Gupta et al. Molecular Autism (2017) 8:19
DOI 10.1186/s13229-017-0133-0

Molecular Autism

RESEARCH Open Access

Neurogenetic analysis of childhood disintegrative disorder

Abha R. Gupta[1,2*†], Alexander Westphal[2,3†], Daniel Y. J. Yang[2], Catherine A. W. Sullivan[1], Jeffrey Eilbott[2], Samir Zaidi[4], Avery Voos[2], Brent C. Vander Wyk[2], Pam Ventola[2], Zainulabedin Waqar[1], Thomas V. Fernandez[2,3], A. Gulhan Ercan-Sencicek[2], Michael F. Walker[2], Murim Choi[4], Allison Schneider[2], Tammy Hedderly[5], Gillian Baird[5], Hannah Friedman[2], Cara Cordeaux[2], Alexandra Ristow[2], Frederick Shic[2], Fred R. Volkmar[2] and Kevin A. Pelphrey[2]

Abstract

Background: Childhood disintegrative disorder (CDD) is a rare form of autism spectrum disorder (ASD) of unknown etiology. It is characterized by late-onset regression leading to significant intellectual disability (ID) and severe autism. Although there are phenotypic differences between CDD and other forms of ASD, it is unclear if there are neurobiological differences.

Methods: We pursued a multidisciplinary study of CDD ($n = 17$) and three comparison groups: low-functioning ASD ($n = 12$), high-functioning ASD ($n = 50$), and typically developing ($n = 26$) individuals. We performed whole-exome sequencing (WES), copy number variant (CNV), and gene expression analyses of CDD and, on subsets of each cohort, non-sedated functional magnetic resonance imaging (fMRI) while viewing socioemotional (faces) and non-socioemotional (houses) stimuli and eye tracking while viewing emotional faces.

Results: We observed potential differences between CDD and other forms of ASD. WES and CNV analyses identified one or more rare de novo, homozygous, and/or hemizygous (mother-to-son transmission on chrX) variants for most probands that were not shared by unaffected sibling controls. There were no clearly deleterious variants or highly recurrent candidate genes. Candidate genes that were found to be most conserved at variant position and most intolerant of variation, such as *TRRAP*, *ZNF236*, and *KIAA2018*, play a role or may be involved in transcription. Using the human BrainSpan transcriptome dataset, CDD candidate genes were found to be more highly expressed in non-neocortical regions than neocortical regions. This expression profile was similar to that of an independent cohort of ASD probands with regression. The non-neocortical regions overlapped with those identified by fMRI as abnormally hyperactive in response to viewing faces, such as the thalamus, cerebellum, caudate, and hippocampus. Eye-tracking analysis showed that, among individuals with ASD, subjects with CDD focused on eyes the most when shown pictures of faces.

Conclusions: Given that cohort sizes were limited by the rarity of CDD, and the challenges of conducting non-sedated fMRI and eye tracking in subjects with ASD and significant ID, this is an exploratory study designed to investigate the neurobiological features of CDD. In addition to reporting the first multimodal analysis of CDD, a combination of fMRI and eye-tracking analyses are being presented for the first time for low-functioning individuals with ASD. Our results suggest differences between CDD and other forms of ASD on the neurobiological as well as clinical level.

Keywords: Autism spectrum disorder (ASD), Childhood disintegrative disorder (CDD), Regression, Intellectual disability (ID), Genetics, Functional magnetic resonance imaging (fMRI), Eye tracking

* Correspondence: abha.gupta@yale.edu
†Equal contributors
[1]Department of Pediatrics, Yale School of Medicine, New Haven, Connecticut, USA
[2]Child Study Center, Yale School of Medicine, New Haven, Connecticut, USA
Full list of author information is available at the end of the article

Background

Autism spectrum disorder (ASD) is defined by deficits in social communication and interaction and restricted, repetitive patterns of behavior, interests, or activities [1]. Decades before Kanner published his landmark paper describing autism [2], Heller reported on six normally developing children who experienced a severe regression in skills between 3 and 4 years of age leading to global impairments, including autistic features [3, 4]. Heller termed the condition *dementia infantilis*, which was included in the ICD-10 [5] and DSM-IV [6] as childhood disintegrative disorder (CDD). CDD was defined by normal development for at least the first 2 years of life followed by regression before age 10 years in at least two of the following areas: (1) expressive or receptive language, (2) social skills or adaptive behavior, (3) bowel or bladder control, (4) play, and (5) motor skills. There has been much debate as to whether CDD is a late-onset variant of autism or a distinct entity [7, 8]. CDD was subsumed by the diagnosis ASD in the DSM-5 [1], since there was little scientific basis for including CDD as a separate disorder [9].

There are, however, important phenotypic differences between CDD and other forms of ASD [8, 10–15]. While symptoms of ASD are usually recognized by 2 years of age, the onset of symptoms in CDD is usually between 3 and 4 years of age. While approximately a third of children with ASD experience a regression in skills, again usually by age 2 years [16], CDD is defined by regression, which is characteristically of later onset, more global in extent, and more severe in degree. Indeed, children with CDD generally have the poorest outcome among individuals with ASD, usually with severe loss of cognitive and communication skills [8, 11]. In contrast to CDD, children who are diagnosed with ASD later than the typical age range tend to be higher functioning, leading to the delay in diagnosis, and early subtle abnormalities are often noted in retrospect [10, 12]. The majority of children with CDD experience a distinct prodrome characterized by bouts of anxiety and terror [3, 4, 8, 17]. No consistent medical, environmental, or psychosocial triggers have been associated with CDD [8].

Our overarching question is whether there are neurobiological features that distinguish CDD from other forms of ASD. The genetic basis, neuroimaging abnormalities, and social phenotype of ASD are being intensively studied, but no similarly comprehensive studies have been published examining CDD for two important reasons. First, CDD is rare. While the prevalence of ASD is reported to be 1/68 [18], the prevalence of CDD is estimated to be 1–2/100,000 [19]. Second, conducting experimental protocols such as non-sedated functional magnetic resonance imaging (fMRI) and eye tracking with low-functioning subjects is extremely challenging. To study CDD, we used a multidisciplinary approach encompassing: (1) expert clinical characterization; (2) the identification of candidate genes and gene expression analysis; (3) an analysis of brain function, via fMRI, in response to viewing socioemotional (fearful faces) and non-socioemotional (houses) stimuli; and (4) the precise quantification of the social behavioral phenotype using eye tracking. This study is novel not only in examining the neurobiological features of CDD but also in obtaining a combination of reliable non-sedated fMRI and eye-tracking data from low-functioning individuals on the autism spectrum.

Methods

A detailed description of all methods can be found in Additional file 1: Supplementary information. We studied four cohorts: (1) subjects with CDD ($n = 17$, Table 1), (2) low-functioning [full-scale IQ (FSIQ) ≤ 75] subjects with ASD (LFASD, $n = 12$) and early-onset delays (<2 years old), (3) high-functioning (FSIQ ≥ 75) subjects with ASD (HFASD, $n = 50$) and early-onset delays (<2 years old), and (4) typically developing subjects (TD, $n = 26$). The genetics analysis focused on the CDD cohort whereas the fMRI and eye-tracking analyses included subsets of each cohort. We performed whole-exome sequencing (WES) and copy number variant (CNV) analyses of 15 families affected by CDD, which included 15 probands, 13 unaffected sibling controls, and their parents (Additional file 2: Table S1), to identify three types of rare [novel or found at most once across 1000 Genomes (May 2011 release), NHLBI GO ESP Exome Variant Server (ESP6500SI-V2), and in-house database of 2500 exomes] protein-changing variants: (1) de novo, (2) homozygous, and (3) hemizygous (mother-to-son transmission on chrX). We included one additional category for family CDD17 since the father and paternal grandfather reportedly have high-functioning autism: paternally inherited likely gene-disrupting (LGD) variants (premature stop codon, splice site disruption, deletion). We used the human BrainSpan exon-array transcriptome dataset [20] to plot the brain expression profile of CDD candidate genes and conduct co-expression analysis. To study neural systems, we used non-sedated fMRI, and a blocked design involving the presentation of grayscale fearful face (NimStim set of facial expressions) [21] and house (lab database) images to determine brain activation patterns across the four cohorts [CDD ($n = 7$), LFASD ($n = 7$), HFASD ($n = 14$), and TD ($n = 19$)]. To quantify the social phenotype of our four cohorts [CDD ($n = 5$), LFASD ($n = 7$), HFASD ($n = 32$), and TD ($n = 14$)], we collected eye-tracking data as they viewed emotional faces [21].

Table 1 CDD cohort: clinical characteristics and the modalities by which each proband was studied

Subject	Sex	Age at onset (months)	Anxiety prodrome	Multiple regressions	Length of regression (months)[a]	Level of ID	IQ	Cognitive/ developmental test[b]	Language loss	Social or adaptive loss	Incontinence	Play loss	Motor skills loss	Seizures	1^0 relative with ASD	Genetics	fMRI	ET	Age at fMRI, ET (months)
CDD03-03	M	30	No	No	5	Profound	18	Mullen	Yes	Yes	Yes	Yes	Yes	No	No	Yes	No	No	n/a
CDD05-03	M	37	Yes	No	9	Severe	16	Mullen	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	176
CDD07-03	M	45	Yes	No	4	Profound	9	Mullen	Yes	Yes	Yes	Yes	Yes	No	No	Yes	No	No	n/a
CDD08-03	M	80	Yes	Yes	3	Moderate	46	DAS	Yes	Yes	No	Yes	Yes	No	No	Yes	Yes	Yes	139, 154
CDD09-03	M	44	No	No	6	Mild	64	DAS	Yes	Yes	No	Yes	No	No	No	Yes	Yes	No	52
CDD10-03	M	38	Yes	No	3	Profound	19	Mullen	Yes	Yes	Yes	Yes	Yes	No	No	Yes	No	Yes	122
CDD11-03	M	40	Yes	No	4	Profound	8	Mullen	Yes	Yes	No	Yes	Yes	No	No	Yes	Yes	No	206
CDD12-03	F	36	Yes	No	3	Severe	33	Mullen	Yes	Yes	Yes	Yes	Yes	No	No	Yes	No	Yes	87
CDD13-03	F	31	No	No	80	Profound	18	Mullen	Yes	Yes	Yes	Yes	Yes	No	MZ twin (CDD)[c]	Yes	Yes	Yes	92
CDD13-04	F	28	No	Yes	4	Profound	16	Mullen	Yes	Yes	Yes	Yes	Yes	No	MZ twin (CDD)[c]	Yes	Yes	No	92
CDD15-03	M	84	Yes	Yes	2	Severe	n/a	n/a	Yes	Yes	Yes	Yes	No	No	No	Yes	No	No	n/a
CDD16-03	M	53	Yes	Yes	6	Moderate	n/a	n/a	Yes	Yes	Yes	Yes	No	No	Brother (ASD)	Yes	No	No	n/a
CDD17-03	M	68	Yes	No	3	Moderate	44	DAS	Yes	Yes	No	Yes	No	No	Father (autistic features)	Yes	No	No	n/a
CDD19-03	M	31	No	No	24	Severe	26	Mullen	Yes	Yes	No	Yes	No	No	Brother (autistic features)	No	Yes	No	60
CDD20-03	M	51	Yes	Yes	9	Mild	74	DAS	Yes	Yes	Yes	Yes	No	No	MZ twin (ASD)	Yes	No	No	n/a
CDD21-03	F	30	Yes	No	5	Severe	30	Mullen	Yes	Yes	No	Yes	Yes	No	No	Yes	No	No	n/a
CDD22-03	M	48	Yes	No	4	Severe	30	Leiter	Yes	Yes	Yes	Yes	Yes	No	No	Yes	No	No	n/a

Abbreviations: *ASD* autism spectrum disorder, *CDD* childhood disintegrative disorder, *DAS* Differential Ability Scales, *ET* eye tracking, *F* female, *fMRI* functional magnetic resonance imaging, *ID* intellectual disability, *IQ* intelligence quotient, *M* male, *MZ* monozygotic

[a]If multiple regressions occurred, the number indicates the length of the first period of regression

[b]For subjects who received the Mullen Scales, their full-scale IQ is based on a ratio IQ

[c]Subjects CDD13-03 and 13-04 are monozygotic twin girls with CDD who are counted as one proband for the genetics studies

Results

Study subjects

Clinical characteristics of the CDD cohort and the number of subjects examined by each study modality are shown in Table 1. The sex ratio of 3.25 males to 1 female is similar to that reported for ASD. The mean and median age at onset of symptoms was 46 and 40 months, respectively, with a range of 28 to 84 months. Seventy percent of subjects experienced a prodrome of anxiety and terror. Thirty percent of subjects had multiple episodes of regression. The length of the first regressive episode ranged from 2 months to almost 7 years in one subject. Most subjects have severe to profound ID, with the mean and median IQ being 30 and 26, respectively, with a range of 8 to 74. All had loss of language skills, loss of social skills or adaptive behavior, and loss of play skills. Sixty-five percent had loss of bowel or bladder control, and the same proportion had loss of motor skills. Although CDD has been reported to be almost always sporadic, a few of our subjects have immediate family members with ASD or autistic features, including two sets of monozygotic twins. Both members of one pair (CDD13-03/04) have CDD; in the other pair (CDD20-03/04), one has CDD and the other has ASD.

Genetics

Given the rarity, severity, and apparently sporadic transmission seen in most CDD cases, we hypothesized that rare variants of large effect contribute to the etiology. Indeed, there is abundant evidence for the contribution of rare variants to ASD [22–24]. As shown in Table 2, we found one or more rare variants for all but one proband, which were not shared by any unaffected sibling controls (Additional file 2: Table S2). We also looked for compound heterozygous variants in subjects by searching for additional variants in genes affected by de novo variants but did not find any. The rates of all high-probability (Bayesian quality score ≥ 50) de novo variants were 0.80/proband exome and 0.92/sibling exome (Additional file 2: Table S3), which are similar to the overall rates calculated from 11 recent WES studies of neurodevelopmental disorders: 1.00/proband exome, $n = 2358$; 0.82/control exome, $n = 731$ [25]. There were no significant differences in the rates of non-synonymous de novo, homozygous, and hemizygous variants (Additional file 2: Table S3); the rate of brain-expressed genes affected; phylogenetic *P* value (PhyloP) conservation scores at variant positions; Residual Variation Intolerance Scores (RVIS); and polymorphism phenotyping v2 (PolyPhen-2) scores (Additional file 2: Table S2) between the probands and siblings.

We found one de novo genic CNV in a proband (0.07/proband, Table 2), which is similar to rates previously reported for ASD [26], and none in siblings. The proband CNV is a 2 kb heterozygous deletion of the 3′UTR of *OGDHL*, which encodes a component of a mitochondrial protein complex implicated in neurodegeneration [27]. One gene, *SUPT20HL2*, and two gene families, *USP* and *BBS*, are affected in more than one CDD proband. Two hemizygous missense variants were identified in *SUPT20HL2*, which encodes a putative transcription factor but could be a pseudogene according to the UniProtKB database (http://www.uniprot.org/). Three members of the *USP* (*ubiquitin-specific peptidase*) gene family are affected in CDD probands: *USP9X* (hemizygous missense), *USP9Y* (paternally-inherited non-sense), and *USP26* (hemizygous missense). They encode deubiquitinating enzymes that prevent the degradation of proteins. Two members of the Bardet-Biedel Syndrome (BBS) gene family, which is involved in ciliogenesis, have de novo missense variants in CDD probands: *BBS5* and *BBS9*. Although the specific protein-changing variants identified in CDD subjects were rare and not previously associated with disease, we reviewed the literature and found some overlap between CDD candidate genes and genes potentially associated with other neurological disorders (Table 2).

There were no clearly deleterious variants in the CDD probands. To identify potentially pathogenic variants, we considered a combination of factors: (1) positive brain expression, (2) PhyloP score ≥ 1.30 ($P = 0.05$ for conservation), (3) negative RVIS (gene intolerant of variation), and (4) PolyPhen-2 classification of probably damaging missense (or n/a due to a variant other than missense). Of the 47 CDD candidate genes, 14 met all of these criteria: *NRK*, *TBC1D8B*, *TRRAP*, *NAV2*, *OGDHL*, *ZNF236*, *PRKCSH*, *MTMR8*, *BCOR*, *SRPK3*, *USP9Y*, *KIAA2018*, *CXorf57*, *and ALG13* (Table 2). To further refine this list, inspection of sequencing data from the Exome Aggregation Consortium (http://exac.broadinstitute.org/) revealed that: (1) the variants in all of the genes except *NAV2*, *MTMR8*, and *ALG13* are novel or found at most once in the dataset and (2) among the remaining 11 genes, 4 are among the 5% most intolerant: *TRRAP*, *ZNF236*, *BCOR*, and *KIAA2018*.

TRRAP (*transformation/transcription domain-associated protein*)affected by a de novo missense variant in a male CDD proband; it encodes a component of histone acetyltransferase complexes and is involved in DNA transcription and repair. It is not associated with an OMIM disorder, but de novo variants have been identified in other neurological disorders (Table 2). *ZNF236* (*Zinc Finger Protein 236*) is also affected by a de novo missense variant in a male proband; it may be involved in transcriptional regulation (UniProtKB) but is not associated with a known disorder. *BCOR* (*BCL6 Corepressor*) is affected by a hemizygous missense variant in a male CDD proband; it encodes a transcriptional corepressor. It is associated with syndromic

Table 2 Rare variants unique to CDD probands

Sex	Inheritance	Gene	AA change	Genomic coordinate (hg19)	Reference	Variant	Brain exp	PhyloP[b]	RVIS	PolyPhen2	Gene-associated OMIM disorder (#, inheritance)	Variants identified in ID, ASD, EE, SCZ, DD: #, inheritance, variant type [reference]
CDD03-03												
M	Hemizygous	*NRK*	K336R	chrX:105150568	A	G	Yes	5.88	n/a	0.999 (Mis3)	n/a	ID: 1 DN, 7 n/a missense [42–44]
M	Hemizygous	*TBC1D8B*	L653I	chrX:106093374	C	A	Yes	4.84	-1.15	0.999 (Mis3)	n/a	ASD: 1 DN missense [23]; ID: 1 DN, 2 n/a missense [42, 43]
M	Hemizygous	*NKRF*	K50R	chrX:118725239	T	C	Yes	1.39	-0.54	0.005 (Mis1)	n/a	ID: 3 n/a missense [44]
M	Hemizygous	*SAGE1*	A362V	chrX:134989926	C	T	Yes	-0.03	0.01	0.000 (Mis1)	n/a	ASD: 1 DN missense [23]; ID: 4 n/a missense [42, 44]
CDD05-03												
M	Homozygous	*TCTEX1D2*	R110C	chr3:196022930	G	A	Yes	2.69	0.28	1.000 (Mis3)	n/a	SCZ: 1 DN, 20 n/a deletion [45, 46]
M	Hemizygous	*USP26*	P683L	chrX:132160201	G	A	No	-1.09	0.46	0.000 (Mis1)	n/a	ID: 4 n/a missense [42, 44]
CDD07-03												
M	De novo	*BBS9*	L376P	chr7:33376163	T	C	Yes	5.72	1.05	0.222 (Mis1)	Bardet-Biedl syndrome 9 (615986, AR)	Syndromic ID
M	De novo	*TRRAP*	P1781S	chr7:98548580	C	T	Yes	7.70	-6.14	1.000 (Mis3)	n/a	ASD: 3 DN missense [23]; EE: 1 DN missense [47]; SCZ: 1 DN missense, 1 DN SS [48]; DD: 3 DN missense [49]
M	Homozygous	*DNMT3B*	A364T	chr20:31383238	G	A	Yes	0.31	-1.70	0.000 (Mis1)	Immunodeficiency-centromeric instability-facial anomalies syndrome 1 (242860, AR)	Syndromic ID
M	Hemizygous	*CDR1*	G161R	chrX:139866051	C	T	Yes	0.37	0.15	0.995 (Mis3)	n/a	ID: 2 n/a missense [42]
M	Hemizygous	*FAM50A*	E136K	chrX:153674872	G	A	Yes	6.75	-0.01	0.010 (Mis1)	n/a	ID: 1 n/a missense [42]; EE: 1 DN missense [47]
CDD08-03												
M	Hemizygous	*CNGA2*	F385S	chrX:150912129	T	C	No	7.56	0.49	0.999 (Mis3)	n/a	ID: 5 n/a missense [42, 44]
CDD09-03												
M	Homozygous	*NOP9*	A161_E162insEE	chr14:24769850	*	I:AGGAGG	Yes	0.59	0.98	n/a	n/a	DD: 1 DN indel [49]

Table 2 Rare variants unique to CDD probands *(Continued)*

M	Hemizygous	*ZXDA*	V264L	chrX:57936065	C	A	Yes	-0.44	n/a	0.000 (Mis1)	n/a	ID: 1 DN, 1 n/a missense [42, 43]
M	Hemizygous	*SPANXN2*	L41F	chrX:142795555	C	A	No	-1.83	0.41	0.946 (Mis2)	n/a	ID: 1 n/a missense [42]
CDD10-03												
M	De novo	*NAV2*	R2046W	chr11:20122629	C	T	Yes	3.78	-1.56	1.000 (Mis3)	n/a	ASD: 2 DN missense [23]
CDD12-03												
F	Homozygous	*ADAMTS18*	S660N	chr16:77359816	C	T	Yes	4.90	-0.09	0.001 (Mis1)	Microcornea, myopic chorioretinal atrophy, and telecanthus (615458, AR)	
CDD13-03 and 04 (MZ twins)[a]												
F	De novo	*BBS5*	I76M	chr2:170344335	A	G	Yes	1.29	0.04	0.82 (Mis2)	Bardet-Biedl syndrome 5 (615983, AR)	Syndromic ID
F	De novo	*NSD1*	R964W	chr5:176639097	C	T	Yes	0.27	-1.55	0.997 (Mis3)	Sotos syndrome 1 (117550, AD), Beckwith-Wiedemann syndrome (130650, AD)	Syndromic ID and ASD
F	De novo	*OGDHL*	3'UTR deletion	chr10:50940935-50943068	*	CNV: deletion	Yes	5.43	-0.54	n/a	n/a	
F	Hemizygous	*TUBGCP5*	T274S	chr15:22846945	A	T (paternal allele deleted)	Yes	8.29	-0.77	0.891 (Mis2)	n/a	ASD: 1 DN missense [23]; ID: 1 inh deletion [50]; EE: 1 inh deletion [50]; SCZ: 1 inh deletion [50]
CDD15-03												
M	De novo	*ZNF236*	L1599H	chr18:74659496	T	A	Yes	5.43	-1.53	0.999 (Mis3)	n/a	
M	Hemizygous	*SUPT20HL2*	G472R	chrX:24330019	C	T	No	1.32	n/a	0.884 (Mis2)	n/a	
CDD16-03												
M	Homozygous	*PRKCSH*	E314_E315del	chr19:11558343-11558348	*	D:GGAGGA	Yes	1.61	-0.89	n/a	Polycystic liver disease 1 (174050, AD)	
M	Hemizygous	*SUPT20HL2*	P795A	chrX:24329050	G	C	No	-0.55	n/a	0.047 (Mis1)	n/a	
M	Hemizygous	*USP9X*	A843G	chrX:41027363	C	G	Yes	7.62	-1.62	0.102 (Mis1)	Mental retardation, X-linked 99 (300919, XLR and 300968, XLD)	Syndromic ID and EE

Table 2 Rare variants unique to CDD probands *(Continued)*

M	Hemizygous	*GATA1*	A55T	chrX:48649679	G	A	Yes	1.06	-0.60	0.993 (Mis3)	Anemia (300835, XLR); Thrombocytopenia (314050, XLR and 300367, XLR)	
M	Hemizygous	*MTMR8*	N265K	chrX:63564995	G	T	Yes	4.00	-0.77	1.000 (Mis3)	n/a	ASD: 1 DN missense [23], 1 hemizygous SS [51]; ID: 5 n/a missense [42, 44]
M	Hemizygous	*DCX*	M51V	chrX:110654052	T	C	Yes	0.66	-0.10	0.002 (Mis1)	Lissencephaly, X-linked (300067, XL)	Syndromic ID and EE
M	Hemizygous	*CT47B1*	L236P	chrX:120008818	A	G	No	-1.15	n/a	0.001 (Mis1)	n/a	ASD: 1 DN missense [23]; ID: 2 n/a missense [44]
CDD17-03												
M	Heterozygous	*CLUL1*	5′SS disruption	chr18:618107	G	A (paternal)	Yes	6.04	0.66	n/a	n/a	
M	Hemizygous	*BCOR*	F1023L	chrX:39930395	G	C	Yes	8.97	-3.03	0.999 (Mis3)	Microphthalmia, syndromic 2 (300166, XLD)	Syndromic ID
M	Hemizygous	*BEX2*	Y97C	chrX:102564711	T	C	Yes	0.78	0.64	0.046 (Mis1)	n/a	ASD: 1 n/a missense [52]; ID: 1 n/a missense [42]; SCZ: 1 n/a missense [52]
M	Hemizygous	*VSIG1*	Q88R	chrX:107304707	A	G	Yes	1.19	-0.18	0.856 (Mis2)	n/a	ASD: 2 hemizygous SS [51]; ID: 1 n/a missense [44]
M	Hemizygous	*SRPK3*	R619Q	chrX:153050929	G	A	Yes	7.69	-0.33	1.000 (Mis3)	n/a	ID: 7 n/a missense [42, 44, 53]
M	Hemizygous	*USP9Y*	R122X	chrY:14837084	C	T	Yes	2.63	n/a	n/a	Spermatogenic failure, Y-linked, 2 (415000, YL)	
CDD20-03[a]												
M	Homozygous	*KIAA2018*	Q1472del	chr3:113376112-113376114	*	D:GCT	Yes	3.58	-0.71	n/a	n/a	ASD: 1 DN nonsense, 4 n/a missense [22]; EE: 1 DN missense [47]; SCZ: 1 DN missense [54]
M	Homozygous	*BRIP1*	V193I	chr17:59924512	C	T	Yes	-0.69	-0.64	0.000 (Mis1)	Breast cancer, early-onset (114480, AD); Fanconi anemia, complementation group J (609054, AR)	
M	Hemizygous	*PDK3*	K410Q	chrX:24557261	A	C	Yes	2.44	0.13	0.001 (Mis1)	Charcot-Marie-Tooth disease, 6 (300905, XLD)	
M	Hemizygous	*ARSF*	R386Q	chrX:3021857	G	A	Yes	6.00	0.36	0.999 (Mis3)	n/a	ASD: 1 n/a nonsense [51]; ID: 4 n/a missense, 3 n/a indel, 2 n/a SS [42, 44]
M	Hemizygous	*ALAS2*	V193I	chrX:55047546	C	T	Yes	3.26	0.13	0.803 (Mis2)	Anemia, sideroblastic, 1 (300751, XLR); Protoporphyria, erythropoietic (300752, XLD)	

Table 2 Rare variants unique to CDD probands *(Continued)*

M	Hemizygous	*STARD8*	3 bp deletion	chrX:67938303-67938305	*	D:AAG	Yes	3.59	0.74	n/a	n/a	ID: 2 n/a missense, 1 n/a indel [42, 44]
M	Hemizygous	*CXorf57*	R545T	chrX:105882817	G	C	Yes	6.24	-0.40	1.000 (Mis3)	n/a	ID: 2 n/a missense [42, 44]
M	Hemizygous	*ALG13*	S891F	chrX:110980084	C	T	Yes	4.77	-1.00	0.986 (Mis3)	Epileptic encephalopathy, early infantile, 36 (300884, XLD)	Syndromic ID and EE
CDD21-03												
F	Homozygous	*RAD51C*	T287A	chr17:56798128	A	G	Yes	5.16	0.11	0.988 (Mis3)	Fanconi anemia, complementation group O (613390, AR)	
F	Homozygous	*SHANK3*	G277R	chr22:51117800	G	A	Yes	1.95	n/a	0.002 (Mis1)	Phelan-McDermid syndrome (606232, DN); Schizophrenia 15 (613950, DN)	Syndromic ID, ASD, EE, SCZ
CDD22-03												
M	Homozygous	*DAP3*	G5E	chr1:155679584	G	A	Yes	1.86	-0.22	0.890 (Mis2)	n/a	ASD: 1 DN missense [23]
M	Hemizygous	*ENOX2*	E12K	chrX:129843232	C	T	Yes	-0.35	-0.43	0.000 (Mis1)	n/a	ID: 1 hemizygous, 1 n/a missense [42, 55]

Abbreviations: *AA* amino acid, *AD* autosomal dominant, *AR* autosomal recessive, *ASD* autism spectrum disorder, *CDD* childhood disintegrative disorder, *CNV* copy number variant, *DD* developmental disorders, *DN* de novo, *EE* epilepsy/epileptic encephalopathy, *F* female, *indel* insertion/deletion, *inh* inherited, *ID* intellectual disability, *M* male, *Mis1* Missense1 (benign), *Mis2* Missense2 (possibly damaging), *Mis3* Missense3 (probably damaging), *MZ* monozygotic, *OMIM* Online Mendelian Inheritance in Man, *PhyloP* phylogenetic P-value, *PolyPhen-2* polymorphism phenotyping v2, *RVIS* Residual Variation Intolerance Score, *SCZ* schizophrenia/childhood-onset schizophrenia, *SS* splice site, *XL* X-linked, *XLD* X-linked dominant, *XLR* X-linked recessive, *YL* Y-linked

[a]No discordant variants were confirmed between monozygotic twin girls CDD13-03 and 04 (CDD) and between monozygotic twin boys CDD20-03 (CDD) and 04 (ASD)

[b]PhyloP scores for indels and CNVs were calculated by averaging the PhyloP scores for all positions affected

*wild-type sequence

microphthalmia, which can have the feature of ID but otherwise does not characterize the CDD proband and is usually caused by truncating mutations in females. *KIAA2018* is affected by a homozygous one-amino acid deletion in a male CDD proband; it is also known as USF3 (*upstream transcription factor 3*). It is not associated with an OMIM disorder, but de novo variants have been identified in other neurological disorders (Table 2). Of note, all of these top candidate genes either play a role or may be involved in transcription, which characterizes many ASD-associated genes as well [22].

Using the human BrainSpan exon-array transcriptome dataset [20], we plotted the median expression level of the CDD candidate genes as a group for all the brain regions available from embryonic to late adulthood stages ($n = 40$ genes represented once in the core probe set, Additional file 2: Table S4). As shown by the expression profile in Fig. 1, CDD candidate genes are more highly expressed in non-neocortical regions [hippocampus (HIP), amygdala (AMY), striatum (STR), mediodorsal nucleus of the thalamus (MD), and/or cerebellar cortex (CBC)] compared to neocortical regions across the lifespan (Additional file 2: Table S5). Moreover, there are increasing levels of expression in the AMY, STR, and HIP during periods 10 (1–6 years old) and 11 (6–12 years-old), the range that encompasses the age of onset of symptoms in our CDD cohort.

Given this observation, we compared the difference in median expression levels between non-neocortical and neocortical regions for genes affected by non-synonymous and synonymous variants in CDD probands, their unaffected siblings, and ASD probands from the Simons Simplex Collection (SSC) with and without regression [23] matched by sex, age at evaluation, IQ, and autism symptom severity (see Additional file 1: Supplementary information for cohort selection details). The expression profile of CDD candidate genes is qualitatively distinct from the other gene sets, except that it is similar to the profile of genes affected by non-synonymous variants in SSC probands with regression, even though they have only one gene, *NAV2*, in common (Fig. 2, Additional file 2 : Tables S4 and S6). The difference in expression, non-neocortical minus neocortical, reaches a maximum positive value at mid-fetal stages. For CDD candidate genes, this occurs at period six [19–24 postconceptual weeks (PCW)]; permutation testing with 100,000 iterations of 40 randomly selected genes from the BrainSpan dataset confirmed the significance of this differential expression ($P = 0.0022$). We extended the analysis to several other gene sets, such as those identified in SSC probands and unaffected siblings with non-synonymous, synonymous, and LGD variants; genes most significantly associated with ASD by three recent large WES and CNV studies [22–24]; and all genes in the BrainSpan dataset. The expression profile of genes affected by non-synonymous variants in CDD probands and SSC probands with regression is qualitatively distinct from these other sets as well (Additional file 1: Figure S1, Additional file 2: Tables S4 and S6).

We also investigated whether CDD candidate genes are coexpressed with each other. Of the 40 candidate genes, 11 are coexpressed with at least one other candidate gene across all brain regions and time periods with a Pearson correlation coefficient $r \geq 0.7$ (Fig. 3, Additional file 2: Table S7). There are 23 such connections, for a mean of 2.09 correlations/gene and a mean coefficient of 0.779. Permutation testing with 100,000 iterations of 40 randomly selected genes from the BrainSpan dataset revealed that observing 11 genes with at least 2.09 correlations/gene is significant ($P = 0.036$), as is observing 11 genes with a mean correlation coefficient of at least 0.779 ($P = 0.019$). Meeting both thresholds is also significant ($P = 0.0059$). Since all 11 CDD candidate genes which are coexpressed with each other have positive brain expression as per the BrainSpan dataset, permutation testing with 100,000 iterations was also performed with 40 randomly selected brain-expressed genes from BrainSpan. While observing 11 genes with at least 2.09 correlations/gene is not significant ($P = 0.066$), observing 11 genes with a mean correlation coefficient of at least 0.779 is significant ($P = 0.022$) as is meeting both thresholds ($P = 0.011$). Comparing the set of 11 coexpressed CDD candidate genes with the remaining set of 29 which are not coexpressed revealed no significant differences between the rate of brain-expressed genes, PhyloP scores, or PolyPhen-2 scores; however, the coexpressed genes are significantly more intolerant of variation (average RVIS −1.42 versus −0.15, $t(35) = -2.91$, $P = 0.0062$, independent t test, two-tailed). Gene ontology enrichment analysis using the Database for Annotation, Visualization and Integrated Discovery v6.8 (https://david.ncifcrf.gov/) for the whole set of CDD candidate genes, and the subset of 11 coexpressed genes did not identify significant enrichment of GO terms after Benjamini-Hochberg correction of P values.

Neural systems

Given the universality of social deficits in ASD, dysfunction in brain systems subserving social perception, including the perception of faces, is a key focus of ASD research. Face and house visual stimuli reliably activate and dissociate systems involved in socioemotional (fearful faces) and non-socioemotional (houses) information processing. We studied four cohorts: CDD ($n = 7$), LFASD ($n = 7$), HFASD ($n = 14$), and TD ($n = 19$). Even though individuals with LFASD are more numerous than those with CDD, our sample size was still limited by the difficulty of obtaining high-quality neuroimaging (and

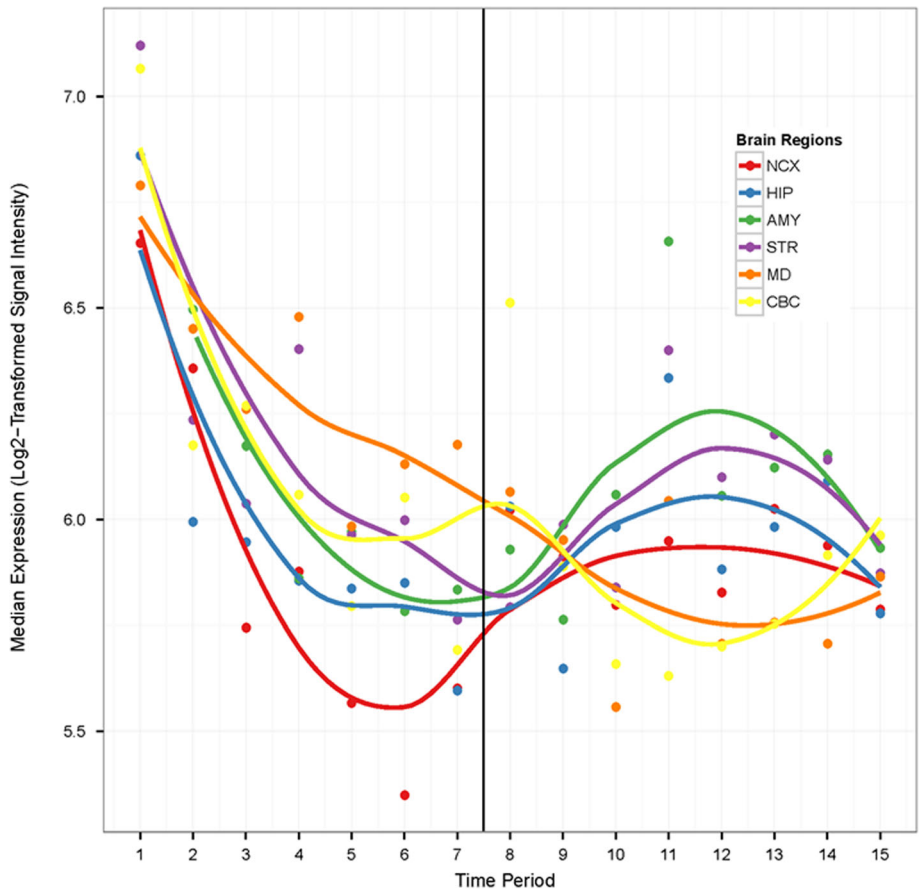

Fig. 1 Median expression levels of CDD candidate genes ($n = 40$) by brain region and time period (Additional file 2: Table S5) using the human BrainSpan exon-array transcriptome dataset [20]. The dark vertical line indicates birth. Log_2-transformed signal intensity ≥ 6 in at least one sample is considered positive expression [20]. *AMY* amygdala, *CBC* cerebellar cortex, *HIP* hippocampus *MD* mediodorsal nucleus of the thalamus, *NCX* neocortex, *STR* striatum

eye-tracking) data in low-functioning subjects. That being said, to our knowledge, this is the first ever presentation of non-sedated fMRI data from individuals with ASD and marked ID.

There were no significant differences in sex, age, intracranial volume, and head movement in the scanner between the four cohorts. The CDD and LFASD groups were also not significantly different by IQ and autism severity, and the HFASD and TD groups were not significantly different by IQ (Additional file 2: Tables S8 and S9). First, we utilized a discovery sample of 12 of our 19 TD subjects in a whole-brain analysis for an independent localization of regions of interest involved in processing faces relative to houses. Figure 4a illustrates regions of ventrolateral occipitotemporal cortex where TD subjects exhibited significant faces > houses activation (Additional file 2: Table S10). These regions included the expected locations of well-known nodes of the occipitotemporal face-sensitive network including the fusiform face area [28, 29] and the occipital face area [30]. As shown in Fig. 4b and Additional file 2: Table S11, extraction of the mean percent signal change (faces > houses) for each of the four groups [TD:validation (the remaining 7 of the 19 TD subjects), HFASD, LFASD, and CDD] indicated an absence of group differences in the response to faces versus houses in these independently defined regions of interest when comparing the TD:validation and HFASD groups [$t(19) = 0.17$, $P = 0.87$, Cohen's $d = 0.08$] and when comparing the LFASD and CDD groups [$t(12) = 0.97$, $P = 0.35$, Cohen's $d = 0.56$]. The faces > houses response within the CDD group was not significantly greater than zero [$t(6) = 0.80$, $P = 0.45$, Cohen's $d = 0.30$], suggesting an overall lack of sensitivity to faces in the occipitotemporal face-sensitive network as a whole. There is a well-established finding of hypoactivation to faces (versus houses) in the right, middle fusiform gyrus in HFASD relative to TD [31]. We were able to replicate this finding in our cohorts [$t(31) = 3.54$, $P = 0.0013$, Cohen's $d = 1.29$]. However, comparison of faces > houses activity

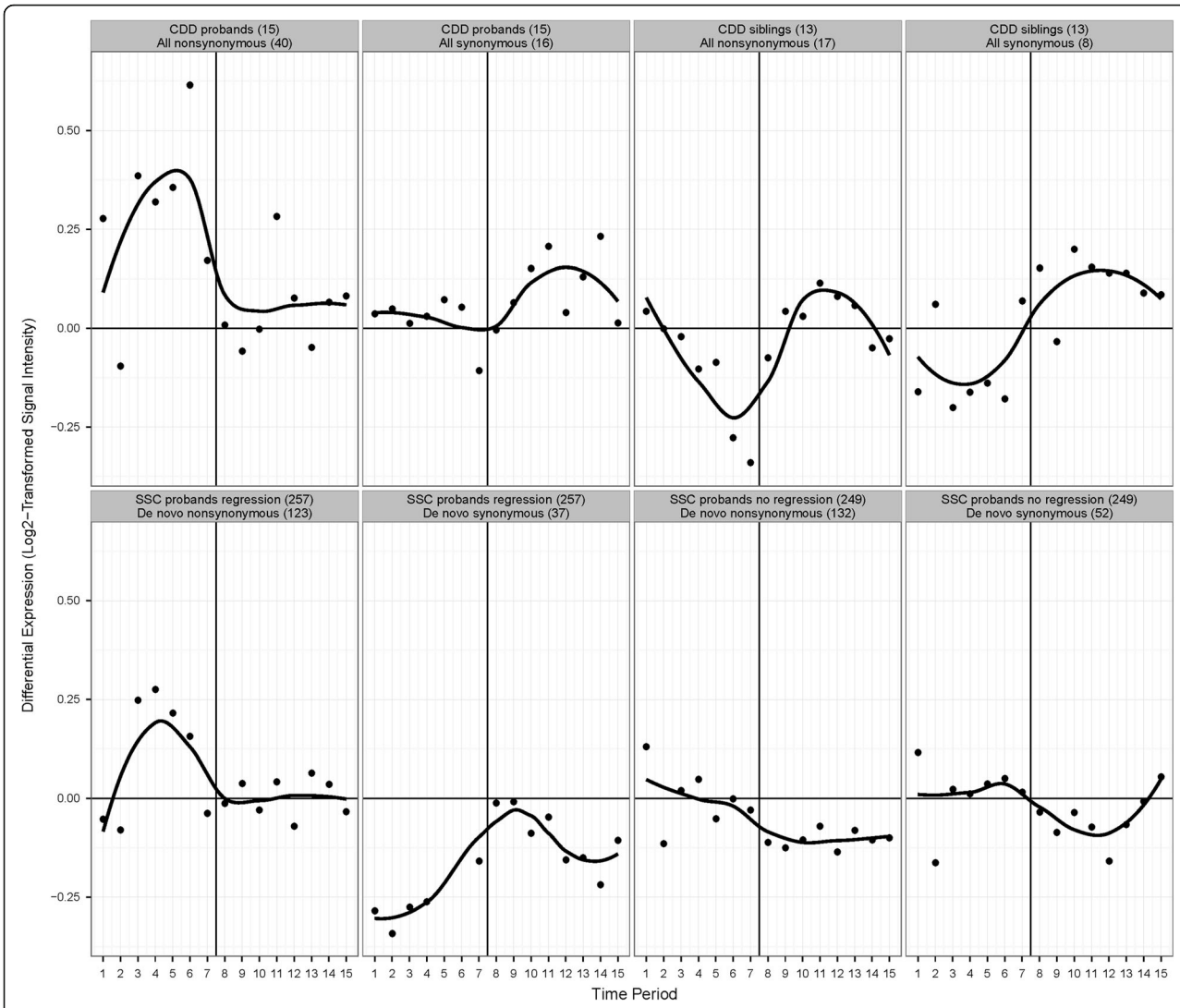

Fig. 2 Differential expression levels of various gene sets. The difference in median expression levels (non-neocortical minus neocortical brain regions) is shown for genes affected by non-synonymous or synonymous variants in CDD probands, their unaffected siblings, SSC probands with regression, and SSC probands without regression. The number in parentheses indicates the number of subjects or variants, and the *dark vertical line* in each panel indicates birth. For potential CDD candidate genes, the difference reaches a maximum positive value at period six (mid-fetal stages); significance was confirmed by permutation testing with 100,000 iterations of 40 randomly selected genes ($P = 0.0022$). *CDD* childhood disintegrative disorder, *SSC* Simons Simplex Collection

in CDD relative to TD revealed no significant difference [$t(24) = 1.18$, $P = 0.25$, Cohen's $d = 0.54$], as did the comparison of the LFASD and TD groups [$t(24) = 1.10$, $P = 0.28$, Cohen's $d = 0.51$] (Additional file 1: Figure S2 and Additional file 2: Table S12).

Given the possible lack of sensitivity to faces in the ventrolateral occipitotemporal cortex in CDD, we next conducted a whole-brain evaluation of the CDD subjects to localize the neuroanatomical substrates of face perception in these individuals. As illustrated in Fig. 5a, CDD subjects exhibited faces > houses activity in the middle frontal gyrus, precentral gyrus, caudate (striatum), thalamus, hippocampus, and cerebellum (Additional file 2: Table S13). These overlap with brain regions determined to have the highest levels of CDD candidate gene expression (Fig. 1). As shown in Fig. 5b and Additional file 2: Table S14, comparison of the mean percent signal change (faces > houses) from these regions of interest revealed a significant difference between CDD and HFASD [$t(19) = 2.98$, $P = 0.0076$, Cohen's $d = 1.45$], but no significant difference between CDD and LFASD [$t(12) = 1.71$, $P = 0.11$, Cohen's $d = 0.99$]. The LFASD group showed an intermediate phenotype to that of HFASD and CDD groups (Fig. 5b).

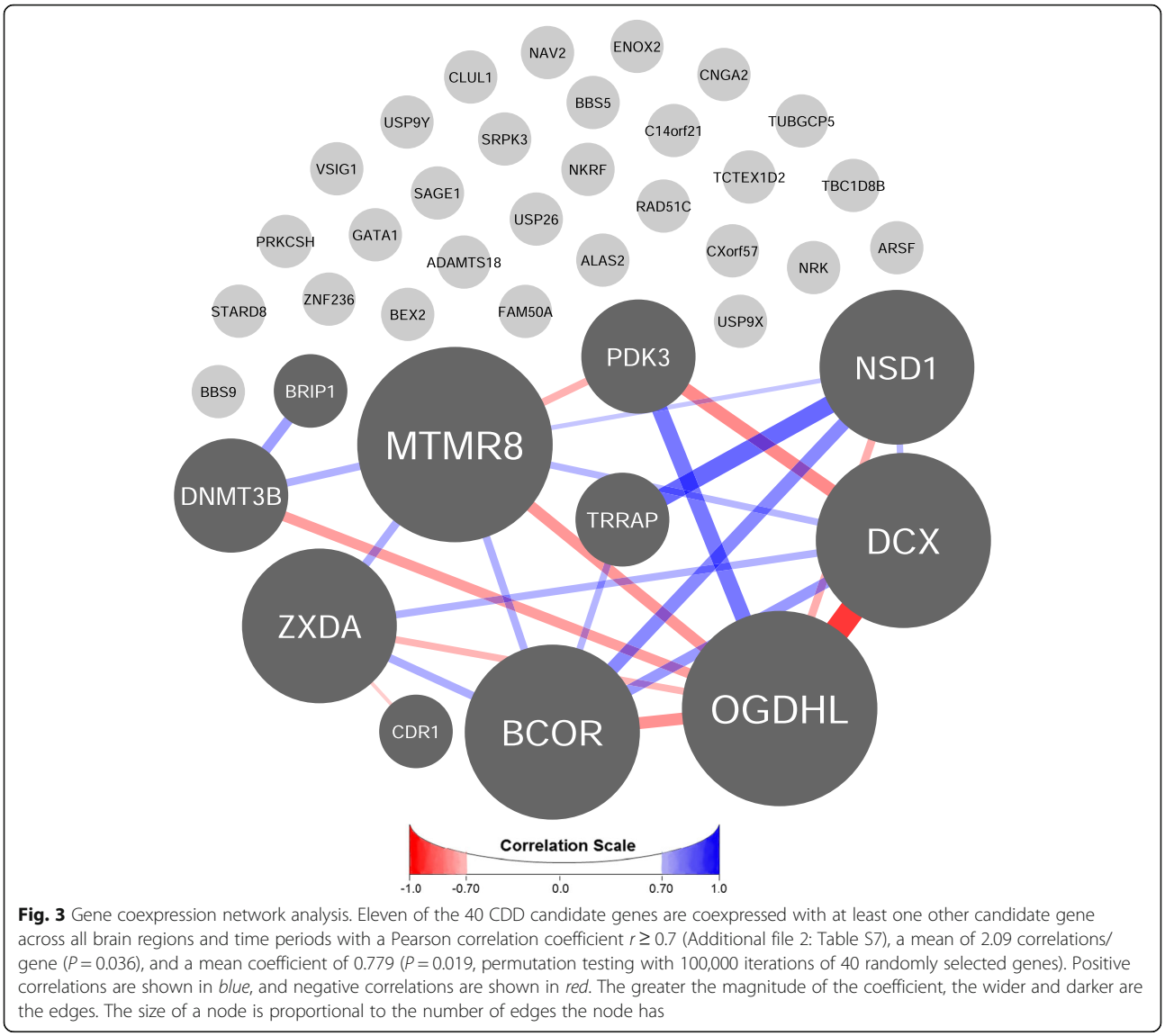

Fig. 3 Gene coexpression network analysis. Eleven of the 40 CDD candidate genes are coexpressed with at least one other candidate gene across all brain regions and time periods with a Pearson correlation coefficient $r \geq 0.7$ (Additional file 2: Table S7), a mean of 2.09 correlations/gene ($P = 0.036$), and a mean coefficient of 0.779 ($P = 0.019$, permutation testing with 100,000 iterations of 40 randomly selected genes). Positive correlations are shown in *blue*, and negative correlations are shown in *red*. The greater the magnitude of the coefficient, the wider and darker are the edges. The size of a node is proportional to the number of edges the node has

Eye-gaze behavior

We collected eye-tracking data to quantify the social phenotype of our four cohorts [CDD ($n = 5$), LFASD ($n = 7$), HFASD ($n = 32$), and TD ($n = 14$)] as they viewed emotional faces [21]. As shown in Additional file 2: Tables S15 and S16, the groups were not significantly different by sex, age, and total fixation duration on the image. The CDD and LFASD groups were also not significantly different by IQ and autism severity, and the HFASD and TD groups were not significantly different by IQ. As shown in Fig. 6, we replicate prior findings [32–34] of decreased fixation on the eyes [$t(44) = -2.28$, $P = 0.03$, Cohen's $d = 0.77$] and increased fixation on the mouth [$t(44) = 2.16$, $P = 0.04$, Cohen's $d = 0.76$] in HFASD relative to TD. However, while the percentage of time subjects with LFASD spent looking at the eyes did not differ significantly from the HFASD group [$t(37) = 0.43$, $P = 0.67$, Cohen's $d = 0.17$], CDD subjects fixated eyes significantly more than the HFASD group [$t(35) = 2.19$, $P = 0.04$, Cohen's $d = 1.08$]. Compared to each other, CDD and LFASD subjects did not differ significantly in time spent looking at the eyes [$t(10) = 1.35$, $P = 0.21$, Cohen's $d = 0.87$]. As with the fMRI results (Fig. 5b), the LFASD group showed an intermediate phenotype to that of the HFASD and CDD groups (eye-mouth ratio, Additional file 2: Table S16).

Discussion

We are reporting the first multimodal analysis of CDD, a rare form of ASD characterized by late-onset, severe regression. The small cohort size due to its low prevalence, and the challenges of obtaining interpretable data from non-sedated fMRI and eye tracking in subjects with ASD and significant ID necessitated an exploratory study.

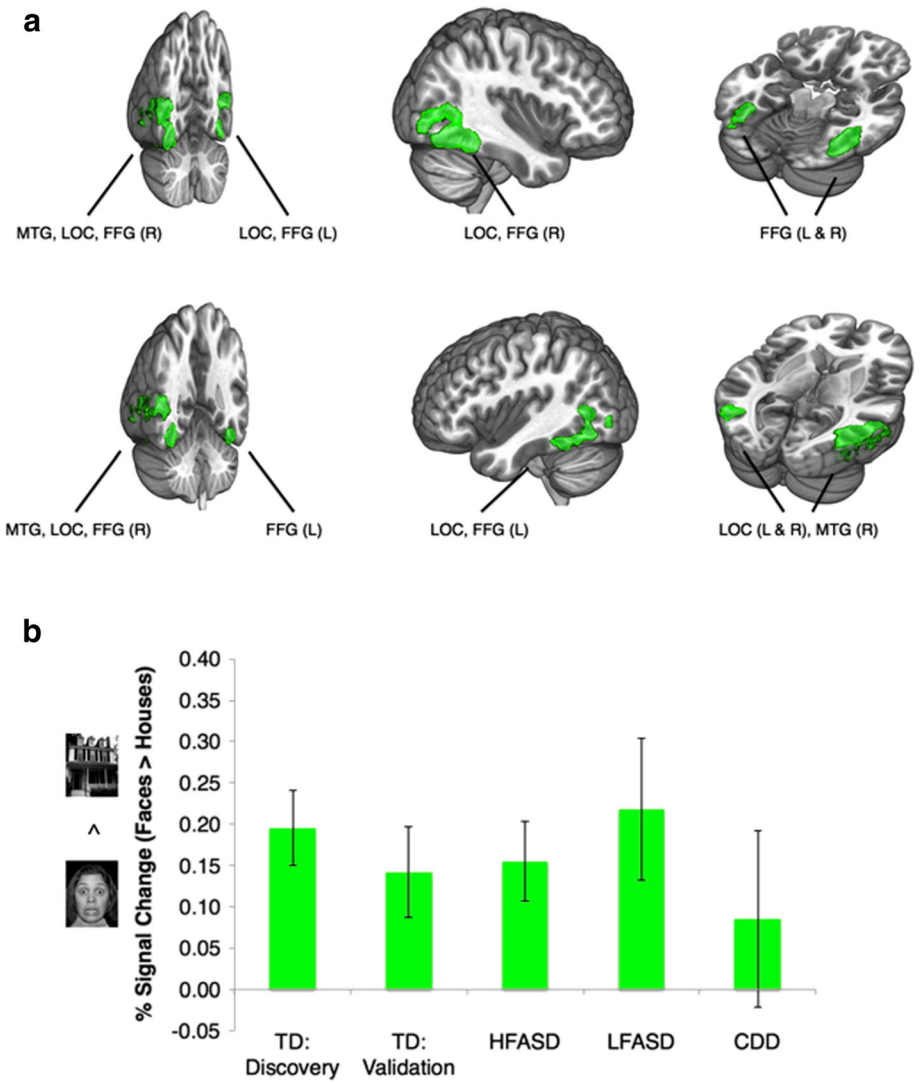

Fig. 4 Brain regions of interest (ROIs) involved in processing socioemotional (*fearful face*) versus non-socioemotional (*house*) visual stimuli. **a** The *green* color brain map indicates regions of significant faces > houses activation in a discovery sample of 12 TD subjects ($Z > 3.09$, whole-brain corrected at the cluster-level $P < 0.05$). **b** These independently defined ROIs were then utilized for comparisons across the four remaining cohorts, a TD:validation sample ($n = 7$), HFASD ($n = 14$), LFASD ($n = 7$), and CDD ($n = 7$). The *bar graph* indicates the mean % signal change (faces > houses) for each cohort. Group differences were not significant when comparing the TD:validation and HFASD groups [$t(19) = 0.17$, $P = 0.87$, Cohen's $d = 0.08$] and when comparing the LFASD and CDD groups [$t(12) = 0.97$, $P = 0.35$, Cohen's $d = 0.56$]. The faces > houses response within the CDD group was not significantly greater than zero [$t(6) = 0.80$, $P = 0.45$, Cohen's $d = 0.30$]. *Error bars* indicate standard error of the mean. All *P* values were calculated by independent *t* test and are two-tailed. *FFG* fusiform gyrus, *L* left, *LOC* lateral occipital cortex, *MTG* middle temporal gyrus, *R* right

There is a relative deficiency of reports using these protocols with low-functioning individuals on the autism spectrum. For the first time, a combination of fMRI and eye-tracking analyses are being presented for such individuals to help fill this gap.

Gene expression, neuroimaging, and social behavior analyses suggest that there are neurobiological differences which may underlie the distinct clinical features of CDD. Although no clearly deleterious variants or highly recurrent candidate genes were identified, candidate genes most conserved at variant position or most intolerant of variation, such as *TRRAP*, *ZNF236*, and *KIAA2018*, play a role or may be involved in transcription, which characterizes many ASD-associated genes as well [22]. Gene expression analysis provided some potential insights into CDD. The expression profile of CDD candidate genes resembled that of SSC probands with regression but not SSC probands without regression (matched by IQ), suggesting a pattern relevant to regression. A significant number of CDD candidate genes are co-expressed and may interact in pathways important to the pathophysiology of the disorder.

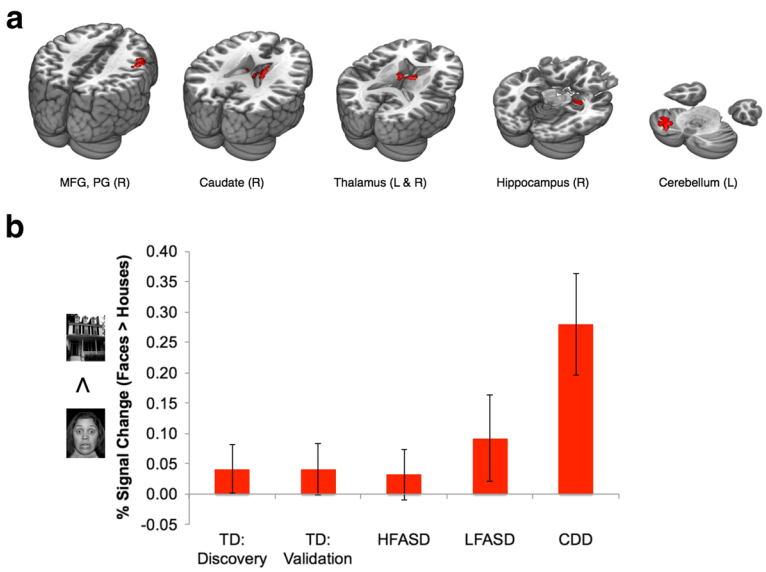

Fig. 5 CDD whole-brain fMRI analysis. **a** The *red color* brain map indicates regions of significant faces > houses activation in the CDD subjects ($Z > 3.09$, whole-brain corrected at the cluster-level $P < 0.05$). **b** The *bar graph* indicates the mean % signal change (faces > houses) within these areas for each cohort: TD:discovery ($n = 12$), TD:validation ($n = 7$), HFASD ($n = 14$), LFASD ($n = 7$), and CDD ($n = 7$). The CDD cohort differed significantly from HFASD [$t(19) = 2.98$, $P = 0.0076$, Cohen's $d = 1.45$] but not from LFASD [$t(12) = 1.71$, $P = 0.11$, Cohen's $d = 0.99$]. *Error bars* indicate standard error of the mean. All *P* values were calculated by independent *t* test and are two-tailed. *MFG* middle frontal gyrus, *PG* precentral gyrus

Interestingly, expression of the candidate genes overlapped with face-evoked brain hyperactivity in CDD in non-neocortical regions, such as the thalamus, cerebellum, caudate (striatum), and hippocampus. These regions are known to be involved in distributing eye movements (and thus attention) to socially meaningful stimuli, including faces, early in development. Increased face-evoked activity in CDD was paralleled by increased attention to the eyes of faces, culminating in a normal distribution of attention to the eyes. Still unresolved,

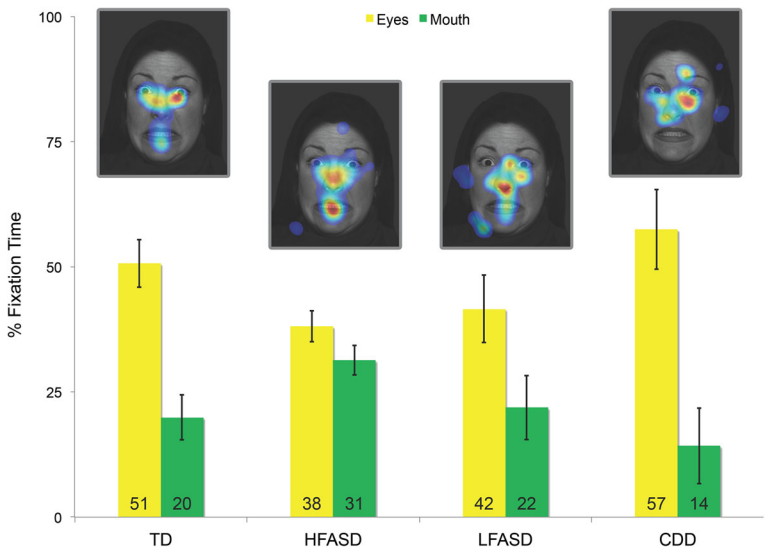

Fig. 6 Behavioral analysis through eye tracking. The *yellow* and *green bars* of the graph represent the mean % of time spent fixating (y axis) on the eyes and mouth of the faces, respectively, by cohort (x axis): TD ($n = 14$), HFASD ($n = 32$), LFASD ($n = 7$), CDD ($n = 5$). The *gaze heat maps* illustrate the group-level gaze data overlaid on one of the images at which subjects looked. Compared to TD subjects, HFASD subjects show decreased fixation on the eyes [$t(44) = -2.28$, $P = 0.03$, Cohen's $d = 0.77$] and increased fixation on the mouth [$t(44) = 2.16$, $P = 0.04$, Cohen's $d = 0.76$]. The % of time subjects with LFASD spent looking at the eyes did not differ from HFASD [$t(37) = 0.43$, $P = 0.67$, Cohen's $d = 0.17$], but CDD subjects fixated eyes significantly more than HFASD [$t(35) = 2.19$, $P = 0.04$, Cohen's $d = 1.08$]. *Error bars* indicate standard error of the mean. All *P* values were calculated by independent *t* test and are two-tailed

though, is how a more typical viewing pattern relates to the poor outcomes which characterize CDD.

Individuals with ASD with greater communicative competence show a more atypical pattern of attention toward faces comprised of decreased looking at the eyes and increased looking at the mouth [35, 36], while individuals with ASD and language impairment have been reported to not differ from typical peers [36]. We found a similar discontinuity in face information processing behaviors, with atypical face-viewing strategies evidenced most clearly for the (most able) HFASD group, and a more typical pattern for the CDD group. While CDD did not differ significantly from LFASD on eye-tracking measures, CDD showed the strongest between group differences in effect sizes referenced against the atypical looking patterns of HFASD. The presence of intact orientation to the eyes and unusual face-sensitive brain activation suggest an alternative developmental pathway for face processing in CDD.

Coinciding with the onset of canonical babbling, the typical infant's transition from looking at the eyes of a speaker to looking at the mouth is between 4 and 8 months of age [37]. This bias reverts back to the eyes by 12 months for infants viewing people speaking their native language (but not a foreign language), an effect probably driven by advancing expertise and perceptual narrowing. Preferences for looking at mouths in HFASD may reflect higher-order cognitive compensatory mechanisms with scaffolding functions analogous to the 4–8 month transitional period in typical development or biases for second languages later in infancy [33, 38], whereas LFASD and CDD may lie on opposite sides of the 4–8 month divide. The unique face-evoked activity that localized to a set of subcortical structures and the cerebellum in CDD suggests a neoteny in the development of the face-processing system whereby subcortical mechanisms thought to control orienting and attention to faces [39, 40] early in human development remain abnormally involved or cease to be inhibited by top-down regulation following regression. This may represent a marker of the unique developmental process underlying CDD, thereby suggesting a target for studies utilizing eye tracking for early identification and stratification of behaviorally and biologically heterogeneous forms of ASD [36, 41].

It is important to note that our investigations occurred months to years after the onset of symptoms. Since CDD is typically a diagnosis of exclusion, subjects come to our attention for the purposes of research long after the regressive period. Therefore, how the neurobiological features of CDD that we identified relate to the course of regression is unknown. It will be essential to confirm our results in larger cohorts. Ideally, subjects with CDD would be studied before and after the regression to better identify neurobiological correlates; however, this is challenging with such a rare disorder. Since regression is frequently described in ASD, prospective studies of more typical cases of regression may determine whether our results are relevant to regression in the autism spectrum more broadly. It would also be interesting to conduct our studies in regressive disorders such as Rett syndrome. Furthermore, since the fMRI and eye-tracking results revealed that the LFASD subjects had phenotypes intermediate to those of CDD and HFASD, it will be important to study an ID cohort without ASD to better attribute group differences to the effects of ASD versus ID. A major future challenge will be to elucidate the mechanisms by which variants in a set of genes may lead to areas of brain hyperactivity and an apparently normal attention to eyes but, in the end, the severe autism which characterizes CDD.

Conclusions

In summary, we pursued a multidisciplinary, multi-level approach comprising genetic, brain, and behavioral analyses to conduct an exploratory study of CDD, a rare and severe condition of unclear etiology. Although CDD and other forms of ASD have clinical similarities, the unique natural history of CDD may mark some unique neurobiological features. The clinical and genetic heterogeneity of ASD are well established; our results suggest that there is also heterogeneity of biomarkers, such as affected brain regions and neural circuits. Biomarkers established for high-functioning individuals with ASD may not apply to the substantial proportion of individuals on the spectrum who have ID. Ultimately, the recognition of an increasing number of specific ASD biotypes may translate into more targeted diagnostic tests and treatments.

Additional files

Additional file 1: Supplementary information includes supplementary methods, references, and figures. (PDF 700 kb)

Additional file 2: Table S1. CDD families for genetic analysis. **Table S2.** Rare non-synonymous and synonymous variants from WES unique to probands or unaffected sibling controls. **Table S3.** Rates of variants from WES in CDD probands and unaffected sibling controls. **Table S4.** Genes represented once in the core probe set (Kang et al. 2011) and used for expression analysis. **Table S5.** Median expression values (Log2-transformed signal intensity) for CDD candidate genes by time period and brain region. **Table S6.** Difference in median expression values (Log2-transformed signal intensity) between non-neocortical and neocortical regions for gene sets by time period. **Table S7.** Pearson correlation coefficients for all pairwise combinations of CDD candidate genes. **Table S8.** Clinical characteristics of subjects studied by neuroimaging. **Table S9.** Features of CDD, LFASD, HFASD, and TD cohorts for neuroimaging analysis. **Table S10.** List of brain regions where TD:discovery exhibits significant faces > houses activation. **Table S11.** Mean % signal change (faces > houses) for each cohort in Fig. 4b. **Table S12.** Mean % signal change (faces > houses) for each cohort in Figure S2. **Table S13.** List of brain regions where CDD exhibits significant faces > houses activation. **Table S14.** Mean % signal change (faces > houses)

for each cohort in Fig. 5b. **Table S15.** Clinical characteristics of subjects studied by eye tracking. **Table S16.** Features of CDD, LFASD, HFASD, and TD cohorts in eye-tracking analysis. **Table S17.** Whole-exome sequencing quality metrics. **Table S18.** Features of SSC probands with and without regression by IQ and autism severity. (XLSX 364 kb)

Abbreviations

AMY: Amygdala; ASD: Autism spectrum disorder; CBC: Cerebellar cortex; CDD: Childhood disintegrative disorder; CNV: Copy number variant; FFG: Fusiform gyrus; fMRI: Functional magnetic resonance imaging; FSIQ: Full-scale IQ; HFASD: High-functioning ASD; HIP: Hippocampus; ID: Intellectual disability; IQ: Intelligence quotient; LFASD: Low-functioning ASD; LGD: Likely gene-disrupting; LOC: Lateral occipital cortex; MD: Mediodorsal nucleus of the thalamus; MFG: Middle frontal gyrus; MTG: Middle temporal gyrus; OMIM: Online Mendelian Inheritance in Man; PG: Precentral gyrus; PhyloP: Phylogenetic *P* value; PolyPhen-2: Polymorphism Phenotyping v2; ROI: Region of interest; SSC: Simons Simplex Collection; STR: Striatum; TD: Typically developing; WES: Whole-exome sequencing

Acknowledgements

We are very grateful to the families who participated in this study. We are thankful to the following colleagues for assistance and comments: Kaya Bilguvar, Marie Claudel, Jessica Connelly, Charles Farber, Richard Lifton, Francesc Lopez, Shrikant Mane, John Murdoch, Laura Politte, Stephan Sanders, Ariella Ritvo Slifka, Michele Spencer-Manzon, Matthew State, Irina Tikhonova, Jeremy Willsey, Nicole Wright, and Andrew Zimmerman.

Funding

This research was funded by the Simons Foundation (206929 R10981 to ARG and KAP), the National Institutes of Health (K08MH087639 to ARG; K08MH099424-02 to TVF; Clinical Investigation and the Clinical and Translational Science UL1 TR000142 and KL2 TR000140 from the National Center for Research Resources to AGE-S; CTSA UL1 RR024139, Expedition in Computing award 1139078, and R21MH102572 to FS; RR19895 and RR029676-01 to the Yale University Biomedical High Performance Computing Center), the Alan B. Slifka Foundation and Foundation Rumsey Cartier to AW, and Autism Speaks Meixner Postdoctoral Fellowship in Translational Research (#9284) and Hilibrand Autism Fellowship to DYJY.

Availability of data and materials

The data generated and analyzed during this study are included in this article (and its additional files) and are available from the corresponding author upon request.

Authors' contributions

ARG conceived and designed the genetics study, collected and analyzed the genetic data, evaluated the medical records, and co-authored the manuscript. AW recruited and performed clinical characterization of the study subjects, co-designed the fMRI and eye-tracking experiments, and collected and analyzed the fMRI and eye-tracking data. DYJY analyzed the fMRI data. CAWS collected and analyzed WES and CNV data. JE analyzed fMRI and eye-tracking data. SZ collected and analyzed WES data. AV collected and analyzed fMRI and eye-tracking data. BCVW analyzed fMRI data. PV performed clinical characterization of study subjects. ZW collected and analyzed WES and CNV data. TVF analyzed genotyping data and performed ancestry mapping. AGES analyzed and confirmed CNV data. MFW collected and analyzed WES data. MC collected and analyzed WES data. AS analyzed fMRI data. TH recruited and performed the clinical characterization of study subjects. GB recruited and performed clinical characterization of study subjects. HF analyzed clinical characteristics of study subjects and managed fMRI and eye-tracking databases. CC recruited and performed clinical characterization of study subjects. AR analyzed the clinical characteristics of study subjects. FS analyzed eye-tracking data. FRV recruited and performed the clinical characterization of study subjects. KAP conceived and designed the overarching gene-brain-behavior approach, co-designed the fMRI and eye-tracking experiments, and co-authored the manuscript. All authors read, edited, and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

The HIC protocol provides for consent for publication. However, no individually identifying information is presented in this report.

Ethics approval and consent to participate

This research was approved by the Yale University Institutional Review Board, and informed consent was obtained for all subjects (Human Investigations Committee (HIC) Protocol # 0301024156 and 1004006656).

Publisher's Note

Author details

[1]Department of Pediatrics, Yale School of Medicine, New Haven, Connecticut, USA. [2]Child Study Center, Yale School of Medicine, New Haven, Connecticut, USA. [3]Department of Psychiatry, Yale School of Medicine, New Haven, Connecticut, USA. [4]Department of Genetics, Yale School of Medicine, New Haven, Connecticut, USA. [5]Evelina London Children's Hospital, Guy's and St. Thomas' Trust, Kings Health Partners AHSC, London, UK.

Received: 22 December 2016 Accepted: 15 March 2017
Published online: 04 April 2017

References

1. American Psychiatric Association. Autism spectrum disorder. In: Diagnostic and Statistical Manual of Mental Disorders: DSM-5. 5th ed. Washington: American Psychiatric Association; 2013. p. 50–9.
2. Kanner L. Autistic disturbances of affective contact. Nerv Child. 1943;2:217–50.
3. Heller T. Dementia infantilis. Zeitschrift fur die Erforschung und Behandlung des Jugenlichen, Schwachsinns. 1908;2:17–28.
4. Westphal A, Schelinski S, Volkmar F, Pelphrey K. Revisiting regression in autism: Heller's dementia infantilis. J Autism Dev Disord. 2013;43:265–71.
5. World Health Organization. The ICD-10 classification of mental and behavioural disorders: clinical descriptions and diagnostic guidelines. Geneva: World Health Organization; 1992.
6. American Psychiatric Association. Pervasive Developmental Disorders. In: Diagnostic and statistical manual of mental disorders: DSM-IV. 4th ed. Washington: American Psychiatric Association; 1994. p. 65–71.
7. Hendry CN. Childhood disintegrative disorder: should it be considered a distinct diagnosis? Clin Psychol Rev. 2000;20:77–90.
8. Rosman NP, Bergia BM. Childhood disintegrative disorder: distinction from autistic disorder and predictors of outcome. J Child Neurol. 2013;28:1587–98.
9. Mouridsen SE, Rich B, Isager T. A comparative study of genetic and neurobiological findings in disintegrative psychosis and infantile autism. Psychiatry Clin Neurosci. 2000;54:441–6.
10. Volkmar FR, Cohen DJ. Disintegrative disorder or "late onset" autism. J Child Psychol Psychiatry. 1989;30:717–24.
11. Volkmar FR, Rutter M. Childhood disintegrative disorder: results of the DSM-IV autism field trial. J Am Acad Child Adolesc Psychiatry. 1995;34:1092–5.
12. Fombone E. Prevalence of childhood disintegrative disorder. Autism. 2002;6:149–57.
13. Malhotra S, Gupta N. Childhood disintegrative disorder: re-examination of the current concept. Eur Child Adolesc Psychiatry. 2002;11:108–14.
14. Kurita H, Osada H, Miyake Y. External validity of childhood disintegrative disorder in comparison with autistic disorder. J Autism Dev Disord. 2004;34:355–62.
15. Homan KJ, Mellon MW, Houlihan D, Katusic MZ. Brief report: childhood disintegrative disorder: a brief examination of eight case studies. J Autism Dev Disord. 2011;41:497–504.
16. Barger BD, Campbell JM, McDonough JD. Prevalence and onset of regression within autism spectrum disorders: a meta-analytic review. J Autism Dev Disord. 2013;43:817–28.
17. Kurita H, Koyama T, Setoya Y, Shimizu K, Osada H. Validity of childhood disintegrative disorder apart from autistic disorder with speech loss. Eur Child Adolesc Psychiatry. 2004;13:221–6.
18. Baio J. Prevalence of autism spectrum disorder among children aged 8 years—autism and developmental disabilities monitoring network, 11 sites, United States, 2010. MMWR Surveill Summ. 2014;63:1–21.
19. Fombonne E. Epidemiology of pervasive developmental disorders. Pediatr Res. 2009;65:591–8.

20. Kang HJ, Kawasawa YI, Cheng F, Zhu Y, Xu X, Li M, et al. Spatio-temporal transcriptome of the human brain. Nature. 2011;478:483–9.
21. Tottenham N, Tanaka JW, Leon AC, McCarry T, Nurse M, Hare TA, et al. The NimStim set of facial expressions: judgments from untrained research participants. Psychiatry Res. 2009;168:242–9.
22. De Rubeis S, He X, Goldberg AP, Poultney CS, Samocha K, Cicek AE, et al. Synaptic, transcriptional and chromatin genes disrupted in autism. Nature. 2014;515:209–15.
23. Iossifov I, O'Roak BJ, Sanders SJ, Ronemus M, Krumm N, Levy D, et al. The contribution of de novo coding mutations to autism spectrum disorder. Nature. 2014;515:216–21.
24. Sanders SJ, He X, Willsey AJ, Ercan-Sencicek AG, Samocha KE, Cicek AE, et al. Insights into autism spectrum disorder genomic architecture and biology from 71 risk loci. Neuron. 2015;87:1215–33.
25. Hoischen A, Krumm N, Eichler EE. Prioritization of neurodevelopmental disease genes by discovery of new mutations. Nat Neurosci. 2014;17:764–72.
26. Ronemus M, Iossifov I, Levy D, Wigler M. The role of de novo mutations in the genetics of autism spectrum disorders. Nat Rev Genet. 2014;15:133–41.
27. Bunik V, Kaehne T, Degtyarev D, Shcherbakova T, Reiser G. Novel isoenzyme of 2-oxoglutarate dehydrogenase is identified in brain, but not in heart. FEBS J. 2008;275:4990–5006.
28. Allison T, Ginter H, McCarthy G, Nobre AC, Puce A, Luby M, et al. Face recognition in human extrastriate cortex. J Neurophysiol. 1994;71:821–5.
29. Kanwisher N, McDermott J, Chun MM. The fusiform face area: a module in human extrastriate cortex specialized for face perception. J Neurosci. 1997; 17:4302–11.
30. Pitcher D, Walsh V, Duchaine B. The role of the occipital face area in the cortical face perception network. Exp Brain Res. 2011;209:481–93.
31. Schultz RT, Gauthier I, Klin A, Fulbright RK, Anderson AW, Volkmar F, et al. Abnormal ventral temporal cortical activity during face discrimination among individuals with autism and Asperger syndrome. Arch Gen Psychiatry. 2000;57:331–40.
32. Pelphrey KA, Sasson NJ, Reznick JS, Paul G, Goldman BD, Piven J. Visual scanning of faces in autism. J Autism Dev Disord. 2002;32:249–61.
33. Klin A, Jones W, Schultz R, Volkmar F, Cohen D. Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism. Arch Gen Psychiatry. 2002;59:809–16.
34. Papagiannopoulou EA, Chitty KM, Hermens DF, Hickie IB, Lagopoulos J. A systematic review and meta-analysis of eye-tracking studies in children with autism spectrum disorders. Soc Neurosci. 2014;9:610–32.
35. Campbell DJ, Shic F, Macari S, Chawarska K. Gaze response to dyadic bids at 2 years related to outcomes at 3 years in autism spectrum disorders: a subtyping analysis. J Autism Dev Disord. 2014;44:431–42.
36. Norbury CF, Brock J, Cragg L, Einav S, Griffiths H, Nation K. Eye-movement patterns are associated with communicative competence in autistic spectrum disorders. J Child Psychol Psychiatry. 2009;50:834–42.
37. Lewkowicz DJ, Hansen-Tift AM. Infants deploy selective attention to the mouth of a talking face when learning speech. Proc Natl Acad Sci U S A. 2012;109:1431–6.
38. Norbury CF, Griffiths H, Nation K. Sound before meaning: word learning in autistic disorders. Neuropsychologia. 2010;48:4012–9.
39. Morton J, Johnson MH. CONSPEC and CONLERN: a two-process theory of infant face recognition. Psychol Rev. 1991;98:164–81.
40. Johnson MH. Autism: demise of the innate social orienting hypothesis. Curr Biol. 2014;24:R30–1.
41. Jones W, Klin A. Attention to eyes is present but in decline in 2–6-month-old infants later diagnosed with autism. Nature. 2013;504:427–31.
42. Tarpey PS, Smith R, Pleasance E, Whibley A, Edkins S, Hardy C, et al. A systematic, large-scale resequencing screen of X-chromosome coding exons in mental retardation. Nature. 2009;41:535–43.
43. Rauch A, Wieczorek D, Graf E, Wieland T, Endele S, Schwarzmayr T, et al. Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. Lancet. 2012;380:1674–82.
44. Hu H, Haas SA, Chelly J, Van Esch H, Raynaud M, de Brouwer AP, et al. X-exome sequencing of 405 unresolved families identifies seven novel intellectual disability genes. Mol Psychiatry. 2016;21:133–48.
45. Kirov G, Rees E, Walters JTR, Escott-Price V, Georgieva L, Richards AL, et al. The penetrance of copy number variations for schizophrenia and developmental delay. Biol Psychiatry. 2014;75:378–85.
46. Mulle JG, Dodd AF, McGrath JA, Wolyniec PS, Mitchell AA, Shetty AC, et al. Microdeletions of 3q29 confer high risk for schizophrenia. Am J Hum Genet. 2010;87:229–36.
47. Allen AS, Berkovic SF, Cossette P, Delanty N, Dlugos D, Eichler EE, et al. De novo mutations in epileptic encephalopathies. Nature. 2013;501:217–21.
48. Xu B, Ionita-Laza I, Roos JL, Boone B, Woodrick S, Sun Y, et al. De novo gene mutations highlight patterns of genetic and neural complexity in schizophrenia. Nat Genet. 2012;44:1365–9.
49. Fitzgerald TW, Gerety SS, Jones WD, van Kogelenberg M, King DA, McRae J, et al. Large-scale discovery of novel genetic causes of developmental disorders. Nature. 2015;519:223–8.
50. Ahn K, Gotay N, Andersen TM, Anvari AA, Gochman P, Lee Y, et al. High rate of disease-related copy number variations in childhood onset schizophrenia. Mol Psychiatry. 2014;19:568–72.
51. Lim ET, Raychaudhuri S, Sanders SJ, Stevens C, Sabo A, MacArthur DG, et al. Rare complete knockouts in humans: population distribution and significant role in autism spectrum disorders. Neuron. 2013;77:235–42.
52. Piton A, Gauthier J, Hamdan FF, Lafreniere RG, Yang Y, Henrion E, et al. Systematic resequencing of X-chromosome synaptic genes in autism spectrum disorder and schizophrenia. Mol Psychiatry. 2011;16:867–80.
53. Niranjan TS, Skinner C, May M, Turner T, Rose R, Stevenson R, et al. Affected kindred analysis of human X chromosome exomes to identify novel X-linked intellectual disability genes. PLoS One. 2015;10:e0116454.
54. Fromer M, Pocklington AJ, Kavanagh DH, Williams HJ, Dwyer S, Gormley P, et al. De novo mutations in schizophrenia implicate synaptic networks. Nature. 2014;506:179–84.
55. de Ligt J, Willemsen MH, van Bon BW, Kleefstra T, Yntema HG, Kroes T, et al. Diagnostic exome sequencing in persons with severe intellectual disability. N Engl J Med. 2012;367:1921–9.

Additional file 1: Supplementary Information

Supplementary methods

Subjects. This research was approved by the Yale University Institutional Review Board, and informed consent was obtained for all subjects. We studied four cohorts: (1) subjects with CDD (n = 17) who were referred to the Yale Child Study Center (YCSC), (2) low-functioning (FSIQ ≤ 75) subjects with ASD (LFASD, n = 12) and early-onset (< 2 years-old) delays, (3) high-functioning (FSIQ ≥ 75) subjects with ASD (HFASD, n = 50) and early-onset (< 2 years-old) delays, and (4) typically-developing subjects (TD, n = 26). A multidisciplinary (child psychiatrist, developmental-behavioral pediatrician, child psychologist, and speech/language pathologist) team of expert clinicians at the YCSC evaluated records for each child with suspected ASD to determine whether they met DSM-IV criteria for ASD (to remain consistent with the use of the DSM-IV to diagnose CDD). Clinical judgment was supplemented with the Autism Diagnostic Observation Schedule (ADOS) and the Autism Diagnostic Interview-Revised (ADI-R). Given the challenge of differentiating ASD and ID from ID alone in very low-functioning individuals and the limitations of current diagnostic instruments for this population, the differential diagnosis was made based on the consensus clinical judgment of experienced clinicians specializing in low-functioning individuals with ASD and ID and with ID alone. In addition, for suspected CDD, expert clinicians (child psychiatrist, developmental-behavioral pediatrician, child neurologist, and/or child psychologist) conducted a comprehensive interview with parents and viewed any available home videos to characterize the nature of each subject's possible regression. Besides meeting full criteria for ASD, CDD was defined as: (1) Loss of language skills after previous period of typical development. The child must have used at least two-word phrases daily and spontaneously prior to the regression. We required language loss, as it is the most objective and quantifiable domain as reported by parents. (2) Loss in at least one other domain: social skills or adaptive behavior, bowel or bladder control, play, and motor skills (reflective of the DSM-IV criteria for CDD). (3) Loss must have occurred after 24 months of age. (4) The child must not have regained level of skill prior to loss. (5) FSIQ ≤ 75. Although loss of cognitive ability was not part of the DSM-IV criteria for CDD, severe regression is most noticeable in the cognitive domain, and, in practice, we only diagnose CDD when there is comorbid ID.

DNA samples. Genomic DNA was prepared from 15 families affected by CDD (Additional file 2: Table S1). Both biological parents were available for all probands as well as 13 unaffected siblings. DNA was extracted from blood for all families except CDD07, for which only lymphoblastoid cell line DNA was available. For CDD21 and CDD22, only whole-genome amplified DNA from blood was available.

Whole-exome sequencing and quality control. All DNA samples were sequenced at the Yale Center for Genome Analysis. Exonic sequences were selected by the NimbleGen v2.0 exome capture reagent (Roche, Basel, Switzerland) and sequenced on the HiSeq 2000 (75 bp paired-end reads; Illumina, San Diego, CA, USA). Reads were mapped to the human reference genome (hg19) using CASAVA v1.8 (ELAND v2). Quality metrics are shown in Additional file 2: Table S17, indicating high quality whole-exome sequencing data. Single nucleotide variants and small insertions/deletions (indels) were identified and assigned quality scores (QS) using SAMtools (http://samtools.sourceforge.net/). All variants were annotated for impact on the encoded protein (synonymous, missense, nonsense, splice site, frameshift) and frequency using dbSNP141/1000 Genomes (May 2011 release), NHLBI GO ESP Exome Variant Server (ESP6500SI-V2), and an in-house database of 2500 exomes. Each exome matched the recorded sex of the subject. Family relationships were validated using an in-house Perl script comparing the overlap of novel heterozygous variants between members of each family. All reported family structures were confirmed.

Ancestry mapping. EIGENSTRAT (http://genepath.med.harvard.edu/~reich/EIGENSTRAT.htm) was used to compare SNP genotypes of CDD family members to individuals of known ancestry in HapMap3 (http://hapmap.ncbi.nlm.nih.gov/). 1,824 SNPs from whole-exome sequencing data were pre-defined: (1) minor allele frequency (MAF) > 5%, (2) not in significant linkage disequilibrium with other SNPs analyzed,

(3) 100 kb apart, (4) not in a region of segmental duplication, (5) satisfy Hardy-Weinberg equilibrium ($P < 0.001$), and (6) contain high Fst values (different frequencies across major ethnic groups in HapMap3). Eigenvalues of the first two principal components, which contributed the greatest amount of variation relative to the other principal components, were plotted against each other (Additional file 1: Fig. S3). The principal component analysis correctly separated and distinguished ancestry groups in HapMap3 samples and confirmed the self-reported race and ethnicities of the subjects (Additional file 2: Table S1).

Genotyping. Subjects were genotyped using the HumanOmni2.5M BeadChip (Illumina). All DNA samples were hybridized and scanned simultaneously on the Illumina iScan to minimize batch effects and variation. All subjects had a genotyping call rate > 95%. Genotyping data were analyzed by PLINK v1.07 [1] (http://pngu.mgh.harvard.edu/~purcell/plink/) and confirmed the recorded sex and family relationships of each subject.

***De novo* and inherited sequence variant detection.** Three types of rare [novel or found at most once across 1000 Genomes (May 2011 release), NHLBI GO ESP Exome Variant Server (ESP6500SI-V2), and in-house database of 2500 exomes] protein-changing variants from whole-exome sequencing were prioritized for study: (1) *de novo*, (2) homozygous, and (3) hemizygous (mother-to-son transmission on chrX). The Exome Aggregation Consortium (ExAC) Browser was not used to filter out variants by frequency since approximately 23% of subjects come from neuropsychiatric studies (exac.broadinstitute.org/faq); still, all of our genotypes of interest (*de novo*, homozygous, or hemizygous) have frequency $\leq 0.36\%$ in this database. All *de novo* variants were confirmed by Sanger sequencing in both forward and reverse directions. All homozygous and hemizygous variants at $\leq 1\%$ general population frequency that were unique to probands (not shared by unaffected siblings) were visualized by in-silico inspection and/or Sanger validated. Sequencing chromatograms were aligned and analyzed using Sequencher v4.9 (Gene Codes, Ann Arbor, MI, USA).

De novo variants were identified using a Bayesian algorithm as previously described [2]. Virtually 100% of *de novo* variants with a Bayesian quality score (BQS) ≥ 50 validate by Sanger sequencing [2]. For the purposes of comparing *de novo* rates between probands and siblings, only variants with a BQS ≥ 50 were considered. To maximize our discovery of *de novo* variants in probands, all variants with a BQS ≥ 1 were both inspected computationally by the visualization of plot reads and by Sanger sequencing. 100% (5/5) of *de novo* variants with a BQS ≥ 50 and 11% (2/18) with a BQS between 1 and 50 were confirmed.

Homozygous variants were required to have: (1) SAMtools QS ≥ 60 (94% of such variants confirm by Sanger sequencing in our experience), (2) heterozygous genotypes (SAMtools QS ≥ 60) in both parents, and (3) the homozygous genotype seen at most once in 1000 Genomes (May 2011 release) and NHLBI GO ESP Exome Variant Server (ESP6500SI-V2).

In male probands, hemizygous variants (mother-to-son transmission on chrX) were required to have: (1) SAMtools QS ≥ 100, (2) the proband's father was required to have the hemizygous reference genotype and the proband's mother was required to have a heterozygous genotype with SAMtools QS ≥ 100, and (3) the hemizygous (in males) and homozygous (in females) genotypes seen at most once in 1000 Genomes (May 2011 release) and NHLBI GO ESP Exome Variant Server (ESP6500SI-V2).

Copy number variant (CNV) detection. Three types of rare (novel or seen at most once in the Database of Genomic Variants) CNVs from genotyping data were prioritized for study: (1) *de novo*, (2) homozygous, and (3) hemizgyous (mother-to-son transmission on chrX). CNV detection was performed using three algorithms, PennCNV Revision 220, QuantiSNP v1.1, and GNOSIS, as previously described [3]. PennCNV and QuantiSNP are based on the Hidden Markov Model. GNOSIS uses a continuous distribution function to fit the intensity values from the HapMap data and determine thresholds for significant points in the tails of the distribution that are used to detect copy number changes. Analysis and merging of the CNV predictions were performed using an in-house Perl script. All rare ($\geq 50\%$ of CNV at $\leq 1\%$ frequency in the Database of Genomic Variants) genic CNVs that were unique to probands (not shared by unaffected siblings) and predicted by at least PennCNV and QuantiSNP were tested by quantitative PCR (qPCR) as

previously described [3].

Gene expression levels. Gene-level brain expression data (Platform GPL5175, Affymetrix GeneChip Human Exon 1.0 ST Array) [4], which were generated as part of the BrainSpan project (www.hbatlas.org), were downloaded from the NCBI GEO database (accession number GSE25219) in the form of $\log_2$-transformed signal intensity values. Affymetrix uses background probes with matching GC content for background correction for all probes on the array (http://media.affymetrix.com/support/technical/whitepapers/exon_background_correction_whitepaper.pdf). Genes represented once in the core probe set were identified for the following groups: (1) all nonsynonymous variants (n = 40) in CDD probands (n = 15), (2) all synonymous variants (n = 16) in CDD probands (n = 15), (3) all nonsynonymous variants (n = 17) in unaffected siblings (n = 13) of CDD probands, (4) all synonymous variants (n = 8) in unaffected siblings (n = 13) of CDD probands, (5) *de novo* nonsynonymous variants (n = 123) in SSC probands with regression (n = 257) [5], (6) *de novo* synonymous variants (n = 37) in SSC probands with regression (n = 257) [5], (7) *de novo* nonsynonymous variants (n = 132) in SSC probands without regression (n = 249) [5], (8) *de novo* synonymous variants (n = 52) in SSC probands without regression (n = 249) [5], (9) *de novo* nonsynonymous variants (n = 1526) in SSC probands (n = 2508) [5], (10) *de novo* synonymous variants (n = 503) in SSC probands (n = 2508) [5], (11) *de novo* nonsynonymous variants (n = 1044) in unaffected siblings (n = 1911) of SSC probands [5], (12) *de novo* synonymous variants (n = 389) in unaffected siblings (n = 1911) of SSC probands [5], (13) *de novo* LGD variants (n = 297) in SSC probands (n = 2508) [5], (14) *de novo* LGD variants (n = 156) in unaffected siblings (n = 1911) of SSC probands [5], (15) highest-risk genes (n = 67) in SSC, ASC, and AGP probands (n = 8009) [5-7], and (16) all genes in the BrainSpan dataset (n = 16947) [4] (Additional file 2: Table S4). Since the published ASD candidate genes identified by WES and CNV studies [5-7] were not filtered by positive brain expression as determined by BrainSpan, we did not filter by this parameter either across the 16 groups to maintain consistency. To identify SSC probands with and without regression, we queried the SSC v.14 Phenotype Data Set (https://sfari.org/resources/sfari-base). SSC probands with regression were defined as individuals who received maximal scores on two questions from the ADI-R: (1) Question #11 loss of language skills after acquisition: Were you ever concerned that [subject] might have lost language skills during the first years of her/his life? Was there ever a time that s/he stopped speaking for some months after having learned to talk? (0=No, 1=Yes) and (2) Question #20 loss of skills (for at least 3 months): Has there ever been a period when [subject] seemed to get markedly worse or dropped further behind in her/his development? (0=no consistent loss of skills, 1=probable loss of skill but of a degree that falls short of specified criteria, 2=account of definite loss of skills over a period of time). Therefore, SSC probands with regression were defined as individuals with a total score of 3. SSC probands without regression were defined as individuals who received scores of 0 on both questions. SSC probands with and without regression were matched by sex, age at evaluation, IQ, and autism symptom severity (Additional file 2: Table S18). Equality of variances was determined by Levene's test.

The median expression value for genes affected by nonsynonymous variants in CDD probands and represented once in the core probe set (n = 40) across all brain samples was determined and plotted using ggplot2 in R for each brain region (NCX, HIP, AMY, STR, MD, CBC) and time period. The difference in median expression values between non-neocortical (HIP, AMY, STR, MD, CBC) and neocortical (NCX) brain regions for all 16 gene groups described above was also plotted using ggplot in R for each time period. Local polynomial regression fitting was used to smooth the scatter plots. This difference reached a maximum value at period 6 for genes affected by nonsynonymous variants in CDD probands (Fig. 2). Permutation testing with 100,000 iterations was performed to determine the significance of this difference. Random sets of 40 genes were selected from the BrainSpan dataset, and the differential expression between non-neocortical and neocortical brain regions was calculated at period 6. The P value was determined by the number of times the differential expression was greater than or equal to the difference observed for CDD candidate genes.

Gene coexpression analysis. For the 40 CDD candidate genes, a gene coexpression matrix was constructed

using the mean expression value for each gene in each brain region for each brain sample and calculating the Pearson correlation coefficient for each pairwise gene combination across all data points. Two genes were considered coexpressed if they had a correlation coefficient $r \geq 0.7$. The number of genes that were coexpressed with at least one other gene from the set as well as the number of correlations/gene and the mean coefficient value were determined. Permutation testing with 100,000 iterations of 40 random genes from the BrainSpan dataset was performed to determine the significance of these values. The *P* value was calculated by the number of iterations that resulted in a greater or equal number of genes being coexpressed, correlations/gene, and mean coefficient value. To visualize the gene coexpression network, edges were drawn between two genes if their correlation coefficient $r \geq 0.7$, using the organic layout function of Cytoscape [8]. Positive correlations are shown in blue, and negative correlations are shown in red. The greater the magnitude of the coefficient, the wider and darker are the edges. The size of a node is proportional to the number of edges the node has.

Non-sedated fMRI data acquisition and paradigm. Images were collected on a Siemens 3T Tim Trio scanner located at the Yale University Magnetic Resonance Research Center. High-resolution, T1-weighted anatomical images were acquired using a magnetization-prepared rapid gradient echo (MPRAGE) sequence (TR = 1,900 ms, TE = 2.96 ms, flip angle = 9°, matrix = 256×256, voxel size = 1x1 mm^2, field of view = 256×256 mm^2, slice thickness = 1.00 mm, 160 slices, interleaved acquisition). Whole-brain functional images were acquired using a single-shot, gradient-recalled echo planar pulse sequence (TR = 2,000 ms, TE = 25 ms, flip angle = 60°, matrix = 64×64, voxel size = 3.44×3.44 mm^2, field of view = 220×220 mm^2, slice thickness = 4.00 mm, 34 slices, interleaved acquisition) sensitive to blood oxygenation level-dependent (BOLD) contrast. The fMRI task consisted of ten 12 s blocks of images (5 blocks of fearful faces and 5 blocks of houses). The blocks, consisting of either faces or houses, alternated, with the fearful faces presented first. Each block included six images and each image was presented for 2 s.

fMRI data collection, processing, and analyses. The neurobiology of CDD has not been elucidated in part due to the technical difficulty of conducting experimental protocols with very low-functioning subjects. To obtain this data, we implemented an individualized training protocol to accustom subjects to the scanner environment as well as to provide training and reinforcement for compliance with the requirement to remain very still during fMRI and eye tracking. We utilized the following training procedures: (1) preparation for scanning and eye tracking through videos sent home before the visits; (2) preparation for use of earphones and earplugs in the scanner by sending home earphones and earplugs and asking parents to help their children learn to wear them properly for increasing periods of time; (3) providing a list of "statue" and "let's-take-a-picture" games for parents to engage their children in at home before and between the training protocol sessions, to help children learn to "pretend to be a statue"/"pretend to have a picture taken," involving earning rewards for holding still for increasing lengths of time; (4) gradual introduction to experimental procedures through interaction with, first, a "toy" scanner used on a stuffed animal, then a mock scanner before entering the scanning environment; (5) helping subjects become familiar and comfortable with the pictures to be presented in the scanner by providing analogues of all stimuli, first on a tabletop, then in a mock scanner before moving to the scanner; (6) using picture schedules to accompany mock scanner sessions and as reminders prior to scanning and eye-tracking sessions; (7) utilizing visual transition signals between "statue"/"picture taking" and "move" conditions; (8) providing comforting activities and rewards to assist children in overcoming distress, along with parental support. Approximately 70% of low-functioning subjects who were able to progress to the real scanner were able to give usable data. Across all cohorts, no subject had an active seizure disorder, since this is an exclusion criterion for our MRI studies. Subjects were not taking medications that are known to affect the fMRI BOLD signal.

Data were processed and analyzed using FEAT v6.00 (FMRI Feat Analysis Tool) of FSL 5.0.6, via a data processing pipeline implemented in the Yale University High-Performance Computing clusters. The pipeline consisted of: (1) motion correction using MCFLIRT, (2) interleaved slice timing correction, (3) BET brain extraction, (4) spatial smoothing using a kernel of FWHM 5 mm, and (5) high-pass temporal filtering using 100 s. The first and last 10 volumes were fixation (no images were presented). The remaining

60 volumes were analyzed.

The EPI data was registered to the subject's structural scan (with the brain extracted using BET) via linear boundary-based registration and then registered to the MNI152 standard brain with linear transformation with 12 degrees of freedom. Artifact removal was performed with FSL's FIX (FMRIB's ICA-based Xnoiseifier). The standard denoising classifier from FSL's FIX package was applied to the raw results from FSL's MELODIC ICA (Independent Component Analysis) to identify artifact components such as head movement, respiratory motion, and scanner artifacts. General Linear Model (GLM)-based analyses, where normality is assumed, were conducted for each subject to assess task-related BOLD responses. We did not include motion regressors in GLM to avoid over (duplicated) correction of head movement. To create predictors for fearful faces and houses conditions, the timing of the corresponding blocks (onset in seconds, duration = 12 s, weighting = 1) was convolved with the default gamma function (phase = 0 s, standard deviation = 3 s, mean lag = 6 s) with temporal derivatives. Time series autocorrelation was estimated using FSL's FILM pre-whitening. Due to the limited number of low-functioning subjects with usable fMRI data, we conservatively limited the statistical inference to our data only and used fixed-effects analysis in the fMRI group analysis [9]. The subject-level parameter estimates were inputs for the group-level fMRI analyses. Correction for multiple comparisons was implemented with a highly stringent voxel-level threshold of $Z > 3.09$ ($P < 0.001$, one-sided) and a cluster-level threshold of $P < 0.05$ for the main whole-brain analysis, or $Z > 2.58$ ($P < 0.01$, two-sided) for the analysis within FFG and a cluster-level threshold of $P < 0.05$.

Eye-tracking. Eye-tracking data were collected using a Tobii T60 XL monitor-integrated eye tracker. Subjects sat in front of a computer monitor and viewed static images of emotional faces. The images were photographs of 14 adult male and female actors centered on a neutral backdrop (extracted from the NimStim Face Stimulus set) [10] making three different expressions: happy, fearful, or neutral. All stimuli were grayscale, with the mean luminosity of each image adjusted to equal 80% of maximal brightness using Adobe Photoshop. Images were 506 pixels (11.4 degrees) wide and 649 pixels (14.6 degrees) high. Subjects first saw a white fixation cross, centered in the screen for 4 seconds, followed by images from the stimulus set lasting 2 seconds, alternating with a screen with a fixation cross only, lasting either 2 or 3 seconds. The location of the fixation cross varied, appearing at any of the corners of the screen. The purpose of the variations in location and duration of the fixation cross was to encourage subjects to alter the part of the screen they were looking at between stimuli images, so that when they were faced with a new stimulus they would have to reorient their eyes. Faces appeared for a total of 84 s during the paradigm. The experiment was administered two times over consecutive sessions for each subject. Regions of interest for the eyes and mouth were manually defined on the face images and were equal in size across all images of faces.

Approximately 80% of low-functioning subjects who were able to sit in front of the monitor gave usable data. Valid trials were defined as those for which data retention was $\geq$ 50%, with data retention calculated by dividing the number of eye-tracking samples that were identified by Tobii Studio as valid by the total number of samples over which the stimulus was presented. The variables of primary interest were total fixation duration on the image (TFD) as well as the amount of time spent looking at eyes and mouth, separately. In order to adjust for the probability that not all subjects would look at the face images for an equal amount of time, analyses of eye and mouth time were based on the ratio of time spent looking at eyes to the overall face (%Eye), and the ratio of time spent looking at mouth to the overall face (%Mouth). Statistical analyses involved an analysis of variance (ANOVA) approach, with independent sample t-tests for subsequent planned comparisons. Equality of variances was tested using Levene's Test, with subsequent Welch-Satterthwaite correction for degrees of freedom as appropriate. Calibration and calibration quality checks were conducted using Tobii's in-software verification tools (Tobii Studio) with a standard 5-point calibration.

Gaze heat maps were constructed with MATLAB scripts that provided visualization of group-level gaze data overlaid on the images presented to subjects. For each presented stimulus, the associated gaze-points upon that stimulus were collected from all subjects in a given group and spatial smoothed using a Gaussian filter with an approximately 2×2 degree kernel window, which had a standard deviation of

approximately 0.5 degrees.

Statistical methods. *Sample size.* No statistical methods were used to predetermine sample sizes. CDD, LFASD, HFASD, and TD cohort sizes represent the maximum number of subjects who could be recruited and give successful fMRI and/or eye-tracking data. SSC probands with and without regression cohort sizes represent the maximum number of subjects who met ADI-R criteria for regression or no regression, for whom WES data were available, and who could be matched by sex, age at study, IQ, and autism symptom severity.

Genetics. Statistical analysis was performed in R (version 3.2.0). Mutation burden analysis between CDD probands and their unaffected siblings was performed using Fisher exact test (Additional file 2: Table S3). Permutation testing with 100,000 iterations was performed to determine the significance of differential gene expression between non-neocortical and neocortical brain regions for CDD candidate genes (Fig. 2) and for the co-expression analysis (Fig. 3). Equality of variances for SSC probands with and without regression was determined by Levene's test. They were matched by sex, age at study, IQ, and autism symptom severity, as determined by the Fisher exact test or independent t-test as appropriate (Additional file 2: Table S18). All *P* values are two-tailed.

Neuroimaging. Cohorts were compared by sex, age at study, IQ, autism symptom severity, intracranial volume, and relative head motion in the scanner, as determined by chi-square, one-way ANOVA, or independent t-test as appropriate (Additional file 2: Table S9). fMRI analyses involved the commonly employed standard parametric GLM approach in FSL (version 5.0.6), where normality is assumed. Due to the limited number of low-functioning subjects with usable fMRI data, we conservatively limited the statistical inference to our data only and used fixed-effects analysis in the fMRI group analysis [9]. The subject-level parameter estimates were inputs for the group-level fMRI analyses. Correction for multiple comparisons was implemented with a highly stringent voxel-level threshold of $Z > 3.09$ ($P < 0.001$, one-sided) and a cluster-level threshold of $P < 0.05$ for the main whole-brain analysis, or $Z > 2.58$ ($P < 0.01$, two-sided) for the analysis within FFG and a cluster-level threshold of $P < 0.05$. The independent t-test was used for subsequent planned comparisons, and *P* values are two-tailed. All bar graphs show mean and standard error of the mean.

Eye-tracking. Cohorts were compared by sex, age at study, IQ, autism symptom severity, and total fixation duration on the image, as determined by chi-square, one-way ANOVA, or independent t-test as appropriate (Additional file 2: Table S16). Statistical analyses involved an analysis of variance (ANOVA) approach, with independent sample t-tests for subsequent planned comparisons (Additional file 2: Table S16). Equality of variances were tested using Levene's Test, with subsequent Welch-Satterthwaite correction for degrees of freedom as appropriate. All *P* values are two-tailed. All bar graphs show mean and standard error of the mean.

Supplementary references

1. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81:559-75.
2. Zaidi S, Choi M, Wakimoto H, Ma L, Jiang J, Overton JD, et al. De novo mutations in histone-modifying genes in congenital heart disease. Nature. 2013;498:220-3.
3. Sanders SJ, Ercan-Sencicek AG, Hus V, Luo R, Murtha MT, Moreno-De-Luca D, et al. Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. Neuron. 2011;70:863-85.
4. Kang HJ, Kawasawa YI, Cheng F, Zhu Y, Xu X, Li M, et al. Spatio-temporal transcriptome of the human brain. Nature. 2011;478:483-9.
5. Iossifov I, O'Roak BJ, Sanders SJ, Ronemus M, Krumm N, Levy D, et al. The contribution of de novo coding mutations to autism spectrum disorder. Nature. 2014;515:216-21.
6. De Rubeis S, He X, Goldberg AP, Poultney CS, Samocha K, Cicek AE, et al. Synaptic, transcriptional and chromatin genes disrupted in autism. Nature. 2014;515:209-15.

7. Sanders SJ, He X, Willsey AJ, Ercan-Sencicek AG, Samocha KE, Cicek AE, et al. Insights into autism spectrum disorder genomic architecture and biology from 71 risk loci. Neuron. 2015;87:1215-33.
8. Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, Workman C, et al. Integration of biological networks and gene expression data using Cytoscape. Nature Protoc. 2007;2:2366-82.
9. Desmond JE, Glover GH. Estimating sample size in functional MRI (fMRI) neuroimaging studies: statistical power analyses. J Neurosci Methods. 2002;118:115-28.
10. Tottenham N, Tanaka JW, Leon AC, McCarry T, Nurse M, Hare TA, et al. The NimStim set of facial expressions: judgments from untrained research participants. Psychiatry Res. 2009;168:242-9.

Supplementary figures

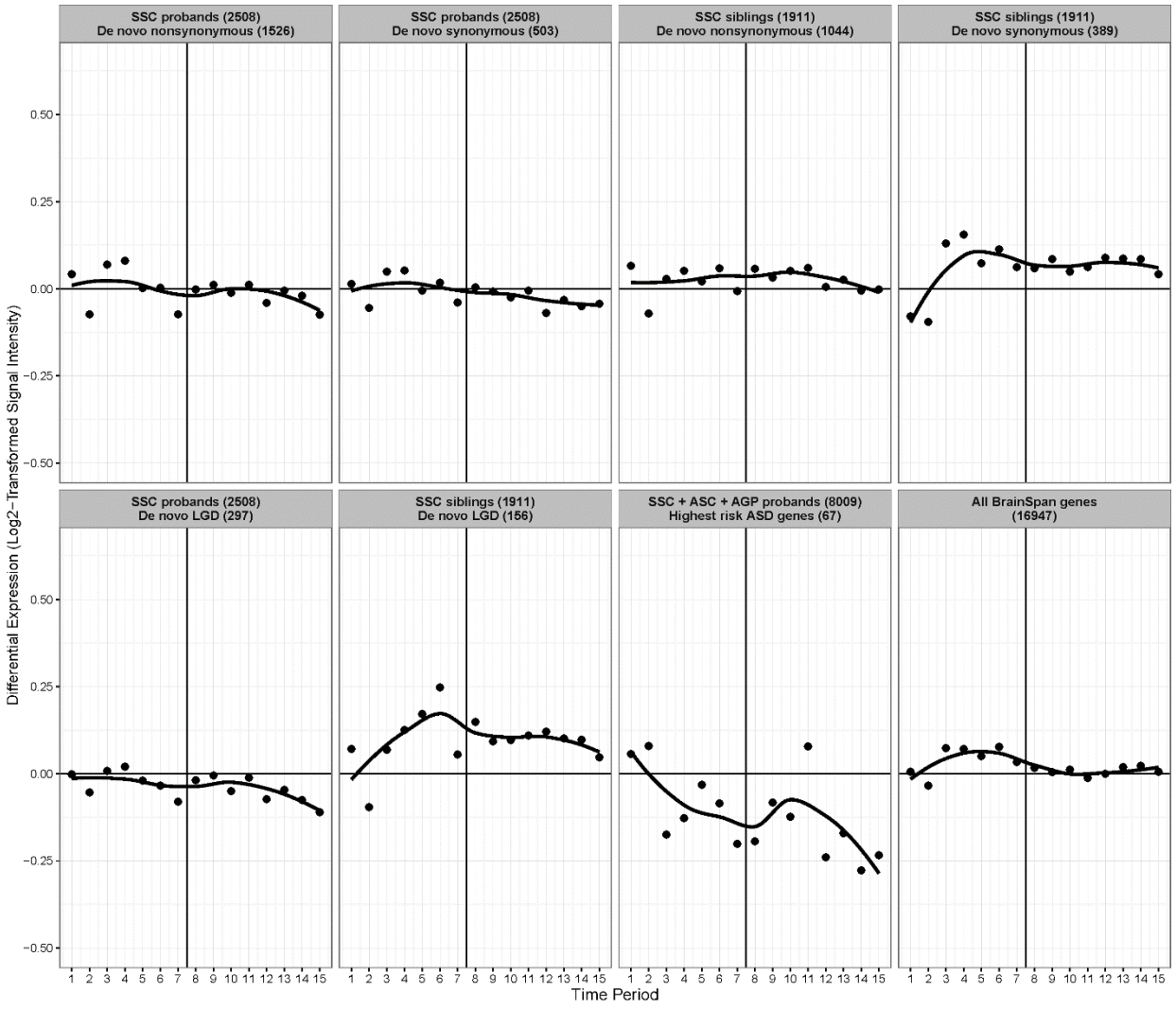

Fig. S1 Differential expression levels of various gene sets. The difference in expression levels (non-neocortical minus neocortical brain regions) is shown for genes affected by nonsynonymous, synonymous, and LGD variants in SSC probands and their unaffected siblings. Also plotted are data for genes most significantly associated with ASD by three recent, large WES and CNV studies [5-7] and all genes in the BrainSpan dataset [4]. The dark vertical line in each panel indicates birth. The number in parentheses indicates the number of subjects or variants. AGP, Autism Genome Project; ASC, Autism Sequencing Consortium; LGD, likely gene disrupting; SSC, Simons Simplex Collection.

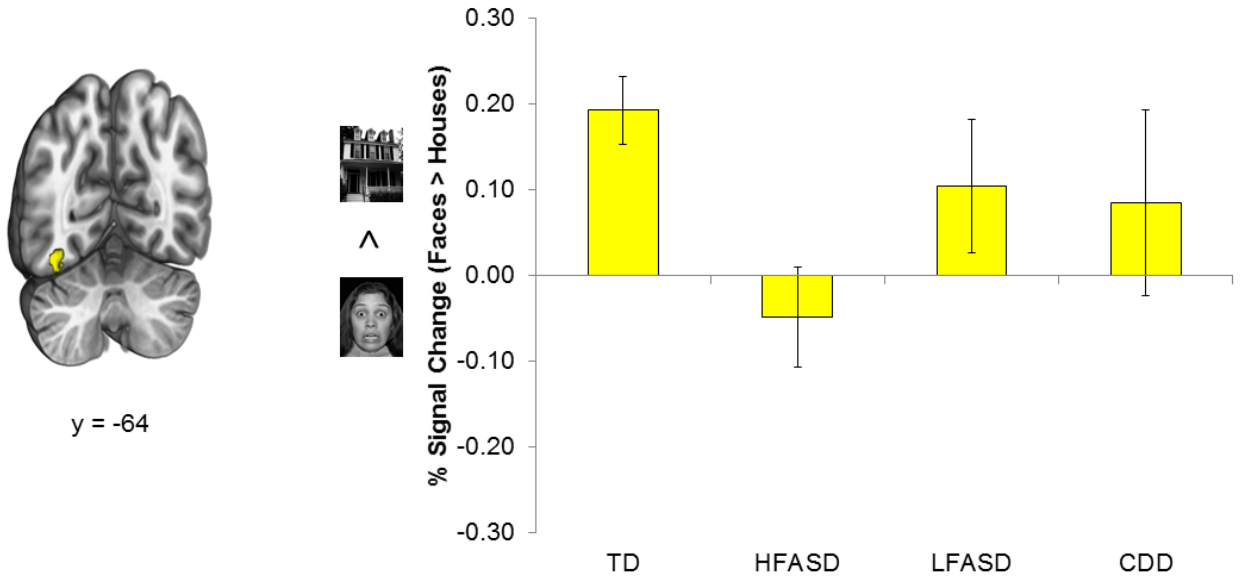

Fig. S2 Comparison of fMRI faces > houses activity in a region within the middle fusiform gyrus (FFG) corresponding to the expected location of the fusiform face area. This region is defined by TD > HFASD in FFG. Left: The yellow color brain map indicates TD > HFASD activity within the FFG (an example slice taken at MNI152 y = -64 mm, 50 voxels), $Z > 2.58$, whole-brain corrected at the cluster-level $P < 0.05$. Right: mean % signal change from the faces > houses contrast within the region of TD > HFASD in the FFG [$t(31) = 3.54$, $P = 0.0013$, Cohen's $d = 1.29$] by all groups: TD ($n = 19$), HFASD ($n = 14$), LFASD ($n = 7$), and CDD ($n = 7$). Comparison of CDD relative to TD revealed no significant difference [$t(24) = 1.18$, $P = 0.25$, Cohen's $d = 0.54$], as did LFASD relative to TD [$t(24) = 1.10$, $P = 0.28$, Cohen's $d = 0.51$]. Error bars indicate standard error of the mean. All P values were calculated by independent t-test and are two-tailed.

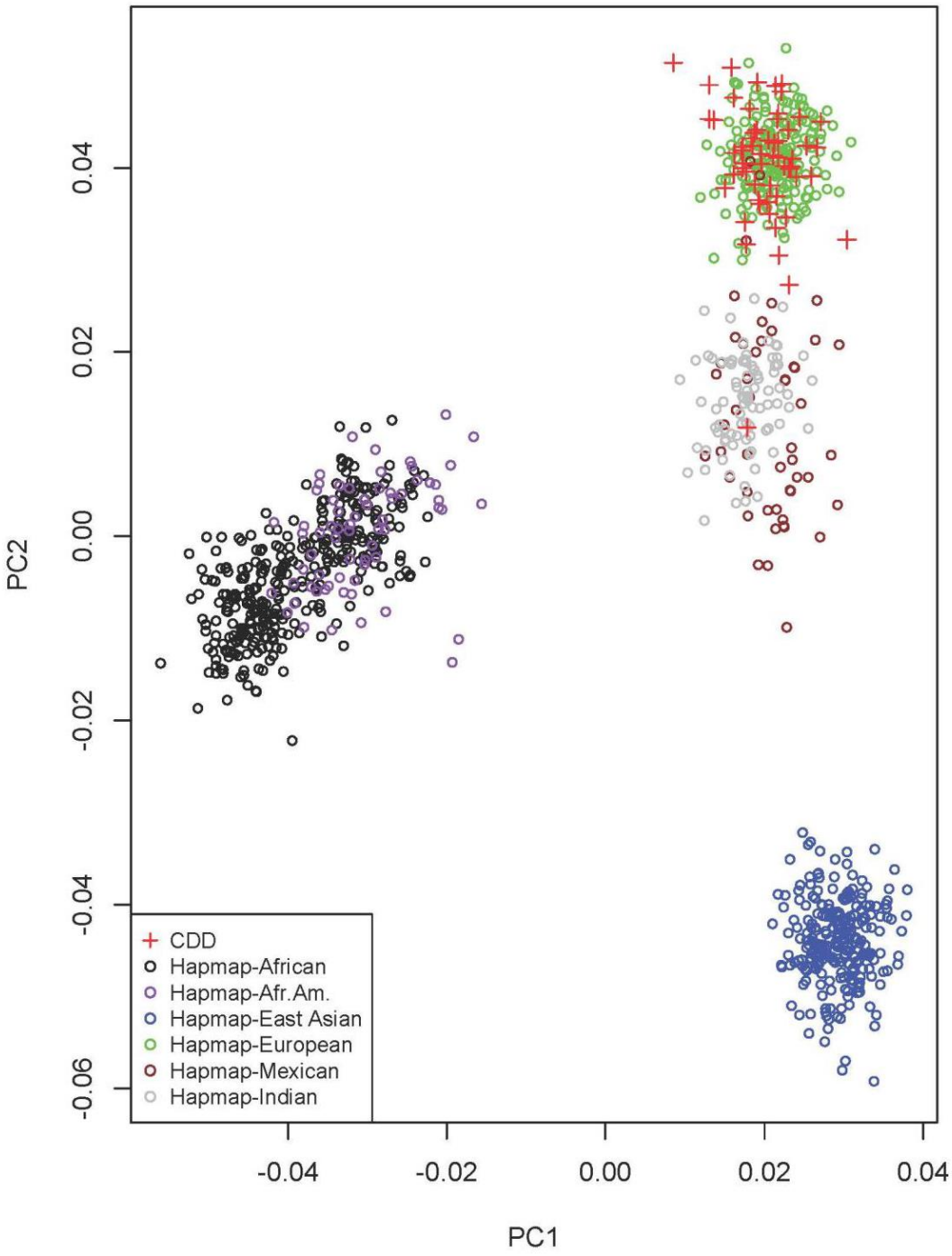

Fig. S3 EIGENSTRAT was used to compare SNP genotypes of CDD family members to individuals of known ancestry in HapMap3. Eigenvalues of the first two principal components (PC1 and PC2), which contributed the greatest amount of variation relative to the other principal components, were plotted against each other.