Human Molecular Genetics, 2013, Vol. 22, No. 14 **2960–2972**
doi:10.1093/hmg/ddt150
Advance Access published on April 10, 2013

An eQTL mapping approach reveals that rare variants in the *SEMA5A* regulatory network impact autism risk

Ye Cheng, Jeffrey Francis Quinn and Lauren Anne Weiss*

Department of Psychiatry and Institute for Human Genetics, University of California, San Francisco, San Francisco, CA, USA

Received October 8, 2012; Revised February 5, 2013; Accepted March 29, 2013

To date, genome-wide single nucleotide polymorphism (SNP) and copy number variant (CNV) association studies of autism spectrum disorders (ASDs) have led to promising signals but not to easily interpretable or translatable results. Our own genome-wide association study (GWAS) showed significant association to an intergenic SNP near Semaphorin 5A (*SEMA5A*) and provided evidence for reduced expression of the same gene. In a novel GWAS follow-up approach, we map an expression regulatory pathway for a GWAS candidate gene, *SEMA5A*, *in silico* by using population expression and genotype data sets. We find that the *SEMA5A* regulatory network significantly overlaps rare autism-specific CNVs. The *SEMA5A* regulatory network includes previous autism candidate genes and regions, including *MACROD2*, *A2BP1*, *MCPH1*, *MAST4*, *CDH8*, *CADM1*, *FOXP1*, *AUTS2*, *MBD5*, 7q21, 20p, *USH2A*, *KIRREL3*, *DBF4B* and *RELN*, among others. Our results provide: (i) a novel data-derived network implicated in autism, (ii) evidence that the same pathway seeded by an initial SNP association shows association with rare genetic variation in ASDs, (iii) a potential mechanism of action and interpretation for the previous autism candidate genes and genetic variants that fall in this network, and (iv) a novel approach that can be applied to other candidate genes for complex genetic disorders. We take a step towards better understanding of the significance of *SEMA5A* pathways in autism that can guide interpretation of many other genetic results in ASDs.

INTRODUCTION

Autism spectrum disorders (ASDs) are neurodevelopmental diseases that affect social and communication skills as well as developmentally appropriate behaviors with an onset in early childhood and are strongly genetic. The genetic basis for autism is supported by twin studies (1–5) and familial correlation in autism-related traits (6–9). Several genome-wide association studies (GWASs) have identified common single nucleotide polymorphism (SNP) and rare copy number variant (CNV) association signals (10–23). However, it is still a challenge to identify the autism susceptibility genes underlying these associations. SNP association signals often fall into intergenic chromosomal regions or show linkage disequilibrium (LD) extending across multiple variants or genes. CNVs associated with disease often contain multiple genes, among which it has been difficult to identify which gene(s) affect risk for ASDs. Limited sample size puts studies at low power to detect any given locus of modest effect size (as are common in replicated GWAS signals) or to assign pathogenicity to any given rare or unique CNV (of which excess numbers are present in neurodevelopmental disease).

Functional enrichment analysis has been performed on several large data sets, but to date the broad categories implicated, such as ubiquitinylation, microtubule cytoskeleton, glycosylation, CNS development/adhesion, cellular proliferation, projection and motility, and GTPase/Ras signaling (18,23), have neither shown consistent signals across data sets nor led to changes in our interpretation of additional genetic signals. Traditional pathway analyses have several major limitations. First, annotation is based on prior knowledge, which is incomplete in the area of developmental neurobiology as well as biased by well-studied processes. It is designed to identify a novel association with a known entity but not to

*To whom correspondence should be addressed at: Langley Porter Psychiatric Institute, Nina Ireland Lab, Box F-0984, 401 Parnassus Avenue, Rm. A101, San Francisco, CA 94143-0984, USA. Tel: +1 415 476 7650; Fax: +1 415 476 7389; Email: lauren.weiss@ucsf.edu

generate novel networks. Second, the categories implicated in autism to date, such as synaptic genes, are often so broad [e.g. more than 1000 synapse proteins are identified in mammals (24)] as to make it unlikely that any given gene in the pathway or category would have a high suspicion of influencing risk. In order to overcome these limitations, we propose a data-driven network analysis using rich expression and genetic resources and application of these results to disease data sets of interest for ASDs.

The genome-wide significant signal in our previous GWAS on the Autism Genetic Resource Exchange and National Institute for Mental Health (AGRE-NIMH) multiplex family sample was with an intergenic SNP ~80 kb upstream of the *SEMA5A* gene (17). The *SEMA5A* gene encodes an axonal guidance protein important in neurodevelopment (25). We and others previously showed down-regulation of *SEMA5A* expression in lymphoblastoid cell line (LCL), blood and brain samples of individuals with ASDs compared with controls (17,26). However, the associated SNP was not sufficient to explain the consistently reduced expression of *SEMA5A*, as this SNP has a low minor allele frequency. We therefore hypothesized that additional regulators of *SEMA5A* might exist, and if so, could be ASD susceptibility candidates.

In this study, we have examined our hypothesis using *in silico* expression quantitative trait locus (eQTL) mapping, or identification of genetic loci associated with *SEMA5A* expression levels, to define an empirical genetic regulatory network for *SEMA5A*. This includes both primary *SEMA5A* eQTL regions and secondary eQTL ('eQTL^2') master regulatory regions. Subsequent permutation-based analyses were used to test whether the *SEMA5A* regulatory network (as a whole) is associated with ASDs in large genome-wide data sets. Our approach provides a robust way to find novel susceptibility genes and a network contributing to complex disease with heterogeneous causes, such as ASDs. This novel data-derived network can inform our understanding of the pathophysiology of ASDs, as well as aid in interpretation of past and future genetic data.

RESULTS

SEMA5A eQTL mapping

Our previous genome-wide study identified an intergenic SNP near *SEMA5A* associated with autism as well as reduced expression of *SEMA5A* in autism. However, the associated SNP (and its LD proxies) near the *SEMA5A* locus on chromosome 5p15 could not explain the reduced expression of *SEMA5A*. Therefore, we sought to identify other genetic regulators of *SEMA5A* expression that might be important in ASD susceptibility. First, we mapped eQTLs for *SEMA5A* in a control (CEU) LCL expression and genetic data set. Using SCAN (www.scandb.org) (27), we identified 12 SNPs near the *SEMA5A* locus on chromosome 5 with $10^{-12} < P < 10^{-4}$ and considered these to comprise *cis* eQTLs. More surprisingly, we identified 908 autosomal SNPs in *trans* associated with *SEMA5A* expression ($10^{-8} < P < 10^{-4}$) (Supplementary Material, Table S1). The 920 SNPs were divided into 245 independent (>1 Mb apart) eQTL clusters (Supplementary Material, Table S1). The eQTL positions were mapped onto genes to identify likely eQTL genes. For eQTL clusters not directly overlapping any genes, we chose the single gene nearest to either of the outer SNPs. This resulted in a total of 321 eQTL-associated genes: 87 eQTLs overlapping a single gene, 18 eQTLs overlapping two or more genes, and 140 eQTLs not overlapping a gene for which the nearest gene was selected. Our top 10 *SEMA5A* eQTLs with smallest minimum *P* are presented in Table 1. A hierarchical clustering of expression of eQTL genes revealed 10 major gene clusters (Fig. 1, Supplementary Material, Fig. S1), and *SEMA5A* was located in one of the clusters, associated closely with 15 other genes (*MAST4*, *FOXP1*, *MBD5*, *C9orf30*, *DNMT3A*, *EVC*, *GPR45*, *GSX1*, *IPO8*, *NSF*, *PROKR1*, *TRAF2*, *TXNL1*, *ZBED5*, *ZNF100*).

SEMA5A eQTL^2 mapping

Because we had identified such an extensive *trans*-regulatory network associated with *SEMA5A* expression, we considered the possibility that a small number of upstream 'master regulators' controlled this network. In order to identify putative master regulators, we performed eQTL mapping for expression level of each of the eQTL-associated genes (above) using the same strategy as our *SEMA5A* eQTL mapping. After obtaining a list of eQTLs for the available expression profiles for 202 eQTL-associated genes, we looked for overlap, specifically regions associated with multiple *SEMA5A* eQTLs. We call these *SEMA5A* secondary eQTLs (eQTL^{2}s), as they were the eQTLs of the eQTLs of *SEMA5A*. Twelve eQTL^{2}s were identified as associated ($P < 10^{-4}$) with the expression of 10 or more primary *SEMA5A* eQTL genes (Table 2, Fig. 1). These 12 regions all contain strong gene regulatory candidates, such as known transcription regulators or genes containing DNA-binding domains (*RARB*, *AUTS2*, *THRB*, *RFPL4B*, *FOXI2*, *ZNF521*, *LMO3*, *SHPRH*), or major regulatory signaling pathway genes (*GNAI1*, *MAGI2*, *NRG3*, *MARCKS*, *CTGF*, *GHITM*, *DOCK1*, *SS18*, *RERGL*, *TCL1A*, *BDKRB2*, *ANGPT4*). Thus, they seem likely candidates for master regulators influencing transcription of multiple genes and ultimately impacting *SEMA5A* expression downstream.

Common variant association between *SEMA5A* eQTLs and autism

In addition to the identification of an extensive regulatory network for *SEMA5A*, we wanted to test our original hypothesis that examining regulators of *SEMA5A* expression would uncover novel association with ASD risk. We used the original AGRE-NIMH data set in which the GWAS association near *SEMA5A* was detected as well as two largely independent autism GWAS data sets (Table 3): the Autism Genome Project (AGP) and Simons Simplex Collection (SSC-V1). We tested the association of combined *SEMA5A* eQTL regions with autism as a single set. We extracted all SNPs genotyped in each data set within the 245 eQTL clusters (1 Mb flank on each side) and performed a set-based transmission disequilibrium test (TDT), as implemented in PLINK (28). The combined eQTL SNPs showed significant set-based association in AGRE-NIMH ($P = 0.020$) and AGP ($P = 0.002$), but not SSC-V1. A control data set (NHGRI-VU type 2 diabetes) did not show evidence for association. In order to be

Table 1. Top 10 SEMA5A eQTLs regions with smallest minimum *P*

Chromosome	Start	End	eQTL gene[a]	Left gene[b]	Right gene[c]	# SNPs[d]	*P*-value[e]
5	8 432 269	10 486 082	***SEMA5A***	*MTRR*	*SNORD123*	9	2×10^{-12}
18	44 894 006	46 979 045	***MIR4320,MYO5B***	*ACAA2*	*CCDC11*	24	2×10^{-8}
9	9 471 065	11 471 205	***PTPRD***	*JMJD2C*	*MPDZ*	2	3×10^{-8}
14	22 917 791	24 973 124	***CBLN3,KHNYN,NFATC4,NYNRIN***	*CMTM5*	*HCD1*	5	5×10^{-8}
4	167 781 602	169 841 939	–	***SPOCK3***	*ANXA10*	12	7×10^{-8}
18	6 084 612	8 084 612	***LAMA1***	*ARGHGAP28*	*LRRC30*	1	2×10^{-7}
20	39 042 867	41 132 741	–	*CHD6*	***PTPRT***	3	2×10^{-7}
12	3 363 712	5 374 681	–	***FGF23***	*FGF6*	4	3×10^{-7}
11	11 310 553	13 324 688	***MICALCL***	*MICAL2*	*PARVA*	5	4×10^{-7}
1	206 278 586	208 290 185	***KCNH1***	*HHAT*	*RCOR3*	8	5×10^{-7}

The top 10 eQTL regions with the smallest association *P*-value are shown here with their chromosome locations, overlapping or nearest Refseq genes, number of SNPs along with the smallest *P*-value in each eQTL regions. The coordinates are based on human genome assembly NCBI36/hg18.
[a]eQTL gene that is located in the eQTL cluster. A "–" indicates that no gene overlaps the eQTL cluster.
[b]eQTL gene that is located on the left side (p-terminal) of the eQTL cluster.
[c]eQTL gene that is located on the right side (q-terminal) of the eQTL cluster. Selected eQTL gene(s) bolded for each region.
[d]Number of SNPs that are located in the eQTL cluster.
[e]Minimum *P*-value of association test of *SEMA5A* expression in the eQTL cluster.

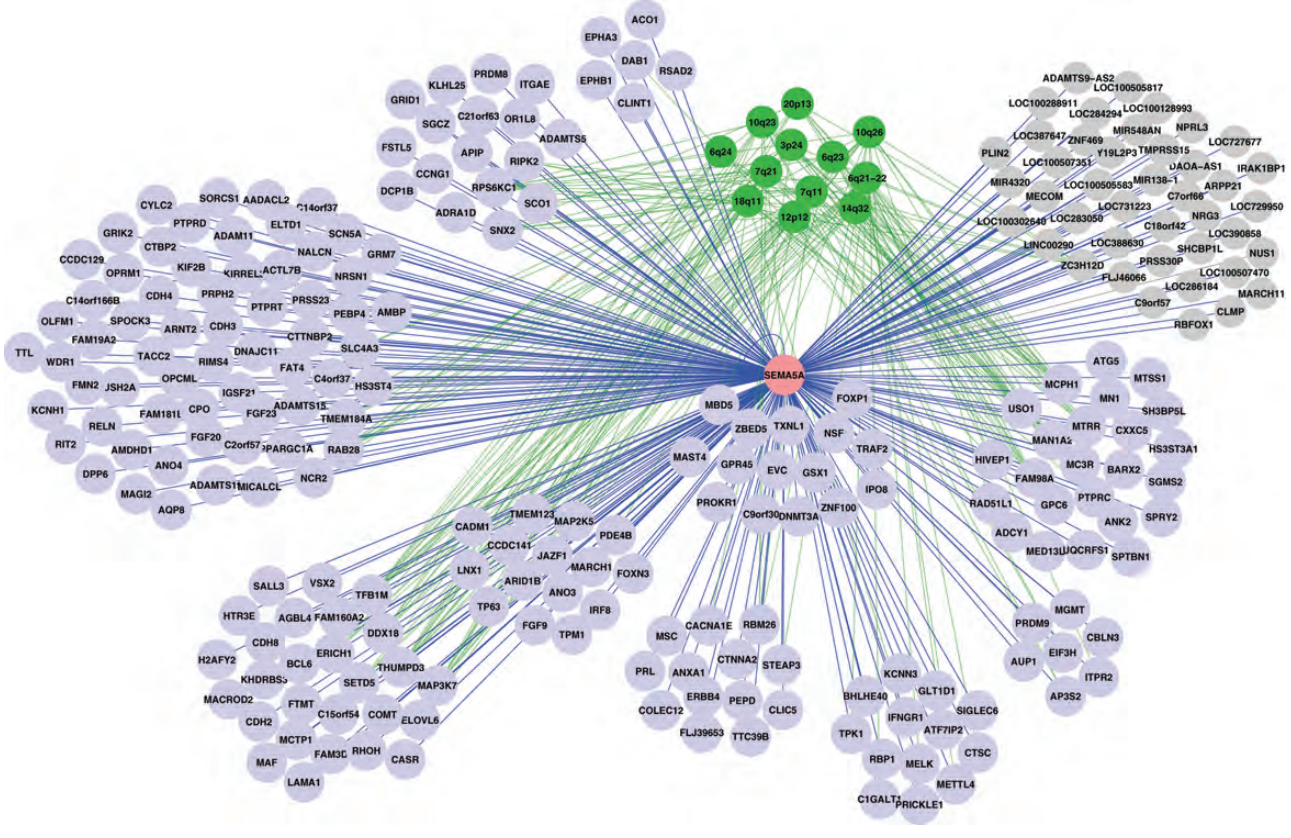

Figure 1. Gene clustering in the *SEMA5A* expression network. This figure presents a hierarchical clustering using 1- |*r*| (*r*, Pearson's correlation coefficient) as a measurement of distance (see Supplementary Material, Fig. S1). The number of clusters is set to 10. eQTLs are shown in purple (with LCL expression data available) and grey (with no LCL expression data available). eQTL[2] loci are shown in green with their chromosomal band position indicated, and their connections to the corresponding eQTL genes in green. *SEMA5A* is shown in pink, with its edge to eQTL nodes in blue. The accession number for the gene-expression data deposited in Gene Expression Omnibus is GSE7851 (80). Only the CEU sample ($N = 87$) data are used to map the network here.

certain the observed association was not an artifact of our eQTL regions being particularly gene dense, we performed a permutation test and found that the eQTL regions are not more gene dense than equivalent genomic regions (Supplementary Material, Table S1). In addition, to ascertain whether the *P*-values derived from gene-dropping in PLINK were truly adequate to test such a large set of potential variants, we performed a permutation-based test using 245

Table 2. eQTL^2 regions associated with ≥10 SEMA5A eQTL-associated genes

Chromosome	Start	End	eQTL^2 gene[a]	Left gene[b]	Right gene[c]	# eQTLs[d]
3	23 923 519	25 923 519	–	*THRB*	*RARB*	15
6	113 135 490	115 228 177	–	*RFPL4B*	*MARCKS*	21
6	145 745 143	147 745 143	*GRM1*	*SHPRH*	*STXBP5*	10
6	131 752 893	133 752 893	*MOXD1*	*CTGF*	*STX7*	10
7	67 061 612	69 123 105	–	*STAG3L4*	*AUTS2*	21
7	78 622 933	80 637 549	*GNAI1*	*MAGI2*	*CD36*	16
10	83 617 431	85 632 820	*NRG3*	*SH2D4B*	*GHITM*	14
10	127 974 317	129 978 367	*DOCK1*	*FAM196A*	*FOXI2*	18
12	16 596 504	18 601 788	–	*LMO 3*	*RERGL*	12
14	94 338 625	96 338 625	–	*TCL1A*	*BDKRB2*	10
18	20 354 815	22 448 785	–	*ZNF521*	*SS18*	18
20	1	1 788 405	–	*FAM110A*	*ANGPT4*	10

The 12 eQTL^2 regions are shown here with chromosome location, overlapping and/or nearest Refseq genes, number of associated eQTLs for each eQTL^2 region. The coordinates were based on human genome assembly NCBI36/hg18.
[a]eQTL^2 gene that is located in the eQTL^2 cluster. A "–" indicates that no gene overlaps the eQTL^2 cluster.
[b]eQTL^2 gene that is located on the left side (p-terminal) of the eQTL^2 cluster.
[c]eQTL^2 gene that is located on the right side (q-terminal) of the eQTL^2 cluster.
[d]Number of eQTLs that are associated with the eQTL^2 cluster.

Table 3. SNP and CNV data sets

Data set	Source[a]	Platform	# Subjects	# Families	# Markers
SNP data sets					
Primary	AGRE-NIMH	Affymetrix 5.0	1756	1033 multiplex	345 429
Replication	AGP-1M	Illumina 1M	1330	1369 simplex and multiplex	842 215
	SSC-V1	Illumina 1M	691	698 simplex	839 246
Meta-analysis	Imputed	Mixed microarray	4222	3444 simplex and multiplex	634 151
Control	NHGRI-VU	Illumina 1M	769	Case–control	1 199 187
Data set	Source[b]	Platform	# Subjects	# Families	
CNV data sets					
Primary	AGP-10K	Affymetrix 10K	196	173 multiplex	
	CIHR	Affymetrix 550K	427	427 simplex and multiplex	
	AGRE-ACC (exonic)	Illumina 550	2542	1802 simplex and multiplex	
	AGRE	Illumina 550	1683	943 simplex and multiplex	
	AGP-1M	Illumina 1M	996	876 simplex and multiplex	
	SSC	Illumina 1M	1124	1124 simplex	
	ACRD	Mixed microarray	–	–	
Control	DGV	Mixed microarray	–	–	
	WTSI	Agilent 105K	450	–	

Five SNP data sets were used for common polymorphism association analysis. The genotyping platform, number of subjects, number of families and number of high quality markers are shown. Seven autism and two control CNV data sets were used for rare variant association analysis. The genotyping platform, number of subjects and number of families are shown.
[a]AGRE-NIMH: combined data sets from AGRE (Autism Genetic Resource Exchange, www.agre.org) and NIMH (National Institute for Mental Health, collections of DNA from multiplex and simplex families with ASD by the NIMH Autism Genetics Initiative) (17). AGP-1M: Autism Genome Project (http://www.autismgenome.org/) (20). SSC-V1: Simons Simplex Collection V1 Public Cohort (https://sfari.org/simons-simplex-collection) (83). Imputed: imputed data sets based on AGRE-NIMH, AGP, SSC-V1 and AGRE-CHOP. NHGRI-VU: Vanderbilt University NUgene Project type 2 diabetes SNP data set (dbGap: phs000237.v1.p1)) (89).
[b]AGP-10K: Autism Genome Project CNV data set (15); CIHR: Marshall *et al.* CNV data set (13); AGRE-ACC: Bucan *et al.* CNV data set (14); AGRE: Autism Genetic Resource Exchange CNV data set identified by PennCNV (82); AGP-1M: Pinto *et al.* CNV data set (23); SSC: Sanders *et al.* CNV data set (12); ACRD: CNV data set from the Autism Chromosome Rearrangement Database (ACRD) (84); DGV: CNV control data set from the DGV (86); WTSI: Conrad *et al*. CNV control data set (29).

equivalent random genomic regions, which showed that the PLINK *P*-values were inflated, despite relatively low overall genomic inflation in our data sets ($\lambda_{\text{AGRE-NIMH}} = 1.021$; $\lambda_{\text{AGP}} = 1.018$; $\lambda_{\text{SSC-V1}} = 1.006$). Our permutation-based estimate suggested no association with the full set of *SEMA5A* eQTLs in any of the three data sets ($P_{\text{AGRE-NIMH}} = 0.44$; $P_{\text{AGP}} = 0.44$; $P_{\text{SSC-V1}} = 0.93$, Table 4).

Common variant association between *SEMA5A* eQTL^{2}s and autism

Similarly, we performed a set-based analysis of association using our permutation-derived estimates of significance between *SEMA5A* eQTL^2 loci and autism to determine whether our putative master regulators are independently

Table 4. Set-based TDT test in autism SNP data sets

Data set	Source[a]	*P*-value[b] eQTLs	eQTL^{2}s
Primary			
With rs10513025	AGRE-NIMH	3.8×10^{-1}	$\mathbf{2.0 \times 10^{-2}}$
Exclude rs10513025	AGRE-NIMH	4.4×10^{-1}	
Replication	AGP-1M	4.4×10^{-1}	1.3×10^{-1}
	SSC-V1	9.3×10^{-1}	9.3×10^{-1}
Meta-analysis	Imputed	2.0×10^{-1}	4.0×10^{-1}
		6.0×10^{-2}	
Control	NHGRI-VU	7.2×10^{-1}	2.6×10^{-1}

The result from a set-based TDT test of our *SEMA5A* eQTL and eQTL^2 regions is shown here for our primary (with and without the original GWAS associated SNP), replication and control data sets. The combined eQTL and eQTL^2 regions were used to perform a set-based TDT test in the meta-analysis data set.
[a]AGRE-NIMH: combined data sets from AGRE (Autism Genetic Resource Exchange, www.agre.org) and NIMH (National Institute for Mental Health, collections of DNA from multiplex and simplex families with ASD by the NIMH Autism Genetics Initiative) (17). AGP-1M: Autism Genome Project (http://www.autismgenome.org/) (20). SSC-V1: Simons Simplex Collection V1 Public Cohort (https://sfari.org/simons-simplex-collection) (83). Imputed: imputed data sets based on AGRE-NIMH, AGP, SSC-V1 and AGRE-CHOP. Both the *P*-values for separate test (above) and combined tests (below) are shown in the table. NHGRI-VU: Vanderbilt University NUgene Project type 2 diabetes SNP data set (dbGap: phs000237.v1.p1)) (89).
[b]Permutation-based *P*-value from set-based association tests of *SEMA5A* eQTLs and eQTL^{2}s. *P*-values < 0.05 are bolded.

Table 5. CNV analysis in autism and control data sets

Data set	Source[a]	#CNVs[b]	Overlap%[c]	OR[d]	*P*-value[e]
Primary	AGP-10K	175	30%	1.1	3.73×10^{-1}
	CIHR	262	30%	1.6	$\mathbf{5.00 \times 10^{-4}}$
	AGRE-ACC (exonic)	315	20%	1.1	4.41×10^{-1}
	AGRE	6303	20%	1.1	2.02×10^{-1}
	AGP-1M	960	21%	1.3	$\mathbf{1.31 \times 10^{-2}}$
	SSC	2,756	20%	1.2	$\mathbf{4.38 \times 10^{-2}}$
	ACRD	355	29%	1.4	$\mathbf{2.00 \times 10^{-4}}$
Control	DGV	36 095	12%	1	6.88×10^{-1}
	WTSI	2248	12%	0.6	9.95×10^{-1}

The result of permutation analysis of rare variant data sets is presented here.
[a]AGP-10K: Autism Genome Project CNV data set (15); CIHR: Marshall *et al.* CNV data set (13); AGRE-ACC: Bucan *et al.* CNV data set (14); AGRE: Autism Genetic Resource Exchange CNV data set by PennCNV; AGP-1M: Pinto *et al.* CNV data set (23); SSC: Sanders *et al.* CNV data set (12); ACRD: CNV data set from the Autism Chromosome Rearrangement Database (ACRD) (84); DGV: CNV control data set from the DGV (86); WTSI: Conrad *et al* CNV control data set (29).
[b]The total number of autism-specific CNVs.
[c]The percentage of CNVs that overlap *SEMA5A* eQTLs and eQTL^{2}s.
[d]Odds ratio calculated in comparison with the median permutation overlapping CNV number.
[e]*P*-value from permutation tests of *SEMA5A* eQTLs and eQTL^{2}s. *P*-values $<$ 0.05 are bolded.

associated with autism risk. Association was found at $P = 0.03$ for the set of SNPs in the top eQTL^2 regions in the AGRE-NIMH data set using the permutation-based set analysis (Table 4). In the AGP and SSC-V1 replication data sets, we found no association for the set of eQTL^2 SNPs (Table 4). The eQTL^2 SNP set was not significantly associated with an unrelated phenotype in the control NHGRI-VU type 2 diabetes data set.

Meta-analysis for SNP association with autism

We performed genotype imputation in order to combine these three autism data sets and carry out meta-analysis. This combined data set has the following desirable features: (i) the existing autism data sets use different genotyping platforms, and imputation can test the same set of markers in the combined data sets; (ii) in a combined data set we can eliminate any sample overlap across data sets and include additional AGRE families genotyped on a different platform (22), and (iii) we achieve greater statistical power through the increase in sample size. We performed a meta-analysis for *SEMA5A* eQTLs and eQTL^{2}s, and we did not observe association (eQTL $P = 0.20$; eQTL^2 $P = 0.40$). In order to be consistent with the combined test for CNVs (described below), we also performed a test for the combined *SEMA5A* eQTL and eQTL^2 regulatory regions and observe near-significant association with ASDs based on this imputed data set, with $P = 0.06$ (Table 4).

Rare variant association between *SEMA5A* regulatory network and autism

Among our initial eQTL and eQTL^2 regions, we recognized several loci that overlapped autism linkage regions, CNVs or point mutations previously observed in autism. Thus, we wanted to formally test the possibility that rare variation in the *SEMA5A* regulatory network might contribute to autism risk. In order to do this, we developed a permutation test utilizing published CNV data in ASDs. From the published results tables of the large genome-wide CNV studies in autism (Table 3), we derived lists of autism-specific CNVs [not observed in controls or less than 80% overlap with records in Database of Genomic Variants (DGV)] and tested whether they significantly overlap with the joint list of *SEMA5A* eQTLs and eQTL^{2}s compared with equivalent genomic regions and compared with other eQTL regions defined in the same way sampled from all RefSeq genes. Due to the limited number of autism-specific CNVs, we decided to test *SEMA5A* eQTL and eQTL^2 regions as a combined regulatory network in order to maximize power. One of the largest subsets from DGV (Wellcome Trust Sanger Institute, WTSI) (29) was tested in addition to the complete DGV record as control data sets. Four of the seven autism data sets showed significant overlap of autism-specific CNVs with *SEMA5A* eQTL and eQTL^2 regions at $P < 0.05$ (Table 5). This provides strong evidence that rare autism CNVs in our eQTL and eQTL^2 regions are associated with autism. The overlap between CNVs in ACRD and eQTL and eQTL^2 regions is displayed in Figure 2 (and Supplementary Material, Table S1). Although the other three autism data sets did not have significant *P*-values, the odds ratios are elevated compared with the permuted data sets. In testing >50 top GWAS candidate genes with eQTL networks ranging from 18 to 330 eQTLs from unrelated phenotypes, we did not observe any eQTL networks associated in 4/7 autism

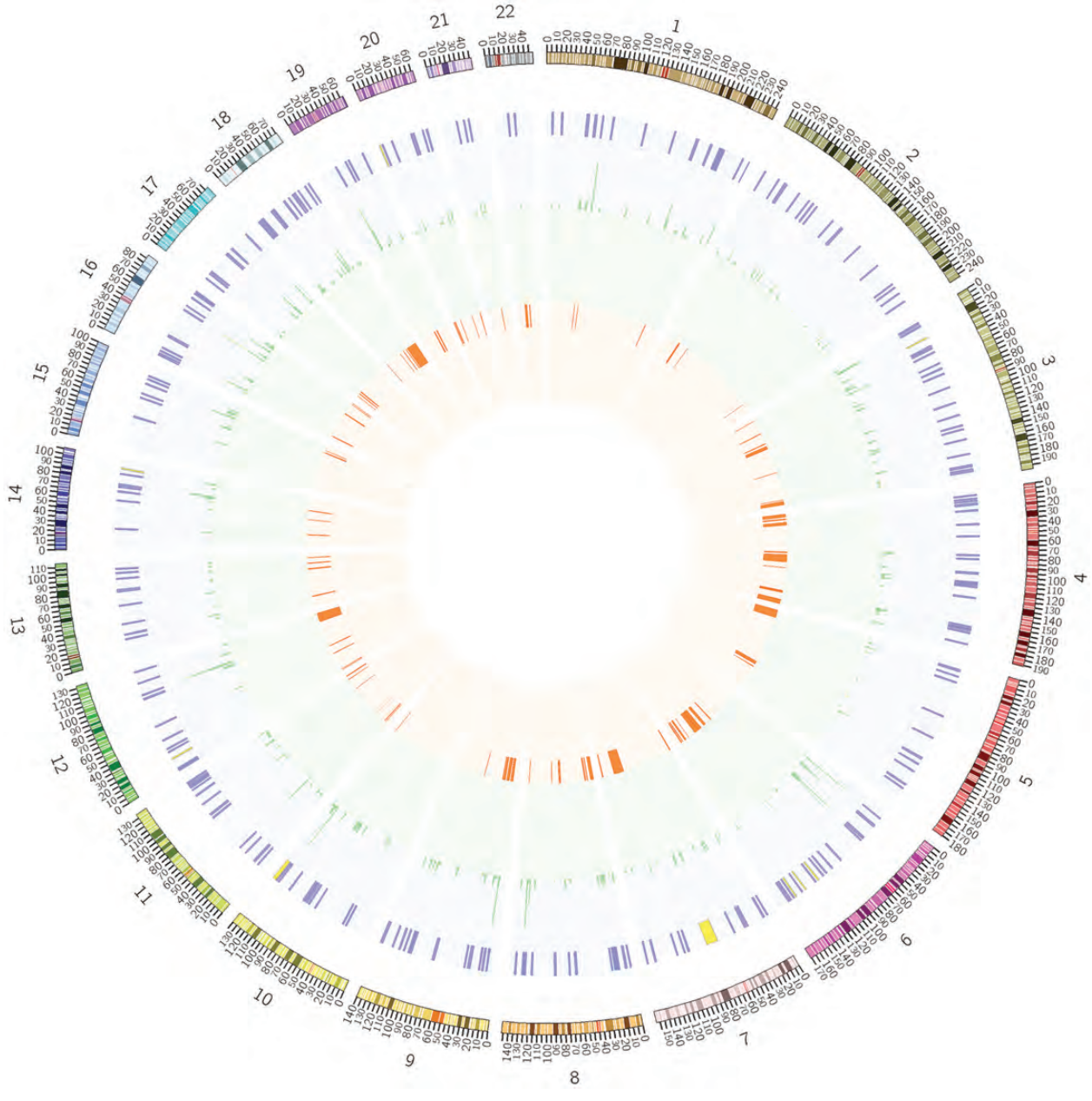

Figure 2. ASD-associated CNVs and SNPs overlap *SEMA5A* eQTL and eQTL^2 regions. The large circle shows CNVs, SNPs, eQTL and eQTL^2 regions for the whole genome. To obtain a more detailed view, small circles (from chromosome 1 to chromosome 22) are provided in the supplementary information (Supplementary Material, Fig. S1) to show CNVs, SNPs, eQTL and eQTL^2 regions for each chromosome. From the outside to the inside of each circle, each sub-circle correspondingly shows: (i) chromosome coordinates with bands highlighted. (ii) eQTL and eQTL^2 regions are represented by each stroke. eQTL regions are indicated by purple, while eQTL^2 regions are indicated by yellow. Overlapping eQTL and eQTL^2 regions are combined. (iii) Significant SNPs identified in the meta-analysis are shown in green. For each SNP, the significance by TDT is highlighted by the length of the stroke on each glyph. From longest to shortest: $P < 0.0001$; $P < 0.001$; $P < 0.01$; $P < 0.05$. The total number of SNPs shown here is 3584. The *P*-value for the combined eQTL and eQTL^2 set-based TDT test in the meta-analysis is 0.06. (iv) CNVs in the ACRD (Autism Chromosome Rearrangement Database) (84) that overlap with eQTLs and eQTL^{2}s are shown in red. The total number of CNVs in ACRD is 371, and 104 of them (27%) overlap with one of the eQTL and eQTL^2 regions. The median overlapping number of 10 000 permutations is 80. The *P*-value is 0.0004 with OR = 1.5.

CNV data sets (data not shown), suggesting that our results are specific to the *SEMA5A* eQTL network.

DISCUSSION

Previous genome-wide studies in autism have identified candidate genomic regions associated with autism risk (10–23). However, it is not easy to identify the autism susceptibility genes in these regions because the associated variants are rarely unambiguously tied to a single gene via a straightforward mechanism. Thus, we designed a novel approach based on genetic and expression data.

We performed *in silico* mapping of eQTLs and eQTL^{2}s for a GWAS candidate gene, *SEMA5A* (17), in order to define an

empirical regulatory network (Fig. 1). We identified association between rare autism-specific CNVs and the combined eQTL and eQTL^2 regions. Although several individual data sets did not provide evidence for association, we cannot rule out low power or heterogeneity due to sampling strategies, and we find the weight of the CNV evidence overall to be strong and consistent.

We observed little evidence for association between common polymorphisms in this regulatory network and autism. A single data set showed association with eQTL^2 regions, and the meta-analysis showed near-significant association with autism. These results in contrast to the strong evidence from CNVs could simply reflect the genetic architecture of autism, which to date has shown stronger and more consistent evidence for CNV contribution to ASD risk. Although autism has been shown to have a polygenic SNP contribution, it is relatively modest (20,30–33). These data could also reflect a stronger magnitude of effect for CNVs on gene expression compared with SNPs (34–38).

Our approach, in fact, relies on two significant assumptions which we know are likely to be imperfect. First, the expression and genetic data set used for eQTL mapping were not derived from the tissue of interest nor from affected individuals. Although recent work suggests that eQTL enrichment in autism GWAS is stronger in the brain than in LCLs (39), we are focused on a specific candidate gene network, and therefore require larger data sets than are currently available for the brain or from individuals with ASDs. Based on previous literature suggesting that although absolute expression level may vary across tissue, genetic regulation of expression is similar in primate blood and brain and relevant autism genes can be identified in LCLs (40–46), we make the assumptions that genetic polymorphisms exert their influence on expression ubiquitously where genes are expressed and that the same relationships observed in normal expression variation in controls will be affected in the disease state of autism. Thus, we are limited to identification of tissue-independent and disease-status independent relationships between genetic variation and expression level. Second, for our eQTL^2 mapping, we do not truly know which gene each primary *SEMA5A* eQTL acts through (or that it acts through a gene at all) and as a conservative unbiased approach have used physical proximity to select eQTL-associated genes, although other approaches such as expression pattern or correlation might introduce less noise into the network to be tested for association. Because we have identified association between genetic variation in the eQTL and eQTL^2 regions and ASD susceptibility under the current design, our assumptions must hold some validity, although we surmise that we have missed some important aspects of the network.

In the new wave of exome and genome sequencing to identify rare sequence variants associated with common, complex disease, it remains an open question whether common polymorphism risk loci likely to have modest individual effects (e.g. subtle expression differences) and rare variants that might have stronger individual effect (e.g. hemizygosity for deletion of multiple genes or protein coding changes) will act in the same pathways. Our novel pathway was entirely defined based on (i) SNP association signal in an autism data set and (ii) SNP data in control LCL expression experiments. Together, these data suggest that rare CNVs show association with autism in a gene expression regulatory pathway defined by SNPs. This pathway might be used to interpret rare variant data, for example, rare CNV or sequence variants overlapping this empirically derived network might have greater *a priori* likelihood to be pathogenic. Examples of previously identified rare variants that fall into this network are translocations in *AUTS2* (eQTL^2) in autism (47–51) and mental retardation (52), linkage regions on chromosome 7q21 (53–55) and 20p (eQTL^2) (17), point mutations in *CADM1* (eQTL) (56), exonic deletion, point mutation and chromosomal abnormalities in *FOXP1* in autism (51,57–59), duplications of *MCPH1* (eQTL) (60,61), duplication or deletion of *MBD5* (eQTL) (51,62–64), deletions of *MAST4* (eQTL) (16), *CDH8* (eQTL) (65), *A2BP1* (eQTL) in autism (10,66,67), rare variants in *KIRREL3* (eQTL) (51), rare complete knockouts in *DBF4B*, *de novo* deletion and loss of function of *USH2A* in autism (68). Likewise, the overall network evidence might be used to prioritize sub-genome-wide significant SNP signals within these regions for follow-up or make functional interpretation more clear. For example, SNP association was seen near 6q22 (eQTL^2) and several regions overlapping eQTLs (chromosomes 2, 4, 5, 11, 13, 20 and 21 and a genome-wide-significant region near *MACROD2*) (17,20). In addition, previous data have shown reduced expression of *RELN* (eQTL) in autism brain tissue (see Supplementary Material, Table S1) (69).

Semaphorins comprise a large family of molecular cues critical to neural development and also having a variety of functions outside the nervous system. *SEMA5A* is a transmembrane semaphorin, whose receptor is plexin-B3 (*PLXNB3*) (70), encoded on the X chromosome. Little is known about the specific function of *SEMA5A*, although it appears to play a neurodevelopmental role in axonal guidance and neurite outgrowth (71,72). In mouse endothelial cells, *SEMA5A* plays general roles in downregulating apoptosis (through Akt), increasing migration (through Met tyrosine kinases) and controlling the extracellular matrix (through matrix metalloproteinase 9) (73). An initial *SEMA5A* knock-out mouse on the 129/Sv/NMRI background showed embryonic lethality (74), and a more recent knock-out mouse on the mixed B6/129P2 background found several genotype by sex effects and behavioral differences, but no differences in cognition or social behavior compared with C57BL/6J control mice (75). Recently, a study in primary human neuronal progenitors from control subjects detected increased expression in *SEMA5A* at the time point of neuronal differentiation and this was coordinately regulated with other ASD candidate genes (76). Our study provides additional testable information about a putative regulatory network of *SEMA5A*, along with support from genetic variation data that it is important in autism risk. In fact, these data suggest the novel hypothesis that *SEMA5A* is the common downstream effector for all the genes in this network and that autism CNVs in this network act through modulation of *SEMA5A* expression. A definitive test of this hypothesis might be in a cellular model of ASDs, such as induced pluripotent stem cell (iPSC) derived neurons. If patient-specific iPSCs were generated from individuals with various CNVs in the *SEMA5A* network defined here and deficits at the neuronal level could be corrected by upregulation of

SEMA5A, our hypothesis would be substantiated. If this hypothesis is indeed proven by further testing, it suggests a new understanding of the biology of autism, as well as new directions for biomarker and treatment targets in ASDs. Construction of networks underlying core endophenotypes can make it possible to identify the fundamental biological processes that are responsive to genetic or environmental perturbation and can lead directly to alteration of disease risk, as has been successfully done for metabolic traits (77,78).

In conclusion, our novel approach successfully took a GWAS-identified candidate gene, *SEMA5A*, and garnered evidence for rare genetic variant risk in its regulatory network. These data can be used in interpretation of SNP association data, as well as interpretation of rare variants. In sum, our study provides a novel framework for identifying the genetic loci that are likely to contribute to autism risk, and the general approach can be extended to other complex heritable disorders.

MATERIALS AND METHODS

Data sets

eQTL mapping

In the SCAN database (http://www.scandb.org/) (27), LCL expression and genetic experimental data from samples of European descent (CEPH from Utah; CEU) were used to map autosomal eQTLs. The published expression data were generated using the Affymetrix Human Exon 1.0 ST Array (79,80). The autosomal SNP genotype data were generated on multiple experimental platforms by the HapMap Project (www.hapmap.org/) (81). The SNP positions were based on dbSNP 129 from NCBI.

SNP data sets

Autism sample banks and consortia have been used to perform past or ongoing GWAS studies (Table 3). These include Autism Genetic Resource Exchange (AGRE, www.agre.org) (82), National Institute for Mental Health (NIMH, collections of DNA from multiplex and simplex families with ASD by the NIMH Autism Genetics Initiative), Simons Simplex Collection (SSC-V1, https://sfari.org/simons-simplex-collection, V1 Public Cohort) (83) and Autism Genome Project (AGP, http://www.autismgenome.org/) (15). These data sets are not completely independent due to partial sample overlap. The AGRE-NIMH GWAS data set was generated by the Broad Institute of MIT and Harvard and Johns Hopkins Center for Complex Disease Genomics on Affymetrix platforms and includes 1756 affected individuals in 1033 multiplex families with 345 429 high quality markers (17). The SSC-V1 data set was genotyped on the Illumina 1M array at Yale University. Before quality control, there were 1 231 154 markers. We removed all markers with (i) $>4\%$ missingness, (ii) minor allele frequency (MAF) $<1\%$, (iii) Hardy–Weinberg equilibrium (HWE) P-value less than 10^{-6}, and (iv) markers with MAF $<5\%$ and $>0.5\%$ missingness. The final SSC-V1 data set includes data on 698 families with a single child affected with ASDs and 839 246 high-quality markers genotyped. The AGP data set was genotyped at Translational Genomics Research Institute (TGEN) and available via dbGAP (phs000267.v1.p1). The quality control steps are: (i) remove SNPs not on chromosomes 1 to 22, and (ii) the following SNPs are removed due to bad genotyping quality as assessed by visual examination of intensity plots: rs12117357, rs2196826, rs2071004, rs1709905 and rs10419948. The final AGP data set includes 1369 families genotyped for 842 215 markers on Illumina 1M, approximately evenly divided between multiplex and simplex families. From the AGP, SSC-V1 and AGRE-NIMH SNP data sets, a combined SNP data set was generated by imputation (see Materials and Methods for meta-analysis protocol used in imputation). Additional AGRE families genotyped on an Illumina 550K platform from Children's Hospital of Philadelphia (AGRE-CHOP) were included in the imputation (22). This imputed data set eliminated any overlapping samples and was used for SNP meta-analysis.

Our control SNP data set was the Vanderbilt University NUgene Project: type 2 diabetes SNP data set (NHGRI-VU), available via dbGap (phs000237.v1.p1). This case–control study had 769 cases and 615 controls with 1 199 187 high-quality markers genotyped on Illumina 1M.

CNV data sets

Seven autism and two control data sets were analyzed here. Six published autism data sets included: 175 CNVs from the AGP-10K data set on Affymetrix 10K (15), 277 CNVs from the CIHR data set on Affymetrix 550K (13), 315 exon-overlapping CNVs from the AGRE-ACC data set on Illumina HumanHap550 (14), 1020 CNVs from the AGP-1M data set on Illumina 1M (23), 3077 CNVs from the SSC data set on Illumina 1M (12) and 371 CNVs from the Autism Chromosome Rearrangement Database (ACRD) on mixed platforms (84).

Because the full AGRE-ACC CNV calls were not available in the Supplementary Materials, we performed CNV calling using the raw data from the AGRE portion of this data set and obtained 4719 CNVs as the seventh autism data set. We applied a hidden Markov model (HMM) implemented in PennCNV (85) to detect CNVs in the AGRE data set. We kept the corresponding parameters the same as Bucan *et al.* (14) to make results comparable. A total of 3554 samples meeting the following criteria were included for individual CNV calling: (i) standard deviation for autosomal log R ratio values (LRR_SD) less than or equal to 0.28; (ii) median B Allele Frequency (BAF_median) larger than or equal to 0.45 and less than or equal to 0.55; (iii) fraction of markers with BAF values between 0.2 and 0.25 or 0.75 and 0.8 (BAF_drift) less than 0.002. We then utilized the family information to perform trio CNV calling for more accurate boundaries. Two additional filtering steps were used to remove CNVs in the repetitive regions: (i) CNVs with more than 50% overlap with four immunoglobulin regions (IGLC1 22q11.22, IGHG1 14q32.33 and IGKC 2p11.2, and the T cell receptor constant chain locus 14q11.2); (ii) CNVs within centromeric and telomeric regions (500 kb flanking windows were added).

Finally, a further filtering process was performed to all the autism data sets to retain autism-specific CNVs with: (i) less than 80% overlap with any variant in the DGV, (ii) greater than 1 kb length, and (iii) on autosomes. The DGV

(variation.hg18.v10.nov.2010) contained 27 113 autosomal CNVs in the healthy population (86) and was used as a control data set. The WTSI data set (29), one of the largest data set components of the DGV, with 2331 autosomal CNVs was tested separately as the second control in order to have a control data set with uniform methods and CNV calling.

Analysis methods

Genotype imputation and meta-analysis protocol

First, the phased Hapmap Phase III data set was obtained from http://hapmap.ncbi.nlm.nih.gov/downloads/phasing/2009-02_phaseIII/HapMap3_r2/ as reference genotypes. Because of population diversity in the autism data sets, we attempted to separate individuals into different populations to improve imputation performance. To accomplish this, each autism data set was merged with the Hapmap data set for identity-by-state clustering and individuals were assigned to one of the five imputation panels based on visual inspection of clustering. The five imputation panels used were built from combinations of the Phased Hapmap III samples. The panels were as follows, Panel 1: CEU + TSI; Panel 2: ASW, LWK, MKK, YRI; Panel 3: JPT, CHB, CHD; Panel 4: MEX; Panel 5: GIH.

Imputation was carried out using the BEAGLE software (http://faculty.washington.edu/browning/beagle/beagle.html#download) (87).

Before the imputation, each autism data set was subjected to stringent quality control and then prepared into subsets of 10 000 continuous SNPs and 300 individuals for feasibility of computation. Each consecutive SNP subset overlapped at least 1000 SNPs. The same chromosomal position cutoffs were applied to the corresponding imputation panel. Genotypes imputed in two SNP subsets were discarded from the subset where its location was closer to the edge to minimize edge effects. All SNPs were aligned to a common strand to ensure imputation accuracy using the BEAGLE strand switching utility (http://www.stat.auckland.ac.nz/~browning/beagle/strand_switching/strand_switching.html). Imputed genotypes with a posterior probability <0.9 were set as uncalled. After imputation, SNPs that met the following criteria were removed: (i) Mendelian error rate $>1\%$ and (ii) HWE P-value $< 1 \times 10^{-10}$. In addition, SNPs in AGRE-NIMH (Affymetrix) and AGRE-CHOP data sets with a call rate $<98\%$ and SNPs in SSC-V1 data set with a call rate $<95\%$ were removed. Finally, four data sets were merged and only SNPs with a combined call rate $>95\%$ were retained.

After imputation, commonly genotyped SNPs were extracted from the imputed AGRE-NIMH, AGRE-CHOP, SSC-V1and AGP data sets. A DerSimonian–Laird random-effects meta-analysis was used to perform a meta-analysis based on the set-based TDT results for each SNPs in the eQTL and eQTL^2 regions (88). We implemented our meta-analysis routines in the Python programming language (www.python.org) and it is available upon request.

eQTL mapping and definition

In order to identify eQTLs for a given gene expression trait (*SEMA5A* or primary eQTL-associated gene expression level), we performed association mapping in SCAN with available CEU LCL expression and genotype data using the default threshold of $P < 10^{-4}$. The eQTL SNPs were organized into clusters according to their genomic position. In order to define independent loci, we walked along the chromosomes and considered any consecutive SNP with a distance of less than 1 Mb to be within the same locus or cluster and a consecutive SNP with a distance of greater than 1 Mb from the nearest SNP to be a new eQTL cluster. The eQTL gene(s) were assigned to each eQTL cluster as the RefSeq gene(s) overlapping the cluster or the single nearest Refseq gene to the eQTL region if none overlapped. We obtained the expression data for eQTL genes used in SCAN through Gene Expression Omnibus (GSE7851) (80), which included 87 CEU samples. We then performed a hierarchical clustering (Methods described in Supplementary Material, Fig. S1) of this data set with 1-$|r|$ (r, Pearson correlation) as a measurement of distance. In addition, we selected more than 50 top GWAS hits from a variety of human traits, including hypertension, hair color, longevity, HIV-1 control, bone mineral density, prostate cancer, breast cancer, leukemia and heart disease (http://hugenavigator.net/HuGENavigator/gWAHit.do). We mapped the eQTL networks for each of these GWAS candidate genes in a similar way as we did for the *SEMA5A* gene. The eQTL networks (with 18–350 eQTLs) were used in our rare variant analysis to determine whether the association with autism CNVs was specific to the *SEMA5A* network.

Set-based SNP analysis and replication

In the first step, we identified all high-quality SNPs within the eQTL and eQTL^2 regions (including a 1 Mb flank in each direction) in each SNP data set. Each data set has different SNP markers directly assayed, therefore the set of SNPs used for association analysis was different across data sets, but no bias in allele frequency or other characteristics was carried over from the SCAN eQTL data set. We then performed a set-based TDT for these markers as implemented in PLINK (28). This consisted of a TDT for all SNPs in each data set, and determination of the LD structure. In each set, the most significant SNP was selected ($P < 0.05$) and then the second most significant SNP was included after removing SNPs in LD ($r^2 \geq 0.5$) with already selected markers. This was repeated until all SNPs with $P < 0.05$ were identified. The test statistic was calculated as the mean of single SNP test statistics. In the PLINK implementation of this test, the alleles of each parent were randomly dropped to offspring with a 50:50 probability to obtain a permuted data set and calculate a P-value. However, after discovery that these P-values were inflated in our data, we derived a permutation-based P comparing the test statistic from the *SEMA5A* eQTL set with test statistics from 100 sets of random genomic regions with the same size as *SEMA5A* eQTLs.

Permutation CNV analysis

We first did a gene density test of this *SEMA5A* network to ensure that our CNV burden analysis would not be biased by eQTL and eQTL^2 regions tending to be gene dense. This was achieved by comparing the number of genes in eQTL and eQTL^2 regions with random genomic regions of similar length (10 000 permutations). Due to the relative small size of eQTL^2 regions and relatively few autism-specific CNVs,

we combined the eQTL and eQTL^2 regions to obtain adequate power for our CNV analysis. The contribution of the *SEMA5A* regulatory network to autism was assessed by estimating the rare CNV burden in this pathway. In each of the nine CNV data sets, we first calculated the total number of CNVs that overlapped with our *SEMA5A* eQTL and eQTL^2 regions. In the second step, random regions with same length to these eQTL and eQTL^2 regions were sampled from the autosomal genome. Permuting segments with the same length corrects for the difference of CNV size among data sets, because large CNVs have a higher probability of overlap with the random regions. Again, a total number of CNVs that overlapped with these random regions was computed. This was repeated for 10 000 permutations and the corresponding *P*-value was calculated. Based on the median overlapping number in the permutations, we calculated the odds ratio of overlap for each data set. In order to exclude bias in regions defined by GWAS eQTL data, we performed an additional control. We mapped eQTLs for all RefSeq genes in SCAN and performed a similar permutation test using randomly selected eQTL regions (instead of any genomic region) defined in the same way and with similar size to *SEMA5A* eQTL regions. The resulting *P*-values for each autism data set were similar to those derived using genomic regions, so these data are not shown.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at *HMG* online.

ACKNOWLEDGEMENTS

We thank all of the families who have participated in and contributed to the public resources that we have used in these studies. We thank the authors of the published CNV data sets/databases used in our analysis and our colleagues at UCSF for their comments on the manuscript. We thank Dr Nancy Cox at University of Chicago and Drs Steve Hamilton and Jane Gitschier at UCSF for helpful advice.

The collection of data and biomaterials that participated in the National Institute of Mental Health (NIMH) Autism Genetics Initiative has been supported by National Institute of Health grants MH52708, MH39437, MH00219 and MH00980; National Health Medical Research Council grant 0034328; and by grants from the Scottish Rite, the Spunk Fund, Inc., the Rebecca and Solomon Baker Fund, the APEX Foundation, the National Alliance for Research in Schizophrenia and Affective Disorders (NARSAD), and the endowment fund of the Nancy Pritzker Laboratory (Stanford); and by gifts from the Autism Society of America, the Janet M. Grace Pervasive Developmental Disorders Fund, and families and friends of individuals with autism. The NIMH collection Principal Investigators and Co-Investigators were: Neil Risch, Richard M. Myers, Donna Spiker, Linda J. Lotspeich, Joachim F. Hallmayer, Helena C. Kraemer, Roland D. Ciaranello, Luigi Luca Cavalli-Sforza (Stanford University, Stanford); William M. McMahon and P. Brent Petersen (University of Utah, Salt Lake City). The Stanford team is indebted to the parent groups and clinician colleagues who referred families and extends their gratitude to the families with individuals with autism who were partners in this research. The collection data and biomaterials also come from the Autism Genetic Resource Exchange (AGRE) collection. This program has been supported by a National Institute of Health grant MH64547 and the Cure Autism Now Foundation. The AGRE collection Principal Investigator is Daniel H. Geschwind (UCLA). The Co-Principal Investigators include Stanley F. Nelson and Rita M. Cantor (UCLA), Christa Lese Martin (Univ. Chicago), T. Conrad Gilliam (Columbia). Co-Investigators include Maricela Alarcon (UCLA), Kenneth Lange (UCLA), Sarah J. Spence (UCLA), David H. Ledbetter (Emory) and Hank Juo (Columbia).

We gratefully acknowledge the resources provided by the Autism Genetic Resource Exchange (AGRE) Consortium and the participating AGRE families. The Autism Genetic Resource Exchange is a program of Autism Speaks and is supported, in part, by grant 1U24MH081810 from the National Institute of Mental Health to Clara M. Lajonchere (PI).

We are grateful to all of the families at the participating SFARI Simplex Collection (SSC) sites, as well as the principal investigators (A. Beaudet, R. Bernier, J. Constantino, E. Cook, E. Fombonne, D. Geschwind, R. Goin-Kochel, E. Hanson, D. Grice, A. Klin, D. Ledbetter, C. Lord, C. Martin, D. Martin, R. Maxim, J. Miles, O. Ousley, K. Pelphrey, B. Peterson, J. Piggot, C. Saulnier, M. State, W. Stone, J. Sutcliffe, C. Walsh, Z. Warren, E. Wijsman).We appreciate obtaining access to phenotypic data on SFARI Base. Approved researchers can obtain the SSC data set described in this study by applying at https://base.sfari.org.

The AGP data sets used for the analysis described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number, phs000267.v1.p1. Submission of the data, phs000267.v1.p1, to dbGaP was provided by Dr Bernie Devlin on behalf of the Autism Genome Project (AGP). Collection and submission of the data to dbGaP were supported by a grant from the Medical Research Council (G0601030) and the Wellcome Trust (075491/Z/04), Anthony P. Monaco, P.I., University of Oxford.

Control data used in this study were provided by the NUgene Project (www.nugene.org). Funding support for the NUgene Project was provided by the Northwestern University's Center for Genetic Medicine, Northwestern University, and Northwestern Memorial Hospital. Assistance with phenotype harmonization was provided by the eMERGE Coordinating Center (Grant number U01HG04603). This study was funded through the NIH, NHGRI eMERGE Network (U01HG004609). Funding support for genotyping, which was performed at The Broad Institute, was provided by the NIH (U01HG004424). Assistance with phenotype harmonization and genotype data cleaning was provided by the eMERGE Administrative Coordinating Center (U01HG004603) and the National Center for Biotechnology Information (NCBI). The data sets used for the analyses described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number phs000237.v1.p1.

Conflict of Interest statement. None declared.

FUNDING

This work is supported by a Young Investigator Award (17379) from NARSAD/Brain and Behavioral Research Foundation and L.A.W. is supported by the International Mental Health Research Organization.

REFERENCES

1. Bailey, A., Lecouteur, A., Gottesman, I., Bolton, P., Simonoff, E., Yuzda, E. and Rutter, M. (1995) Autism as a strongly genetic disorder—evidence from a British twin study. *Psychol. Med.*, **25**, 63–77.
2. Folstein, S. and Rutter, M. (1977) Infantile-autism—genetic study of 21 twin pairs. *J. Child Psychol. Psychiatry*, **18**, 297–321.
3. Steffenburg, S., Gillberg, C., Hellgren, L., Andersson, L., Gillberg, I.C., Jakobsson, G. and Bohman, M. (1989) A twin study of autism in Denmark, Finland, Iceland, Norway and Sweden. *J. Child Psychol. Psychiatry*, **30**, 405–416.
4. Constantino, J.N., Zhang, Y., Frazier, T., Abbacchi, A.M. and Law, P. (2010) Sibling recurrence and the genetic epidemiology of autism. *Am. J. Psychiatry*, **167**, 1349–1356.
5. Hallmayer, J., Cleveland, S., Torres, A., Phillips, J., Cohen, B., Torigoe, T., Miller, J., Fedele, A., Collins, J., Smith, K. *et al.* (2011) Genetic heritability and shared environmental factors among twin pairs with autism. *Arch. Gen. Psychiatry*, **68**, 1095–1102.
6. Constantino, J.N., Hudziak, J.J. and Todd, R.D. (2000) The genetic structure of reciprocal social behavior: support for a population based approach to genetic studies of autism. *Am. J. Med. Genet.*, **96**, 480–480.
7. Constantino, J.N., Gruber, C.P., Davis, S., Hayes, S., Passanante, N. and Przybeck, T. (2004) The factor structure of autistic traits. *J. Child Psychol. Psychiatry*, **45**, 719–726.
8. LeCouteur, A., Bailey, A., Goode, S., Pickles, A., Robertson, S., Gottesman, I. and Rutter, M. (1996) A broader phenotype of autism: the clinical spectrum in twins. *J. Child Psychol. Psychiatry*, **37**, 785–801.
9. Muhle, R., Trentacoste, S.V. and Rapin, I. (2004) The genetics of autism. *Pediatrics*, **113**, E472–E486.
10. Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., Yamrom, B., Yoon, S., Krasnitz, A., Kendall, J. *et al.* (2007) Strong association of de novo copy number mutations with autism. *Science*, **316**, 445–449.
11. Levy, D., Ronennus, M., Yamrom, B., Lee, Y.H., Leotta, A., Kendall, J., Marks, S., Lakshmi, B., Pai, D., Ye, K. *et al.* (2011) Rare de novo and transmitted copy-number variation in autistic spectrum disorders. *Neuron*, **70**, 886–897.
12. Sanders, S.J., Ercan-Sencicek, A.G., Hus, V., Luo, R., Murtha, M.T., Moreno-De-Luca, D., Chu, S.H., Moreau, M.P., Gupta, A.R., Thomson, S.A. *et al.* (2011) Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron*, **70**, 863–885.
13. Marshall, C.R., Noor, A., Vincent, J.B., Lionel, A.C., Feuk, L., Skaug, J., Shago, M., Moessner, R., Pinto, D., Ren, Y. *et al.* (2008) Structural variation of chromosomes in autism spectrum disorder. *Am. J. Hum. Genet.*, **82**, 477–488.
14. Bucan, M., Abrahams, B.S., Wang, K., Glessner, J.T., Herman, E.I., Sonnenblick, L.I., Retuerto, A.I.A., Imielinski, M., Hadley, D., Bradfield, J.P. *et al.* (2009) Genome-wide analyses of exonic copy number variants in a family-based study point to novel autism susceptibility genes. *PLoS Genet.*, **5**, 12.
15. Szatmari, P., Paterson, A.D., Zwaigenbaum, L., Roberts, W., Brian, J., Liu, X.Q., Vincent, J.B., Skaug, J.L., Thompson, A.P., Senman, L. *et al.* (2007) Mapping autism risk loci using genetic linkage and chromosomal rearrangements. *Nat. Genet.*, **39**, 319–328.
16. Weiss, L.A., Shen, Y.P., Korn, J.M., Arking, D.E., Miller, D.T., Fossdal, R., Saemundsen, E., Stefansson, H., Ferreira, M.A.R., Green, T. *et al.* (2008) Association between microdeletion and microduplication at 16p11.2 and autism. *N. Engl. J. Med.*, **358**, 667–675.
17. Weiss, L.A., Arking, D.E., Daly, M.J. and Chakravarti, A. (2009) A genome-wide linkage and association scan reveals novel loci for autism. *Nature*, **461**, 802–808.
18. Glessner, J.T., Wang, K., Cai, G.Q., Korvatska, O., Kim, C.E., Wood, S., Zhang, H.T., Estes, A., Brune, C.W., Bradfield, J.P. *et al.* (2009) Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. *Nature*, **459**, 569–573.
19. Maestrini, E., Pagnamenta, A.T., Lamb, J.A., Bacchelli, E., Sykes, N.H., Sousa, I., Toma, C., Barnby, G., Butler, H., Winchester, L. *et al.* (2010) High-density SNP association study and copy number variation analysis of the AUTS1 and AUTS5 loci implicate the IMMP2L-DOCK4 gene region in autism susceptibility. *Mol. Psychiatry*, **15**, 954–968.
20. Anney, R., Klei, L., Pinto, D., Regan, R., Conroy, J., Magalhaes, T.R., Correia, C., Abrahams, B.S., Sykes, N., Pagnamenta, A.T. *et al.* (2010) A genome-wide scan for common alleles affecting risk for autism. *Hum. Mol. Genet.*, **19**, 4072–4082.
21. Ma, D.Q., Salyakina, D., Jaworski, J.M., Konidari, I., Whitehead, P.L., Andersen, A.N., Hoffman, J.D., Slifer, S.H., Hedges, D.J., Cukier, H.N. *et al.* (2009) A genome-wide association study of autism reveals a common novel risk locus at 5p14.1. *Ann. Hum. Genet.*, **73**, 263–273.
22. Wang, K., Zhang, H.T., Ma, D.Q., Bucan, M., Glessner, J.T., Abrahams, B.S., Salyakina, D., Imielinski, M., Bradfield, J.P., Sleiman, P.M.A. *et al.* (2009) Common genetic variants on 5p14.1 associate with autism spectrum disorders. *Nature*, **459**, 528–533.
23. Pinto, D., Pagnamenta, A.T., Klei, L., Anney, R., Merico, D., Regan, R., Conroy, J., Magalhaes, T.R., Correia, C., Abrahams, B.S. *et al.* (2010) Functional impact of global rare copy number variation in autism spectrum disorders. *Nature*, **466**, 368–372.
24. Collins, M.O., Husi, H., Yu, L., Brandon, J.M., Anderson, C.N.G., Blackstock, W.P., Choudhary, J.S. and Grant, S.G.N. (2006) Molecular characterization and comparison of the components and multiprotein complexes in the postsynaptic proteome. *J. Neurochem.*, **97**, 16–23.
25. Adams, R.H., Betz, H. and Puschel, A.W. (1996) A novel class of murine semaphorins with homology to thrombospondin is differentially expressed during early embryogenesis. *Mech. Dev.*, **57**, 33–45.
26. Melin, M., Carlsson, B., Anckarsater, H., Rastam, M., Betancur, C., Isaksson, A., Gillberg, C. and Dahl, N. (2006) Constitutional downregulation of SEMA5A expression in autism. *Neuropsychobiology*, **54**, 64–69.
27. Gamazon, E.R., Zhang, W., Konkashbaev, A., Duan, S.W., Kistner, E.O., Nicolae, D.L., Dolan, M.E. and Cox, N.J. (2010) SCAN: SNP and copy number annotation. *Bioinformatics*, **26**, 259–262.
28. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.
29. Conrad, D.F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y.J., Aerts, J., Andrews, T.D., Barnes, C., Campbell, P. *et al.* (2010) Origins and functional impact of copy number variation in the human genome. *Nature*, **464**, 704–712.
30. Anney, R., Klei, L., Pinto, D., Almeida, J., Bacchelli, E., Baird, G., Bolshakova, N., Bolte, S., Bolton, P.F., Bourgeron, T. *et al.* (2012) Individual common variants exert weak effects on the risk for autism spectrum disorderspi. *Hum. Mol. Genet.*, **21**, 4781–4792.
31. Klei, L., Sanders, S.J., Murtha, M.T., Hus, V., Lowe, J.K., Willsey, A.J., Moreno-De-Luca, D., Yu, T.W., Fombonne, E., Geschwind, D. *et al.* (2012) Common genetic variants, acting additively, are a major source of risk for autism. *Mol. Autism*, **3**, 9.
32. Devlin, B., Melhem, N. and Roeder, K. (2011) Do common variants play a role in risk for autism? Evidence and theoretical musings. *Brain Res.*, **1380**, 78–84.
33. Devlin, B. and Scherer, S.W. (2012) Genetic architecture in autism spectrum disorder. *Curr. Opin. Genet. Dev.*, **22**, 229–237.
34. Henrichsen, C.N., Chaignat, E. and Reymond, A. (2009) Copy number variants, diseases and gene expression. *Hum. Mol. Genet.*, **18**, R1–R8.
35. Nord, A.S., Roeb, W., Dickel, D.E., Walsh, T., Kusenda, M., O'Connor, K.L., Malhotra, D., McCarthy, S.E., Stray, S.M., Taylor, S.M. *et al.* (2011) Reduced transcript expression of genes affected by inherited and de novo CNVs in autism. *Eur. J. Hum. Genet.*, **19**, 727–731.
36. Luo, R., Sanders, S.J., Tian, Y., Voineagu, I., Huang, N., Chu, S.H., Klei, L., Cai, C., Ou, J., Lowe, J.K. *et al.* (2012) Genome-wide transcriptome profiling reveals the functional impact of rare de novo and recurrent CNVs in autism spectrum disorders. *Am. J. Hum. Genet.*, **91**, 38–55.
37. Gamazon, E.R., Nicolae, D.L. and Cox, N.J. (2011) A study of CNVs as trait-associated polymorphisms and as expression quantitative trait loci. *PLoS Genet.*, **7**, e1001292.

38. Henrichsen, C.N., Vinckenbosch, N., Zollner, S., Chaignat, E., Pradervand, S., Schutz, F., Ruedi, M., Kaessmann, H. and Reymond, A. (2009) Segmental copy number variation shapes tissue transcriptomes. *Nat. Genet.*, **41**, 424–429.
39. Davis, L.K., Gamazon, E.R., Kistner-Griffin, E., Badner, J.A., Liu, C., Cook, E.H., Sutcliffe, J.S. and Cox, N.J. (2012) Loci nominally associated with autism from genome-wide analysis show enrichment of brain expression quantitative trait loci but not lymphoblastoid cell line expression quantitative trait loci. *Mol Autism*, **3**, 3.
40. Hu, V.W., Frank, B.C., Heine, S., Lee, N.H. and Quackenbush, J. (2006) Gene expression profiling of lymphoblastoid cell lines from monozygotic twins discordant in severity of autism reveals differential regulation of neurologically relevant genes. *BMC Genomics*, **7**, 118.
41. Jasinska, A.J., Service, S., Choi, O.W., DeYoung, J., Grujic, O., Kong, S.Y., Jorgensen, M.J., Bailey, J., Breidenthal, S., Fairbanks, L.A. *et al.* (2009) Identification of brain transcriptional variation reproduced in peripheral blood: an approach for mapping brain expression traits. *Hum. Mol. Genet.*, **18**, 4415–4427.
42. Hu, V.W., Nguyen, A., Kim, K.S., Steinberg, M.E., Sarachana, T., Scully, M.A., Soldin, S.J., Luu, T. and Lee, N.H. (2009) Gene expression profiling of lymphoblasts from autistic and nonaffected sib pairs: altered pathways in neuronal development and steroid biosynthesis. *PLoS ONE*, **4**, e5775.
43. Baron, C.A., Liu, S.Y., Hicks, C. and Gregg, J.P. (2006) Utilization of lymphoblastoid cell lines as a system for the molecular modeling of autism. *J. Autism Dev. Disord.*, **36**, 973–982.
44. Sarachana, T., Zhou, R., Chen, G., Manji, H.K. and Hu, V.W. (2010) Investigation of post-transcriptional gene regulatory networks associated with autism spectrum disorders by microRNA expression profiling of lymphoblastoid cell lines. *Genome Med.*, **2**, 23.
45. Nishimura, Y., Martin, C.L., Vazquez-Lopez, A., Spence, S.J., Alvarez-Retuerto, A.I., Sigman, M., Steindler, C., Pellegrini, S., Schanen, N.C., Warren, S.T. *et al.* (2007) Genome-wide expression profiling of lymphoblastoid cell lines distinguishes different forms of autism and reveals shared pathways. *Hum. Mol. Genet.*, **16**, 1682–1698.
46. Nguyen, A., Rauch, T.A., Pfeifer, G.P. and Hu, V.W. (2010) Global methylation profiling of lymphoblastoid cell lines reveals epigenetic contributions to autism spectrum disorders and a novel autism candidate gene, RORA, whose protein product is reduced in autistic brain. FASEB J.*, 24*, 3036–3051.
47. Sultana, R., Yu, C.E., Yu, J., Munson, J., Chen, D.H., Hua, W.H., Estes, A., Cortes, F., de la Barra, F., Yu, D.M. *et al.* (2002) Identification of a novel gene on chromosome 7q11.2 interrupted by a translocation breakpoint in a pair of autistic twins. *Genomics*, **80**, 129–134.
48. Huang, X.L., Zou, Y.S., Maher, T.A., Newton, S. and Milunsky, J.M. (2010) A de novo balanced translocation breakpoint truncating the Autism Susceptibility Candidate 2 (AUTS2) gene in a patient with autism. *Am. J. Med. Genet. A*, **152A**, 2112–2114.
49. Beunders, G., Voorhoeve, E., Golzio, C., Pardo, L.M., Rosenfeld, J.A., Talkowski, M.E., Simonic, I., Lionel, A.C., Vergult, S., Pyatt, R.E. *et al.* (2013) Exonic deletions in AUTS2 cause a syndromic form of intellectual disability and suggest a critical role for the C terminus. *Am. J. Hum. Genet*, **92**, 210–220.
50. Nagamani, S.C., Erez, A., Ben-Zeev, B., Frydman, M., Winter, S., Zeller, R., El-Khechen, D., Escobar, L., Stankiewicz, P., Patel, A. *et al.* (2012) Detection of copy-number variation in AUTS2 gene by targeted exonic array CGH in patients with developmental delay and autistic spectrum disorders. *Eur. J. Hum. Genet*, **21**, 343–346.
51. Talkowski, M.E., Rosenfeld, J.A., Blumenthal, I., Pillalamarri, V., Chiang, C., Heilbut, A., Ernst, C., Hanscom, C., Rossin, E., Lindgren, A.M. *et al.* (2012) Sequencing chromosomal abnormalities reveals neurodevelopmental loci that confer risk across diagnostic boundaries. *Cell*, **149**, 525–537.
52. Kalscheuer, V.M., FitzPatrick, D., Tommerup, N., Bugge, M., Niebuhr, E., Neumann, L.M., Tzschach, A., Shoichet, S.A., Menzel, C., Erdogan, F. *et al.* (2007) Mutations in autism susceptibility candidate 2 (AUTS2) in patients with mental retardation. *Hum. Genet.*, **121**, 501–509.
53. IMGSAC (1998) A full genome screen for autism with evidence for linkage to a region on chromosome 7q. International Molecular Genetic Study of Autism Consortium. *Hum. Mol. Genet.*, **7**, 571–578.
54. IMGSAC (2001) A genomewide screen for autism: strong evidence for linkage to chromosomes 2q, 7q, and 16p. *Am. J. Hum. Genet.*, **69**, 570–581.
55. Badner, J.A. and Gershon, E.S. (2002) Regional meta-analysis of published data supports linkage of autism with markers on chromosome 7. *Mol. Psychiatry*, **7**, 56–66.
56. Zhiling, Y., Fujita, E., Tanabe, Y., Yamagata, T., Momoi, T. and Momoi, M.Y. (2008) Mutations in the gene encoding CADM1 are associated with autism spectrum disorder. *Biochem. Biophys. Res. Commun.*, **377**, 926–929.
57. Hamdan, F.F., Daoud, H., Rochefort, D., Piton, A., Gauthier, J., Langlois, M., Foomani, G., Dobrzeniecka, S., Krebs, M.O., Joober, R. *et al.* (2010) De novo mutations in FOXP1 in cases with intellectual disability, autism, and language impairment. *Am. J. Hum. Genet.*, **87**, 671–678.
58. O'Roak, B.J., Deriziotis, P., Lee, C., Vives, L., Schwartz, J.J., Girirajan, S., Karakoc, E., Mackenzie, A.P., Ng, S.B., Baker, C. *et al.* (2011) Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nat. Genet.*, **43**, 585–589.
59. Palumbo, O., D'Agruma, L., Minenna, A.F., Palumbo, P., Stallone, R., Palladino, T., Zelante, L. and Carella, M. (2012) 3p14.1 de novo microdeletion involving the FOXP1 gene in an adult patient with autism, severe speech delay and deficit of motor coordination. *Gene*, **516**, 107–113.
60. Ozgen, H.M., van Daalen, E., Bolton, P.F., Maloney, V.K., Huang, S., Cresswell, L., van den Boogaard, M.J., Eleveld, M.J., van't Slot, R., Hochstenbach, R. *et al.* (2009) Copy number changes of the microcephalin 1 gene (MCPH1) in patients with autism spectrum disorders. *Clin. Genet.*, **76**, 348–356.
61. Glancy, M., Barnicoat, A., Vijeratnam, R., de Souza, S., Gilmore, J., Huang, S.W., Maloney, V.K., Thomas, N.S., Bunyan, D.J., Jackson, A. *et al.* (2009) Transmitted duplication of 8p23.1–8p23.2 associated with speech delay, autism and learning difficulties. *Eur. J. Hum. Genet.*, **17**, 37–43.
62. Chung, B.H., Mullegama, S., Marshall, C.R., Lionel, A.C., Weksberg, R., Dupuis, L., Brick, L., Li, C., Scherer, S.W., Aradhya, S. *et al.* (2011) Severe intellectual disability and autistic features associated with microduplication 2q23.1. *Eur. J. Hum. Genet.*, **20**, 398–403.
63. Talkowski, M.E., Mullegama, S.V., Rosenfeld, J.A., van Bon, B.W., Shen, Y., Repnikova, E.A., Gastier-Foster, J., Thrush, D.L., Kathiresan, S., Ruderfer, D.M. *et al.* (2011) Assessment of 2q23.1 microdeletion syndrome implicates MBD5 as a single causal locus of intellectual disability, epilepsy, and autism spectrum disorder. *Am. J. Hum. Genet.*, **89**, 551–563.
64. Cukier, H.N., Lee, J.M., Ma, D., Young, J.I., Mayo, V., Butler, B.L., Ramsook, S.S., Rantus, J.A., Abrams, A.J., Whitehead, P.L. *et al.* (2012) The expanding role of MBD genes in autism: identification of a MECP2 duplication and novel alterations in MBD5, MBD6, and SETDB1. *Autism Res.*, **5**, 385–397.
65. Pagnamenta, A.T., Khan, H., Walker, S., Gerrelli, D., Wing, K., Bonaglia, M.C., Giorda, R., Berney, T., Mani, E., Molteni, M. *et al.* (2011) Rare familial 16q21 microdeletions under a linkage peak implicate cadherin 8 (CDH8) in susceptibility to autism and learning disability. *J. Med. Genet.*, **48**, 48–54.
66. Martin, C.L., Duvall, J.A., Ilkin, Y., Simon, J.S., Arreaza, M.G., Wilkes, K., Alvarez-Retuerto, A., Whichello, A., Powell, C.M., Rao, K. *et al.* (2007) Cytogenetic and molecular characterization of A2BP1/FOX1 as a candidate gene for autism. *Am. J. Med. Genet. B Neuropsychiatr. Genet.*, **144B**, 869–876.
67. Davis, L.K., Maltman, N., Mosconi, M.W., Macmillan, C., Schmitt, L., Moore, K., Francis, S.M., Jacob, S., Sweeney, J.A. and Cook, E.H. (2012) Rare inherited A2BP1 deletion in a proband with autism and developmental hemiparesis. *Am. J. Med. Genet. A*, **158A**, 1654–1661.
68. Lim, E.T., Raychaudhuri, S., Sanders, S.J., Stevens, C., Sabo, A., Macarthur, D.G., Neale, B.M., Kirby, A., Ruderfer, D.M., Fromer, M. *et al.* (2013) Rare complete knockouts in humans: population distribution and significant role in autism spectrum disorders. *Neuron*, **77**, 235–242.
69. Fatemi, S.H., Snow, A.V., Stary, J.M., Araghi-Niknam, M., Reutiman, T.J., Lee, S., Brooks, A.I. and Pearce, D.A. (2005) Reelin signaling is impaired in autism. *Biol. Psychiatry*, **57**, 777–787.
70. Artigiani, S., Conrotto, P., Fazzari, P., Gilestro, G.F., Barberis, D., Giordano, S., Comoglio, P.M. and Tamagnone, L. (2004) Plexin-B3 is a functional receptor for semaphorin 5A. *EMBO Rep.*, **5**, 710–714.
71. Matsuoka, R.L., Chivatakarn, O., Badea, T.C., Samuels, I.S., Cahill, H., Katayama, K., Kumar, S.R., Suto, F., Chedotal, A., Peachey, N.S. *et al.* (2011) Class 5 transmembrane semaphorins control selective Mammalian retinal lamination and function. *Neuron*, **71**, 460–473.

72. Kantor, D.B., Chivatakarn, O., Peer, K.L., Oster, S.F., Inatani, M., Hansen, M.J., Flanagan, J.G., Yamaguchi, Y., Sretavan, D.W., Giger, R.J. *et al.* (2004) Semaphorin 5A is a bifunctional axon guidance cue regulated by heparan and chondroitin sulfate proteoglycans. *Neuron*, **44**, 961–975.
73. Sadanandam, A., Rosenbaugh, E.G., Singh, S., Varney, M. and Singh, R.K. (2010) Semaphorin 5A promotes angiogenesis by increasing endothelial cell proliferation, migration, and decreasing apoptosis. *Microvasc. Res.*, **79**, 1–9.
74. Fiore, R., Rahim, B., Christoffels, V.M., Moorman, A.F. and Puschel, A.W. (2005) Inactivation of the Sema5a gene results in embryonic lethality and defective remodeling of the cranial vascular system. *Mol. Cell Biol.*, **25**, 2310–2319.
75. Gunn, R.K., Huentelman, M.J. and Brown, R.E. (2011) Are Sema5a mutant mice a good model of autism? A behavioral analysis of sensory systems, emotionality and cognition. *Behav. Brain Res.*, **225**, 142–150.
76. Konopka, G., Wexler, E., Rosen, E., Mukamel, Z., Osborn, G.E., Chen, L., Lu, D., Gao, F., Gao, K., Lowe, J.K. *et al.* (2011) Modeling the functional genomics of autism using human neurons. *Mol. Psychiatry*, **17**, 202–214.
77. Chen, Y., Zhu, J., Lum, P.Y., Yang, X., Pinto, S., MacNeil, D.J., Zhang, C., Lamb, J., Edwards, S., Sieberts, S.K. *et al.* (2008) Variations in DNA elucidate molecular networks that cause disease. *Nature*, **452**, 429–435.
78. Emilsson, V., Thorleifsson, G., Zhang, B., Leonardson, A.S., Zink, F., Zhu, J., Carlson, S., Helgason, A., Walters, G.B., Gunnarsdottir, S. *et al.* (2008) Genetics of gene expression and its effect on disease. *Nature*, **452**, 423–428.
79. Duan, S., Huang, R.S., Zhang, W., Bleibel, W.K., Roe, C.A., Clark, T.A., Chen, T.X., Schweitzer, A.C., Blume, J.E., Cox, N.J. *et al.* (2008) Genetic architecture of transcript-level variation in humans. *Am. J. Hum. Genet.*, **82**, 1101–1113.
80. Zhang, W., Duan, S., Kistner, E.O., Bleibel, W.K., Huang, R.S., Clark, T.A., Chen, T.X., Schweitzer, A.C., Biume, J.E., Cox, N.J. *et al.* (2008) Evaluation of genetic variation contributing to differences in gene expression between populations. *Am. J. Hum. Genet.*, **82**, 631–640.
81. Gibbs, R.A., Belmont, J.W., Hardenbol, P., Willis, T.D., Yu, F.L., Yang, H.M., Ch'ang, L.Y., Huang, W., Liu, B., Shen, Y. *et al.* (2003) The International HapMap Project. *Nature*, **426**, 789–796.
82. Geschwind, D.H., Sowinski, J., Lord, C., Iversen, P., Shestack, J., Jones, P., Ducat, L. and Spence, S.J. (2001) The Autism Genetic Resource Exchange: a resource for the study of autism and related neuropsychiatric conditions. *Am. J. Hum. Genet.*, **69**, 463–466.
83. Fischbach, G.D. and Lord, C. (2010) The Simons simplex collection: a resource for identification of autism genetic risk factors. *Neuron*, **68**, 192–195.
84. Xu, J., Zwaigenbaum, L., Szatmari, P. and Scherer, S.W. (2004) Molecular cytogenetics of autism. *Curr. Genomics*, **5**, 347–364.
85. Wang, K., Li, M.Y., Hadley, D., Liu, R., Glessner, J., Grant, S.F.A., Hakonarson, H. and Bucan, M. (2007) PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.*, **17**, 1665–1674.
86. Iafrate, A.J., Feuk, L., Rivera, M.N., Listewnik, M.L., Donahoe, P.K., Qi, Y., Scherer, S.W. and Lee, C. (2004) Detection of large-scale variation in the human genome. *Nat. Genet.*, **36**, 949–951.
87. Browning, B.L. and Browning, S.R. (2009) A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.*, **84**, 210–223.
88. DerSimonian, R. and Laird, N. (1986) Meta-analysis in clinical trials. *Control Clin. Trials*, **7**, 177–188.
89. McCarty, C.A., Chisholm, R.L., Chute, C.G., Kullo, I.J., Jarvik, G.P., Larson, E.B., Li, R., Masys, D.R., Ritchie, M.D., Roden, D.M. *et al.* (2011) The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Med. Genomics*, **4**, 13.