (1,836 Affected, 418 Unaffected)
  - 266,770 in Affected (145.30 mean/child)
  - 61,126 in Unaffected (146.23 mean/child)

**Private Inherited**
- Variants that are observed only one iHART family
- AF = 0 in 1000 Genomes Project, ESP6500, ExAC, PSP Controls, and gnomAD
- Variants: 72,999 (coding)
  - Samples: 2,254 (1,836 Affected, 418 Unaffected)
  - 59,157 in Affected (32.22 mean/child)
  - 13,842 in Unaffected (33.11 mean/child)
A

B

C

D

The copyright holder for this preprint which was not certified by peer review is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.
2254 children with fully phaseable parents from 1004 ASD multiplex families: WGS

101 ARC outlier affected samples without rare inherited variants excluded out of 522 ARC outlier affected samples

1836 affected with fully phaseable parents
  Define:
  ~Rare inherited variants: <0.1% public/internal dataset, not in GIAB
  ~Rare de novo variants: absent in all controls, ARC score > 0.4, not in GIAB

1735 affected with fully phaseable parents
  - 421 ARC outlier samples with rare inherited variants (transmitted TADA analysis only)
  - 75 affected monozygotic twins excluded out of 153 affected monozygotic twins

1660 affected with fully phaseable parents
  - 421 considered only transmitted PTV (ARC outliers with inherited PTV/Mis3)
  - 143 had only de novo PTV/Mis3
  - 1096 had de novo PTV/Mis3 and transmitted PTV

Additional Datasets

<table>
<thead>
<tr>
<th>Qualifying variants</th>
<th>Cohorts</th>
</tr>
</thead>
<tbody>
<tr>
<td>De novo PTV/Mis3</td>
<td>2,591 ASD in SSC (Iossifov et al. 2014)</td>
</tr>
<tr>
<td></td>
<td>1,445 ASD in ASC (De Rubeis et al. 2014)</td>
</tr>
<tr>
<td>De novo SmallDel</td>
<td>2,591 ASD in SSC (Sanders et al. 2015)</td>
</tr>
<tr>
<td></td>
<td>2,096 ASD in AGP (Pinto et al. 2014)</td>
</tr>
<tr>
<td>Transmitted and non-</td>
<td>3,046 ASD + 6,842 Control in ASC (De Rubeis</td>
</tr>
<tr>
<td>transmitted PTV</td>
<td>et al. 2014)</td>
</tr>
</tbody>
</table>

TADA Mega-analysis

B

All Genes

Fraction of Subjects per Phenotype

p=0.486

Phenotype

- Affected
- Unaffected