

Bioinformatic analysis of autism positional candidate genes using biological databases and computational gene network prediction

A. L. Yonan[†,‡], A. A. Palmer[†], K. C. Smith[†], I. Feldman[†,††], H. K. Lee[†], J. M. Yonan[§], S. G. Fischer[†], P. Pavlidis[†,††] and T. C. Gilliam[*,†,‡,¶]

[†]*Columbia Genome Center, Columbia University, New York,*
[‡]*Department of Genetics and Development, Columbia University, New York,*
[¶]*Department of Psychiatry, Columbia University and New York State Psychiatric Institute, New York,*
[§]*Division of Molecular Genetics, Departments of Pediatrics and Medicine, Columbia University, New York,*
[††]*Department of Biomedical Informatics, Columbia University, New York, USA*
**Corresponding author: T. C. Gilliam, Columbia Genome Center, 1150 St. Nicholas Avenue, Room 508, New York, NY 10032, USA. E-mail: tcg1@columbia.edu*

Common genetic disorders are believed to arise from the combined effects of multiple inherited genetic variants acting in concert with environmental factors, such that any given DNA sequence variant may have only a marginal effect on disease outcome. As a consequence, the correlation between disease status and any given DNA marker allele in a genomewide linkage study tends to be relatively weak and the implicated regions typically encompass hundreds of positional candidate genes. Therefore, new strategies are needed to parse relatively large sets of 'positional' candidate genes in search of actual disease-related gene variants. Here we use biological databases to identify 383 positional candidate genes predicted by genomewide genetic linkage analysis of a large set of families, each with two or more members diagnosed with autism, or autism spectrum disorder (ASD). Next, we seek to identify a subset of biologically meaningful, high priority candidates. The strategy is to select autism candidate genes based on prior genetic evidence from the allelic association literature to query the known transcripts within the 1-LOD (logarithm of the odds) support interval for each region. We use recently developed bioinformatic programs that automatically search the biological literature to predict pathways of interacting genes (PATHWAYASSIST and GENEWAYS). To identify gene regulatory networks, we search for coexpression between candidate genes and positional candidates. The studies are intended both to inform studies of autism, and to illustrate and explore the increasing potential of bioinformatic approaches as a compliment to linkage analysis.

Keywords: 17q, AGRE sample, autism, association studies, bioinformatics, candidate genes

Received 30 June 2003, revised 20 August 2003, accepted for publication 21 August 2003

Autism is a pervasive neurodevelopmental disorder that severely impairs development of normal social and emotional interactions and related forms of communication. Disease symptoms characteristically include unusually restricted and stereotyped patterns of behaviors and interests. Autism describes the most severe manifestation of a broad spectrum of disorders, known as autism spectrum disorders (ASD) that share these essential features, but vary in their degree of severity and/or age of onset. While it is difficult to accurately estimate the prevalence of ASD, due to an apparent increase over the past few decades (Chakrabarti & Fombonne 2001; Gillberg & Wing 1999; Prior 2003), recent studies suggest that ASD affects 34–60 individuals per 10 000 (Charman 2002; Fombonne 2003; Yeargin-Allsopp *et al*. 2003).

Twin and epidemiological studies show that autism is a highly heritable disorder. When one monozygotic (MZ) twin is diagnosed with autism or ASD, the disease concordance is 70–90%, compared to 0–25% concordance among same-sex dizygotic twins (Bailey *et al*. 1995; Folstein & Rutter 1977; Lauritsen & Ewald 2001; Rutter 2000). The estimated heritability of ASD is believed to be approximately 90%, which is extremely high relative to other complex genetic diseases (Hyttinen *et al*. 2003; Ju *et al*. 2000). The impact of genetic determinants on disease liability is further substantiated by comparing the disease risk for a sibling of a proband diagnosed with ASD (2–6%) with the population prevalence of ASD (0.04–0.1%) (Smalley 1997; Smalley *et al*. 1988; Szatmari *et al*. 1998), yielding a relative risk of 50–100 for ASD (Lamb *et al*. 2000).The rate by which autism and ASD incidence drops among first, second and third degree relatives provides another indication that disease susceptibility arises from the combined effects of multiple, possibly interacting, genes (Lamb *et al*. 2000; Rutter 2000). Therefore, even though autism is clearly among the most heritable of

doi: 10.1046/j.1601-183X.2003.00041.x

all psychiatric disorders, the likely interaction of multiple genes that increase susceptibility to autism, rather than directly cause it, presents formidable challenges for genetic studies.

The search for genetic linkage between DNA markers spanning the entire genome and single-gene disorders with clear Mendelian patterns of inheritance has been enormously successful, in many cases leading to the identification of disease genes and their causal mutations despite years of failure using non-genetic, hypothesis-driven approaches (Botstein & Risch 2003). The success of such studies depends upon the identification of clear recombinant breakpoints that define the boundaries of the disease locus, and typically demarcate a minimal genetic region that harbors the disease gene along with dozens of non-disease related, positional candidate genes (Riordan *et al*. 1989; Rommens *et al*. 1989). Whereas 'single-gene' disorders are typically quite rare, common heritable disorders are believed to arise from the combined effects of multiple predisposing gene variants, presumably in combination with environmental factors. Consequently, the influence of any single gene-variant upon disease status is likely to be small, and therefore difficult to detect using genetic linkage strategies. Moreover, the population prevalence of gene variants with small or negligible individual effects upon reproductive fitness will follow the same stochastic course as neutral polymorphisms, in some instances reaching significant frequencies. This explains in part how heritable disorders with multiple gene etiologies become common, and also why they are elusive gene mapping targets, i.e., it becomes difficult to detect enhanced sharing of disease-related alleles among affected individuals when the same gene variant is prevalent among control individuals. For these reasons and others (Altmuller *et al*. 2001; Lander & Kruglyak 1995; Lander & Schork 1994; Weiss & Terwilliger 2000), evidence for linkage between a common heritable disorder and DNA marker alleles tends to be weak and difficult to distinguish from the type of random statistical fluctuations that inevitably accompany a full genome scan. Consequently, a conservative survey of positional candidate genes based upon whole genome scan analysis typically requires the analysis of positional candidate genes within multiple, broad linkage peaks, often spanning 10–40 million base pairs, and comprising upwards of 50–100 genes.

Consistent with these rather dire predictions, we recently completed the largest whole genome linkage scan of ASD reported to date, and found no statistically significant evidence for linkage between DNA marker alleles and disease status (Yonan *et al*. in press). We did, however, detect 'suggestive' evidence for ASD predisposing loci on chromosomes 17, 5, 11, 4 and 8. Such moderate linkage signals may reflect the marginal contribution to disease risk arising from a given genetic locus, or alternatively, false positive findings that reflect random statistical fluctuation. While independent replication is the standard to distinguish between the two possibilities, the criteria required to declare replication are model and disease dependent, and thus necessarily vague, and at least in theory, replication of a specific linkage finding is many times more complex than detection of any one among several predisposing genetic loci (Lander & Kruglyak 1995).

For reasons outlined above, whole genome linkage analysis of common heritable disorders identifies a large and unmanageable number of positional candidate genes, the vast majority of which are unrelated to the disease target. We propose the use of genomic data-mining strategies to parse these relatively large candidate gene sets with the purpose of identifying a subset of biologically meaningful genes that map to predetermined genetic loci. To illustrate this approach, we have surveyed the top five ASD-linked regions in a recent genomewide linkage study (Yonan *et al*. in press). The strategy is to predict a subset of likely candidate genes mapping to each putative linkage peak. Such candidates would then become the focus of further genetic and biological testing.

There is substantial interest in using bioinformatic resources in conjunction with linkage methodologies to identify the most promising candidate loci within large and sometimes unconfirmed linkage regions, so that they may be examined further (Baron 2002). We chose to use positively associated genes to query known transcripts within peak linkage regions using several complimentary bioinformatic methods. We examined several different bioinformatic approaches in order to identify convergent evidence for specific candidate genes, as well as to explore the future potential and current limitations of these approaches.

Materials and methods

Characterization of putative ASD-linked chromosomal regions

The chromosomal regions examined in this study are shown in Fig. 1. Beginning with 345 families that had two or more siblings diagnosed with either autism or ASD, we used affected sib pair analysis to identify genomewide linkage to ASD (Yonan *et al*. in press). Five chromosomal regions from the genome scan met a cutoff of a pointwise *P*-value of < 0.01, which we interpreted as being 'moderately suggestive'. Here we examine the chromosomal regions defined by the 1-LOD support interval of the 5 most significant peaks. Details of the analysis that lead to the identification of these regions have been described previously (Liu *et al*. 2001; Yonan *et al*. in press).

Association and linkage tables

We performed a search for allelic association between candidate gene allelic variants and autism or ASD using the PubMed database (http://www.ncbi.nlm.nih.gov/). This

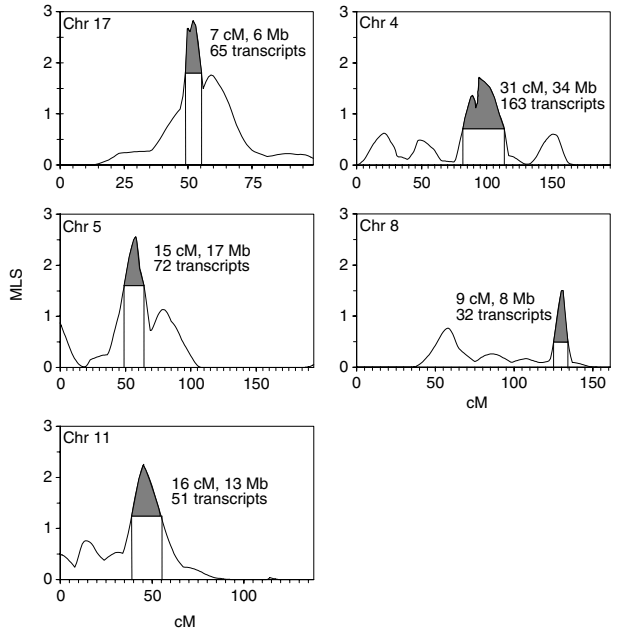

Figure 1: The 1-LOD interval of the five most significant multipoint Maximum Likelihood Score (MLS) regions from genomewide Affected Sib Pair analysis to ASD (Yonan *et al.* in press). The *x*-axis depicts genetic distance in Kosambi centimorgans from pter (zero coordinate) to qter; the *y*-axis represents MLS. The thick line and shading of the peaks demark the 1-LOD interval that defined each region. The size of the 1-LOD intervals are shown in Kosambi centimorgans. The physical distance, as well as the number of transcripts, was taken from the Human Genome Browser for each region, as described in *Materials and methods*.

search strategy was augmented by personal knowledge of the literature and by references from key publications (Table 1). A similar strategy was used to compile the list of genomewide linkage studies for autism and ASD (Table 2).

Manual search strategy

We compiled a comprehensive list of genes (known and predicted from transcripts) in our five most significant regions using the Celera Discovery System (http://www.celeradiscoverysystem.com) and the NCBI Human Genome Project (UCSC Genome Browser; http://genome.ucsc.edu/ version 24 (hg15) April 2003 Freeze) databases. This exhaustive gene list was created by performing database queries against the UCSC Human Genome Browser's annotation database. The table definitions and data of two MySQL (http://www.mysql.com) tables, refGene and refLink, were downloaded from the public FTP site at UCSC (ftp://genome.ucsc.edu/goldenPath/10april 2003/database/) and recreated locally. Genes that mapped to the corresponding intervals in the Celera map were downloaded manually. All genes located within the physical boundaries defined by the 1-LOD unit support intervals on each chromosome were then extracted; the complete list of these 383 genes is available as supplementary material accompanying this paper (see *Supplementary material* section). This list was then further evaluated using several online databases. The Celera database annotates category and family for each gene using the Panther Protein Function. The Human Genome Project provides a gene 'index', a set of links to multiple annotation databases, for each Ref Seq transcript, including to the Online Mendelian Inheritance of Man (OMIM), Locus Link, PubMed, Gene Lynx, Gene Cards and Ace View databases. A short list of 'neural-related' genes was identified based upon evidence of their involvement in neuronal development/ control, neurotransmitter function, transcription regulation and similar functions that made them logical disease-related candidates for the autism spectrum disorders.

Gene ontology methods

Gene Ontology (GO) is a controlled vocabulary designed to describe key aspects of the molecular function, biological process and cellular component of gene products (Bard 2003). Using the complete list of all 383 positional candidate genes (see above) we screened genes for neural-related GO terms in

Table 1: Summary of association studies for autism

Gene name	Physical location	Association found	Phenotype	Study Size and design	Reference	Overlapping linkage-MLS* scores	Linkage reference(s)
DRD5	4p16	No	Autistic disorder	38 families, TDT†	Philippe *et al.* (2002)	1.55	IMGSAC (1998)
DRD2	4q15	No	Autistic disorder	38 families, TDT	Philippe *et al.* (2002)		
HLA	6p21	No	Autistic disorder	20 patients vs. 709 controls	Stubbs *et al.* (1980)		
***HLA-DR* beta 1**	**6p21**	**Yes**	**Autistic disorder**	**50 patients vs. 79 controls**	**Warren *et al.* (1996)**		
GluR6	**6q21**	**Yes**	**Autistic disorder**	**107 trios, TDT and 51 families,**	**Jamain *et al.* (2002)**		

HOXA1	7p15	No	Autistic spectrum disorder (ASD)	ASP case ($n=35$) vs. control ($n=35$)	Talebizadeh *et al.* (2002)		
DLX6	7q21-q22	No	Autistic disorder	196 families, TDT	Nabi *et al.* (2003)	2.2; 3.2	CLSA (1999); IMGSAC (2001a)
PCLO	7q21-q22	No	Autistic disorder	196 families, TDT	Nabi *et al.* (2003)	2.2; 3.2	CLSA (1999); IMGSAC (2001a)
PAI-1	7q22	No	Autistic disorder	167 trios, linkage and association	Persico *et al.* (2001)	3.2	IMGSAC (2001a)
RELN	**7q22**	**Yes**	**ASD – with delayed phrase speech**	**126 families**	**Zhang *et al.* (2002)**	**3.2**	**IMGSAC (2001a)**
RELN	7q23	No	Autistic disorder	167 families, TDT	Krebs *et al.* (2002)		
FOXP2	7q31	No to FOXP2 gene; yes to region	Specific language impairment (SLI)	96 families, linkage and association	O'Brien *et al.* (2003		
GRM8	**7q31**	**Yes (haplotype)**	**Autistic disorder**	**196 families, TDT**	**Serajee *et al.* (2003)**		
WNT2	**7q31–33**	**Yes**	**Autistic with severe language abnormality**	**50 families**	**Wassink *et al.* (2001)**	**2.55–3.55**	**IMGSAC (1998)**
WNT2	7q31–33	No	Autistic or language abnormality	135 singleton and 82 multiplex families	McCoy *et al.* (2002)	2.55–3.55	IMGSAC (1998)
COPG2	7q32	No	Autistic disorder	169 families, TDT	Bonora *et al.* (2002)	2.55–3.55	IMGSAC (1998)
CPA1	7q32	No	Autistic disorder	169 families, TDT	Bonora *et al.* (2002)	2.55–3.55	IMGSAC (1998)
CPA5	7q32	No	Autistic disorder	169 families, TDT	Bonora *et al.* (2002)	2.55–3.55	IMGSAC (1998)
D7S1804	**7q32**	**Yes**	**Autistic spectrum disorder**	**170 multiplex families, TDT with 76 markers**	**IMGSAC (2001b)**	**2.55–3.55**	**IMGSAC (1998)**
PEG1 /MEST	7q32	No	Autistic disorder	169 families, TDT	Bonora *et al.* (2002)	2.55–3.55	IMGSAC (1998)
D7S2533	**7q33**	**Yes**	**Autistic spectrum disorder**	**170 multiplex families, TDT with 76 markers**	**IMGSAC (2001b)**		
EN2	7q36	No	Autistic spectrum disorder	204 AGRE families, TDT	Zhong *et al.* (2003)	3.66	Auranen *et al.* (2002)
PENK	8q11-q12	No	Autistic disorder	38 families, TDT	Philippe *et al.* (2002)		
BDNF	11p13	No	Autistic disorder	38 families, TDT	Philippe *et al.* (2002)		
HRAS	**11p15**	**Yes**	**Autistic disorder**	**case ($n=55$) vs. control ($n=55$)**	**Herault *et al.* (1995)**		
TH	11p15	No	Autistic disorder	38 families, TDT	Philippe *et al.* (2002)		
NCAM	11q22	No	Autistic disorder	38 families, TDT	Philippe *et al.* (2002)		
AVPR1A	12q14	Borderline significance	Autistic disorder	115 trios, MTDT§	Kim *et al.* (2002c)		

GABRA5	15q11-q13	No	Autistic disorder	226 families, PDT¶	Menold *et al.* (2001)		
GABRB3	**15q11-q13**	**Yes**	**Autistic disorder**	**80 families, TDT**	**Buxbaum *et al.* (2002)**		
GABRB3	15q11-q13	No	Autistic disorder	226 families, PDT	Menold *et al.*(2001)		
GABRG3	**15q11-q13**	**Yes**	**Autistic disorder**	**226 families, PDT**	**Menold *et al.* (2001)**		
ATP10C	15q11-q13	No	Autistic disorder	115 trios, TDT	Kim *et al.* (2002b)		
UBE3A	**15q11–q13**	**Yes**	**Autistic disorder**	**94 multiplex families, LD‡**	**Nurmi *et al.* (2001)**		
NF1	17q11	No	Autistic disorder	204 patients vs. 200 controls	Plank *et al.* (2001)	2.34, 2.83	IMGSAC (2001a), Yonan *et al.* (2003)
OMGP	**17q11**	**Yes**	**Autistic disorder (DQ** >30**)**	**case (*n* = 37) vs. control (*n* = 101)**	**Vourc'h *et al.* (2003)**	**2.34, 2.83**	**IMGSAC (2001a), Yonan *et al.* (2003)**
BLMH	**17q11**	**Yes**	**Autistic disorder**	**81 trios, TDT**	**Kim *et al.* (2002a)**	**2.34, 2.83**	**IMGSAC (2001a), Yonan *et al.* (2003)**
5-HTT /SLC6A4	**17q11**	**Yes**	**Autistic disorder**	**81 trios, TDT**	**Kim *et al.* (2002a)**	**2.34, 2.83**	**IMGSAC (2001a), Yonan *et al.* (2003)**
5-HTT /SLC6A4	17q11	No	Hyperserotoninemia in autistic patients	134 autistic patients vs. 291 1st degree relatives	Persico *et al.* (2002)	2.34, 2.83	IMGSAC (2001a) Yonan *et al.* (2003)
5-HTT /SLC6A4	17q11	No	5-HT blood levels	96 families, TDT	Betancur *et al.* (2002)	2.34, 2.83	IMGSAC (2001a), Yonan *et al.* (2003)
5-HTT /SLC6A4	17q11	No	Autistic disorder	98 trios, TDT	Persico *et al.* (2000a)	2.34, 2.83	IMGSAC (2001a), Yonan *et al.* (2003)
HOXB1	17q21	No	Autistic spectrum disorder	case (*n* = 35) vs. control (*n* = 35)	Talebizadeh *et al.* (2002)		
PCSK2	20p11	No	Autistic disorder	38 families, TDT	Philippe *et al.* (2002)		
PDYN	20p12	No	Autistic disorder	38 families, TDT	Philippe *et al.* (2002)		
ADA	**20q13**	**Yes**	**Autistic disorder**	**118 patients vs. 126 controls**	**Bottini *et al.* (2001)**		
ADA	20q13	No	Autistic disorder	91 families, 44 trios, TDT and 91 patients vs. 152 controls	Persico *et al.* (2000b)		
MAO A	Xp11	No	Autistic disorder	38 families, TDT	Philippe *et al.* (2002)		
MAO B	Xp11	No	Autistic disorder	38 families, TDT	Philippe *et al.* (2002)		
GRPR	Xp22	No	Rett syndrome	case (*n* = 25) vs. control (*n* = 100)	Heidary *et al.* (1998)		
HOPA	Xq13	No	Autistic disorder	155 patients vs. 157 controls	Beyer *et al.* (2002)		
DXS287	**Xq23**	**Yes**	**Infantile autism**	**case control**	**Petit *et al.* (1996)**		
FMR-1	Xq27	No	Autistic disorder	123 families	Klauck *et al.* (1997)		

Table summarizes current positive and negative association studies for specific genes and autism disorder or related phenotypes. Positive allelic associations are shown in bold type. Also shown are any whole genome linkage peaks that overlap with a gene tested for association, and their linkage scores.

*MLS = Multipoint LOD score; †TDT = Transmission Disequilibrium Test; ‡LD = Linkage Disequilibrium; ¶PDT = Pedigree Disequilibrium Test; §MTDT = Multiallelic TDT; **DQ = Development Quotient

an effort to identify likely candidates for ASD. Screening was performed with the program PATHWAYASSIST (version 1.1, Stratagene Corp, La Jolla, CA) and the FatiGO website (http://fatigo.bioinfo.cnio.es/).

PATHWAYASSIST and ResNet database

The PATHWAYASSIST software (Ariadne Genomics, Rockville, MD) allows the user to explore gene interaction networks represented in the ResNet (tm) database. ResNet (tm) is a comprehensive database of molecular networks compiled by proprietary natural language processing techniques applied to the whole PubMed database. The database contains more than 100 000 events of regulation, interaction and modification between 15 000 proteins, cell processes and small molecules. The architecture of ResNet and PATHWAYASSIST has been described (http://www.ariadnegenomics.com). PATHWAYASSIST provides a 'front end' that allows the user to query the database, and to direct the construction of specific networks relative to genes of interest.

The complete list of all 383 positional candidate genes was loaded into PATHWAYASSIST. Of those genes, 203 were recognized by the software, and were thus subjected to subsequent analysis. The 'Expand Pathway' feature of PATHWAYASSIST was used to build a network of connections starting with these 203 genes and including all available categories of interaction. This expanded list was then searched to find genes that interacted with neural-related positional candidate genes in the following manner. The genes in the expanded set that had interesting GO terms were identified, and then their interacting 'neighbors' were selected using the 'Select Neighbors' command. Set operations were used to reduce the list to only those genes that were among the original list of 203 positional candidate genes. Nine genes not found in the manual search described above were identified in this manner for further evaluation. Of these, four appeared to be logical candidates, and to have been correctly identified by PATHWAYASSIST as having valid interactions (Method 4, in Table 3) after manual inspection.

Table 2: Summary of genomewide linkage studies for autism

Top regions	Peak position*	Physical location†	LOD score‡	References	*n* (families)
1p13	149 cM	113 Mb	2.15	Risch *et al.* (1999)	90
1q23	164 cM	154 Mb	2.63	Auranen *et al.* (2002)	38
2p12	96 cM	76 Mb	1.60	IMGSAC (2001a)	83 + 69¶
2q31	181 cM	175 Mb	3.74	IMGSAC (2001a)	83 + 69¶
2q31	186 cM	183 Mb	2.39–3.32 (Z)	Buxbaum *et al.* (2001), PSD§	49
3p25	36 cM	11 Mb	1.51	Shao *et al.* (2002)	99
3q26	191 cM	180 Mb	4.81	Auranen (2002)	38
4p16	4.6 cM	3.5 Mb	1.55	IMGSAC (1998)	99
4q21	**94 cM**	**85 Mb**	**1.72**	**Yonan *et al.* (in press)**	**345**
5p13	**58 cM**	**40 Mb**	**2.54**	**Yonan *et al.* (in press)**	**345**
6q13	83 cM	70 Mb	2.23	Philippe *et al.* (1999)	51
7q21	104 cM	91 Mb	2.20	CLSA (1999)	75
7q22	112 cM	100 Mb	3.20	IMGSAC (2001a)	83 + 69¶
7q32	142 cM	128 Mb	2.55–3.55	IMGSAC (1998)	99
7q36	170 cM	153 Mb	3.66	Auranen (2002)	38
8q24	**132 cM**	**125 Mb**	**1.50**	**Yonan *et al.* (in press)**	**345**
11p13	**46 cM**	**34 Mb**	**2.24**	**Yonan *et al.* (in press)**	**345**
13q12	21 cM	30 Mb	2.30	CLSA (1999)	75
13q22	55 cM	73 Mb	3.40	CLSA (1999)	75
16p13	19 cM	10 Mb	1.51–1.97	IMGSAC (1998)	99
16p13	25 cM	12 Mb	2.93	IMGSAC (2001a)	83 + 69¶
17q11	50 cM	28 Mb	2.34	IMGSAC (2001a)	83 + 69¶
17q11	**52 cM**	**29 Mb**	**2.83**	**Yonan *et al.* (in press)**	**345**
Xq21	63 cM	94 Mb	2.54	Shao (2002)	99

Table summarizes genomewide linkage studies for autism or ASD, organized by chromosomal position and showing sample size used. Only the linkage regions with an MLS > 1.4 are shown for consistency of comparison. Linkage regions from Yonan *et al.* (in press), that the current study is based upon, are shown in bold. Liu *et al.* (2001) is not shown since the complete sample (110 families) is included and reanalyzed in Yonan *et al.* (in press).

*Peak position = position of the highest point/marker in Kosambi centimorgans from pter = 0.

†Physical location = position of the highest point/marker as mapped onto the Human Genome Browser.

‡LOD score = usually MLS score, however, Z demarks an NPL Z score.

¶83 + 69 = 89 families were used in the initial genomewide scan and then 69 families were added to follow up in 13 candidate regions.

§PSD = Phrase Speech Delay

PATHWAYASSIST was also used to search for pathway relationships beginning with the 13 genes that have been reported to be positively associated with autism in at least one previous study (Table 1). The PATHWAYASSIST 'Build Pathway' function was used to search for pathways beginning with these genes. Next, the pathway was expanded to examine the connections to any of the positional candidate genes of the current study. As before, 203 of the positional candidates were recognized by the program and used in this analysis, only a few of which showed connections to this pathway (Method 5 in Table 3). Interactions among the 203 positional candidates were excluded from the analysis, as these interactions were unrelated to our hypothesis.

GENEWAYS pathway prediction system

GENEWAYS is a program that uses a natural language processing algorithm to extract relationships between molecules or molecular processes by digesting published research literature and building these relationships into pathways (Rzhetsky *et al*. 2000). Electronic copies of the full text of research articles are downloaded to a local database where biologically important concepts such as names of genes, proteins, processes, small molecules and diseases are extracted from the text (Krauthammer *et al*. 2000) and clarified in relation to the many synonyms and homonyms and other ambiguities that are often applied to an identical term (Hatzivassiloglou *et al*. 2001). An associated program, GENIES is a natural language processing parser (Friedman *et al*. 2001). The output

Table 3: Semi-automated search for candidate genes

Gene name	Full name	Chromosome	Method
ACCN1	Neuronal amiloride-sensitive cation channel 1	17q	1, 2
BLMH	Bleomycin hydrolase	17q	1, 3
CENTA2	Centaurin-alpha 2 protein	17q	1, 2
GIT1	G protein-coupled receptor kinase-interactor 1	17q	1, 2, 6
NF1	Neurofibromin	17q	1, 6
OMG	Oligodendrocyte myelin glycoprotein	17q	1, 3
SLC6A4	Solute carrier family 6 (serontonin transporter)	17q	1, 2, 3, 5, 6
TIAF1	TGFB1-induced antiapoptotic factor 1 isoform 1	17q	2
TNFAIP1	Tumor necrosis factor, alpha-induced protein 1	17q	2
TRAF4	TNF receptor-associated factor 4 isoform 1	17q	2
CARD6	Caspase recruitment domain family, member 6	5p	2
CCL28	Small inducible cytokine A28 precursor	5p	2
GDNF	Glial cell derived neurotrophic factor	5p	1, 2, 5, 6
GHR	Growth hormone receptor	5p	1, 6
IL6ST	Interleukin 6 signal transducer	5p	4
IL7R	Interleukin 7 receptor precursor	5p	1, 2
ITGA2	Integrin alpha 2 precursor	5p	2
LIFR	Leukemia inhibitory factor receptor	5p	4, 6
FYB	FYN binding protein	5p	6
PRLR	Prolactin receptor	5p	1, 2
Nup155	Nucleoporin 155 kDa	5p	1
SLC1A3	Solute carrier family 1, member 3 (glutamate transporter)	5p	1, 2
DAB2	Disabled homolog 2, mitogen-responsive phosphoprotein	5p	5
API5	Apoptosis inhibitor 5	11p	1, 2
CAT	Catalase	11p	1, 2
CHRM4	Cholinergic receptor, muscarinic 4	11p	2
ELF5	E74-like factor 5 (ets domain transcription)	11p	1, 2
MC7	Transcription factor in neuroblasts and developing neurons	11p	1
MDK	Midkine (neurite growth-promoting factor 2)	11p	2
MAPK8IP1	Mitogen-activated protein kinase 8 interacting protein 1	11p	5
CD44	CD44 antigen	11p	6
SLC1A2	Solute carrier family 1, member 2 (glutamate transporter)	11p	1, 2
TRAF6	TNF receptor-associated factor 6	11p	1, 2, 6
ATOH1	Atonal homolog 1	4q	1, 2
BIKE	BMP-2 inducible kinase	4q	1
CDS1	Phosphatidate cytidylyltransferase 1	4q	1
CNOT6L	CCR4-NOT transcription complex, subunit 6-like	4q	1

CXCL1	Chemokine (C-X-C motif) ligand 1	4q	2
EIF4E	Eukaryotic translation initiation factor 4E	4q	4, 5, 6
FGF5	Fibroblast growth factor 5 isoform 1 precursor	4q	2
GRID2	Glutamate receptor, ionotropic, delta 2	4q	1, 2
PTPN13	Protein tyrosine phosphatase, non-receptor type 13	4q	5
HPSE	Heparanase	4q	5
IL8	Interleukin 8 precursor	4q	4
NFKB1	Nuclear factor of kappa light polypeptide gene	4q	1, 2
NK16-1	NK6 transcription factor related, locus 1	4q	1, 2
NUP54	Nucleoporin 54 kDa	4q	1, 2
SHRML	Shroom-related protein	4q	1
SNCA	Alpha-synuclein isoform NACP140	4q	1, 2
SPBP	DNA-binding protein amplifying expression of	4q	1, 2
RAP1GDS1	RAP1, GTP-GDP dissociation stimulator 1	4q	5
PKD2	Polycystic kidney disease 2	4q	6
TACR3	Tachykinin receptor 3	4q	1
UNC5C	Unc-5 homolog C	4q	2
SF2	Otoferlin	8q	1
MTBP	Mdm2, transformed 3T3 cell double minute 2, p53 binding protein	8q	5
TAF2	TBP-associated factor 2	8q	5
ZHX1	Zinc-fingers and homeoboxes 1	8q	1, 2

Table shows all candidate genes within our linkage regions that were found by different search strategies.
Method:
1 = Manual search of biological databases
2 = Gene Ontology (GO) query
3 = Positive association study (Table 1)
4 = PATHWAYASSIST 'neighbors'
5 = PATHWAYASSIST predicted pathway candidates
6 = GENEWAYS predicted pathway candidates

of GENIES is represented with semantic trees. A separate module unwinds these complex output trees into simple binary statements that are saved into the GENEWAYS knowledge base. The GENEWAYS system extracts some percentage of incorrect, redundant or contradictory statements that continue to pose bioinformatic challenges (Krauthammer *et al*. 2002), and currently requires manual curation and annotation. The user can conveniently request information about each interaction and retrieve the complete articles from which the information was extracted.

The pathway built with GENEWAYS was based on two sets of genes. The first consisted of about 20 genes that had been previously identified in the literature as playing a role in autism, either from positive association findings (Table 1), known chromosomal abnormalities or similar methods. The second list was the complete list of 383 positional candidate genes. GENEWAYS was then used to try to identify connections between these two groups of genes and to observe how those potential candidates might interact with each other and with other pathways. Currently, it is only possible to examine the GENEWAYS database by building a pathway out from a single gene, rather than having an exhaustive algorithm systematically identify all possible interactions. GENEWAYS was used to identify and visualize all the meaningful connections from the first list of known autism candidates to any information stored in the database. Several of the identified genes in this pathway were located within our linkage regions. Next, additional positional candidate genes were tested to see if they were connected with the same pathway (Method 6 in Table 3). We added an additional 30 positional candidates that we deemed most likely to contribute to ASD. These were genes that from the manual search made the most logical sense to possibly be involved in ASD phenotypes. Of the 30 genes that we examined, only six had direct connections to other genes in the pathway. Only those 30 candidates were examined using this strategy because our experience with this software suggests that it is important to limit the number of genes examined in order to produce an informative pathway that provides testable connections rather than an exhaustive but unwieldy pathway. Each arrow in Fig.2 represents either a physical or a logical interaction. Logical connections may represent multistep processes that include intermediaries not shown in the diagrams.

Transcription microarray meta-analysis

Whole genome gene expression arrays were used to identify possible functional relationships by searching for genes that are coexpressed with key autism candidate genes and positional candidate genes, based on mRNA expression microarray data. To increase the reliability of coexpression detection,

only patterns of coexpression that were consistent in multiple data sets were used, since a coexpression relationship that is found in two or more independent studies is less likely to be an artifact. Because we did not have access to sufficient quantities of high-quality human brain gene expression data, we analyzed the homologs of our candidate genes in a set of seven independently collected mouse brain gene expression data sets. Of the 383 candidate genes, 170 had known mouse homologs, many of which are curated orthologs, which were then used for further analysis.

Of the seven mouse brain gene expression data sets used for Transcription Microarray Meta-Analysis, five were from unpublished in-house data and two were from published data sets (Sandberg *et al.* 2000; Zhao *et al.* 2001). Except for the dataset of Sandberg, which included data from six brain regions, all samples were from the hippocampus. Zhao *et al.* compared the subfields of the hippocampus. The additional data sets from our group are currently unpublished and consist primarily of test-control studies, with between 8 and 24 microarrays per data set, distributed as biological replicates of each condition. The conditions studied in each of these data sets were as follows: Young vs. old mice (M. Verbitsky, A.L. Yonan, G. Malleret, E.R. Kandel, T.C. Gilliam & P. Pavlidis, submitted); protein kinase C-gamma knockout vs. control mice; mice expressing a dominant negative protein kinase A regulatory subunit (R(AB); Abel *et al.* 1997) vs. control; a separate experiment using R(AB) and control animals to examine the effects of context-cued fear conditioning; and an analysis of mice expressing a dominant-negative inhibitor of CCAAT/enhancer-binding protein-family member transcription factors, compared to controls (Chen *et al.* 2003). Each data set was filtered to remove genes clearly lacking detectable expression, removing 30% of genes with the smallest maximal expression in each data set. Each gene was then analyzed to identify genes it was coexpressed with. For each gene, the Pearson correlation coefficient of all pairs of gene expression profiles in the data set was calculated. A *P*-value was calculated for the Pearson correlation assuming the null distribution follows a t-distribution (Zar 1999). *P*-values for each correlation were Bonferroni corrected, and genes with corrected *P*-values < 0.01 were considered coexpressed with the query gene. We note that this method does not make use of the experimental grouping of the samples (e.g., young vs. old), and thus genes which are coexpressed do not necessarily (indeed, typically do not) have expression patterns that are 'relevant' to the originally defined experimental groups. Pairs of genes that meet the criteria for coexpression were entered in a database. From the seven data sets, for all genes examined by the microarrays ($\sim$10 000), we extracted $\sim$200 000 gene pairs ($< 0.1\%$ of all possible pairs). We then screened this database for pairs involving a positional candidate gene homolog that was identified in at least two of the seven data sets. We also attempted to identify genes that were coexpressed with the 13 genes implicated by positive findings from association studies (Table 1). However, we were unable to identify any genes in our linkage regions that were coexpressed with these genes (data not shown).

Results

Table 1 summarizes results from studies that have sought to detect allelic association between candidate genes and autism or autism-related phenotypes. A total of 13 genes and three markers spanning 10 distinct cytogenetic regions purportedly show positive evidence for allelic association to autism. Of these 10 regions only 17q11 is concordant with the linkage regions identified in Yonan *et al.* in press (Fig. 1).

Table 2 summarizes the results from nine genomewide linkage studies for autism and ASD. Interpretation of genetic linkage to common heritable disorders is fraught with uncertainity and cross-study comparisons are not straightforward (Altmuller *et al.* 2001). All other factors being equal, larger sample studies are less prone to both false positive and false negative errors, thus we focused on the five strongest linkage signals from the large Yonan *et al.* study rather than, for example, choosing the five strongest linkage signals across all nine genomewide scans, or the five regions most supported by independent studies. As shown in Table 2, the Yonan *et al.* study (345 multiplex families) is more than three times the size of other reported genomewide studies. When comparing the results from Yonan *et al.* (in press) with those of other published studies in which evidence for linkage exceeded an MLS > 1.4 ($P < 0.01$; Nyholt 2000), overlap was identified on 17q (IMGSAC 2001a). The five putative ASD linkage regions selected for study are indicated in Fig. 1 (also shown as bold in Table 2).

Semi-automated search for ASD candidate genes

In a first attempt to parse positional candidate genes, we used public and commercial biological databases, together with Gene Ontology formalisms (see *Materials and methods*) to predict a subset of 'neural related' genes of potential relevance to ASD (Table 3). Candidates were selected from the 383 positional candidate genes based upon information gathered by manual search of the public UCSC Human Genome Browser and the proprietary Celera Discovery System together with their related links (Method 1, Table 3). A further search using neural-related GO terms (see *Materials and methods*) identified 11 additional genes (*TIAF1*, *TNFAIP1*, *TRAF4*, *CARD6*, *CCL28*, *ITGA2*, *CHRM4*, *MDK*, *CXCL1*, *FGF5*, *UNC5C*) not already identified by the manual search (Method 2, Table 3). Finally, an additional four candidate genes (*IL6ST*, *LIFR*, *EIF4E*, *IL8*) were identified using the PATHWAYASSIST computational software based upon their predicted network association with neural-related pathway genes (Method 4, Table 3; see *Materials and methods*).

Computational pathway prediction methods

In the present paper, we have attempted to leverage what little information is available about the genes that may contribute to autism in order to identify additional candidate genes for

autism based on the results from our genomewide linkage study. Our hypothesis was that by constructing pathways between the genes already suspected to be involved in autism and our positional candidate genes, we could identify a subset of those positional candidates more likely to be involved in autism.

GENEWAYS' predictions regarding the connections between several of the positional candidate genes and a short list of genes suspected to be involved in autism (including both genes positively associated with autism and biological inferences) are shown in Fig. 2. Interactions among three of the genes positively associated with autism (*GLUR6*, *HRAS1* and *SLC6A4*; shown as circles with red letters) together with connecting pathway genes (blue circles), molecules (red triangles) and processes (yellow rectangle), and 10 positional candidate genes (brown circles) were discovered (Fig. 2; Method 6, Table 3). When using the GENEWAYS program, each connecting line is a 'clickable' link that displays the underlying text that supports the interaction.

Gene networks illustrated in Fig. 3 were developed using a conceptually similar strategy, using PATHWAYASSIST instead of GENEWAYS. The PATHWAYASSIST 'Build Pathway' function found valid connections (as determined by manual inspection) between 2 of the 13 genes that have been positively associated with autism (*GLUR6* and *UBE3A*; Table 1) and a subset of the positional candidate genes. Positional candidates that

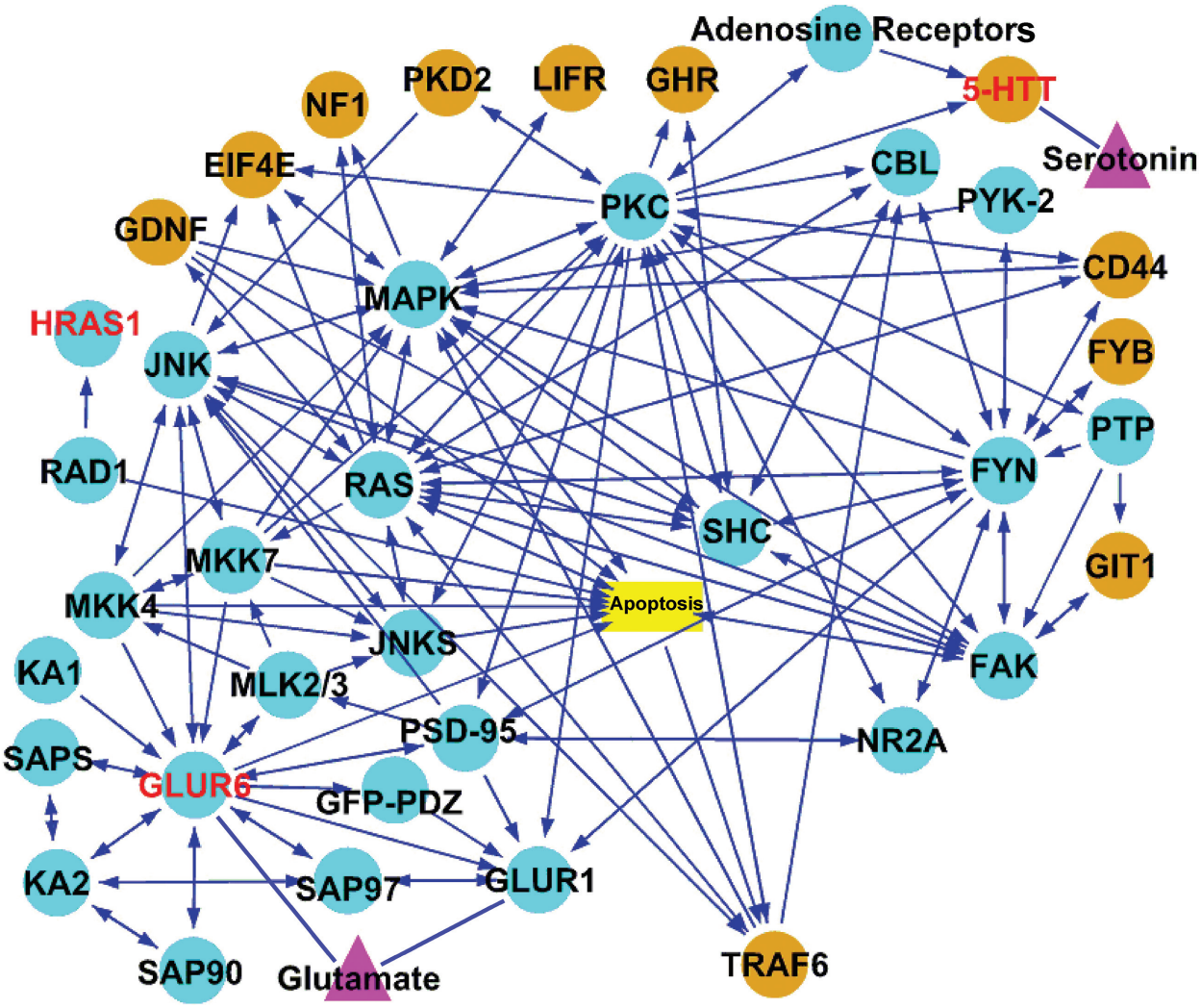

Figure 2: GeneWays pathway showing the interrelationships of several positional candidate genes. This pathway was based on previously identified candidate genes of autism disorder and then built out to show how some of our positional candidate genes may interact. Small molecules are shown as pink triangles, processes are shown as yellow boxes, positional candidate genes found from genomewide linkage study for ASD are shown in brown circles, and all other genes that connect the pathway are shown in turquoise circles. The three gene names that are shown in red text are genes that have been identified as positively associated with autism (Table 1). Note that *5-HTT* (serotonin transporter) is the same gene as *SLC6A4*.

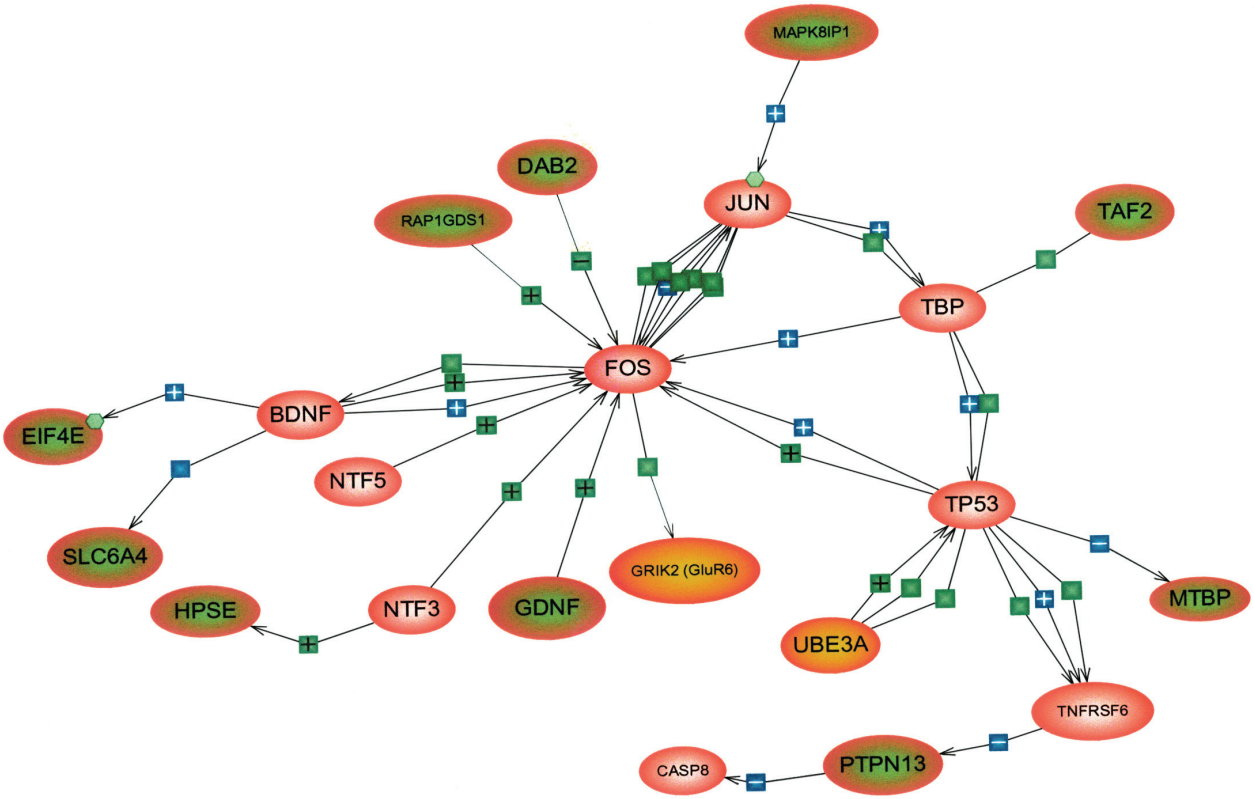

Figure 3: A pathway built using PATHWAYASSIST between genes positively identified in association studies for autism and 203 of the 383 positional candidates. Two of the 13 such positively associated genes (ovals with yellow centers) were found to interact with positional candidate genes (ovals with green centers) via PATHWAYASSIST. The subset of interactions shown here was chosen as being relevant to the pathway originally built out from the positively associated genes.

were found to have valid connections to this pathway are shown as Method 5, Table 3.

Co-expression data, transcription microarray meta-analysis

We analyzed patterns of whole genome gene expression across multiple microarray data sets to identify possible gene regulatory interactions between the selected set of autism candidate genes and a subset of positional candidate genes. Of the 383 candidate genes analyzed, murine homologs for 170 genes were identified, which we then used to query seven independent mouse brain expression data sets. No reliable coexpression patterns were detected among the 13 positively associated autism candidates and the subset of 170 positional candidates. However, 10 of the 170 positional candidates showed highly reliable coexpression with one or more genes that were detected in multiple gene expression data sets (Table 4). A total of 107 genes were coexpressed with the set of 10 query genes. Based on their functions and annotations, we determined that a subset of these 107 genes showed potential relevance to neurodevelopmental disorders (Table 4).

Discussion

In this study we have sought to apply emerging bioinformatic tools to a problem that characterizes nearly all gene-mapping studies that target common, heritable disorders. Common heritable disorders are characteristically multigenic and heterogeneous in nature. Consequently, linkage peaks tend to be broad and weakly significant such that subsequent positional mapping and gene identification is greatly complicated. In a minority of cases, follow-up allelic association analysis has apparently been used successfully to delimit the disease gene region and to identify the disease related genetic variation (Horikawa *et al.* 2000; Ogura *et al.* 2001). The recent sequencing of the human genome, along with the genomes of other well-researched organisms, now makes identification of positionally mapped genes a straightforward bioinformatic exercise. However, knowledge of which genes reside within an interval alone does not significantly change the complexity of gene mapping.

Positional mapping poses unique challenges that are well suited for computational data-mining approaches. Peak linkage findings demarcate chromosomal regions most likely

Table 4: Genes co-expressed with positional candidates based on gene expression data from mouse brain

Index gene	Gene description	Gene accession ID	Linkage region (chromosome)	BP position	Mouse homologue	Number of matches	Co-expressed candidates gene
HNRPDL	heterogeneous nuclear ribonucleoprotein D-like	NM_005463	4	83737143	Mm.195310	17	piccolo (presynaptic cytomatrix protein)* matrin 3
PPP3CA	protein phosphatase 3 (formerly 2B), catalytic	NM_000944	4	102337365	Mm.293	4	Mm.6150 (Highly similar to HAPP_RAT Huntingtin-associated protein-interacting protein)
PKD2	polycystin 2	NM_000297	4	89321599	Mm.6442	2	
PELO	CGI-17 protein	NM_015946	5	52066463	Mm.3241	2	glutamine synthase
NDUFS4	NADH dehydrogenase (ubiquinone) Fe-S protein 4	NM_002495	5	52827009	Mm.14442	1	potassium voltage-gated channel, Shal-related family, member 2
PRLR	prolactin receptor‡	NM_000949	5	35064208	Mm.2752	1	ectonucleotide pyrophosphatase/ phosphodiesterase 2†
ZHX1	zinc-fingers and homeoboxes 1‡	NM_007222	8	123929781	Mm.37216	25	aquaporin 4; quaking; cerebellar postnatal development protein 1
ENPP2	ectonucleotide pyrophosphatase/ phosphodiesterase	NM_006209	8	120238123	Mm.28107	14	prolactin receptor†, ‡; calmodulin-like 4; SLC4A2; TTR
ALDOC	aldolase C, fructose-bisphosphate	NM_005165	17	26752009	Mm.7729	40	Calmodulin; neurochondrin-1; thyroid hormone receptor alpha; protein; procholecystokinin hippocampal amyloid precursor (CCK)
JJAZ1	joined to JAZF1	NM_015355	17	30113956	Mm.21964	1	

Genes that are located within the 1-LOD support interval of our QTL regions (Index Genes) and that belong to classes of coexpressed genes. First the mouse homologue of each index gene was identified (when available). In the absence of appropriate human gene expression data, we utilized 7 independently collected sets of mouse brain gene expression data, consisting of 8–24 microarrays each, to develop classes of coexpressed genes. We identified genes that were reproducibly coexpressed (in two or more of the data sets) with the mouse homologue of the index gene. When an index gene belonged to a functional expression class, the other genes in that class were identified (total # of matches), and the likely candidates from that expression class identified. Candidate genes so identified may be downstream targets of a transcriptional activation pathway common to the index gene and the candidate, with the index gene acting either as a transcription factor (for example, zinc-fingers and homeoboxes 1), or as the modulator of a transcription factor.

* Same gene as PCLO in Table 1 (Nabi et al. 2003).

† These genes are found as both index genes and coexpressed candidates.

‡ Genes also identified in Table 3.

to harbor disease-related genetic variation, yet positional candidate genes pose unique bioinformatic problems: some portion of peak regions will be false positives and harbor no disease related genes, some peaks that do harbor disease related genetic variation will consist of only one disease-related gene among other genes that bear no relationship to the disease, other peaks might obtain their prominence due to the contribution of more than one disease-related gene, and some portion of disease related genes will likely reside outside the identified peak regions.

In addition to positional candidate genes, other types of genetic evidence are typically used to identify common disease causing alleles. Allelic association, or linkage disequilibrium, is used to detect historical association between a candidate gene variant and disease phenotype. Association studies are vulnerable to many of the same genetic complexities that confound genetic linkage studies with the following difference: association studies are robust to locus heterogeneity (since they only test one locus at a time), but confounded by allelic heterogeneity. Association studies are also believed to be quite vulnerable to genotypic differences related to population substructure (background genotypic differences that are unrelated to phenotype) (Hoggart *et al*. 2003). Thus, the 'candidate status' of most candidate genes is subject to uncertainty. Nevertheless, the subset of genes contained within a suggestive linkage peak is likely to be enriched for actual disease genes compared to the genome as a whole.

Both GENEWAYS AND PATHWAYASSIST are pathway prediction methods that are designed to recognize written language and to extract key phrases that describe basic biological relationships between genes, small molecules, cellular processes and similar phenomenon. PATHWAYASSIST reads abstracts, whereas GENEWAYS reads the entire article. Both programs use sophisticated algorithms to predict pathway interactions. Because of problems with interpretation of language, figures and tables, a number of oversights and erroneous conclusions are inevitable in these programs. Thus, human interpretation and curation of these databases and their output remain critically important. Another aspect of natural language processing algorithms is that they must discriminate between physical interactions (binding or cleavage, oxidation, etc.), and logical interactions (e.g., the effect of a drug on gene expression). In the first case, two molecules are known to interact directly, whereas in the latter case, the mechanism of interaction may involve a multistep pathway. Thus, the pathways identified by these algorithms must be carefully filtered and/or checked by an expert user in order to establish the type of experimental data that was collected and to determine what biological experiments are required to further test the proposed pathways.

Identification of candidates by the 'manual' search strategy, the GO strategies and by the PATHWAYASSIST AND GENEWAYS pathway prediction programs all depend upon data from the published literature. In contrast, the transcription microarray meta-analysis is largely based on unpublished experimental data, and thereby provides a completely independent bioinformatic approach to the same positional mapping problem. Since criteria to distinguish correct from incorrect bioinformatic predictions are often lacking, it is desirable to employ independent computational strategies and identify convergent pathway predictions (Eisenberg *et al*. 2000). Yeast whole genome gene expression studies show that coexpressed sets of genes are enriched for functionally related or physically interacting genes (Eisen *et al*. 1998; Ge *et al*. 2001). Genes that are coexpressed may be coregulated by a common transcription factor. Alternatively, one of the genes in a group of transcriptionally coregulated genes may be the transcription factor that drives the expression of the others (e.g., *ZHX1*, Table 4) or it may simply be upstream in a transcriptional cascade of genes that influence the expression of downstream members of the same cascade. Thus, we used this method to identify genes with known function in the brain that may be relevant to ASD, and which are coexpressed with positional candidate genes identified in the genome scan. Such genes might be the downstream targets of transcription factors that reside within the linkage regions and possess functional polymorphisms.

The ability to predict gene expression pathways identified using the method presented in this paper depends heavily on the quality and applicability of the underlying expression data. In the present example, we utilized expression data from mouse brains, rather than human brains, due to availability – an obvious shortcoming. Another limitation is that six of the seven datasets were derived from the hippocampus, rather than the entire brain, or a brain region more relevant to autism (e.g., amygdala). Development of large, well-characterized databases that can store and manage gene expression data and integrate with a range of other heterogeneous data sets, will likely overcome these shortcomings in the near future (e.g., Bader *et al*. 2003). More sophisticated computational programs to predict regulatory motifs (e.g., Bussemaker *et al*. 2001), together with high throughput experimental paradigms for selective and systematic perturbation of well-characterized biological systems (Barstead 2001; Elbashir *et al*. 2001; Ideker *et al*. 2001; McCaffrey *et al*. 2002) will likewise increase the power and scope of this approach. Perhaps the most promising strategies are those that combine the rigor of high throughput experimental paradigms with the speed, power and scope of computational data-mining approaches. Whole genome yeast and *Drosophilia* 'two-hybrid arrays' test every permutation of protein–protein interaction, and despite a high false positive rate, are ideally suited for integrated computational analyses (von Mering *et al*. 2002). Mass spectrometry analysis of purified protein complexes (Ho *et al*. 2002) and the exploration of genetic interactions by identification of synthetic lethal gene combinations in yeast (Tong *et al*. 2001) are both potentially powerful complimentary approaches to the prediction of interacting gene networks and pathways.

It is estimated that upwards of 90% of an individual's liability to develop autism or ASD is determined by genetic factors, yet the disease liability attributable to any single genetic variant may be so small that it is undetectable by current gene mapping strategies. This problem may be addressed to some extent by strategies to predict biological pathways since these strategies may identify interacting sets of genes that together account for a significant portion of heritable disease liability. The role of additive vs. epistatic

gene interactions in the etiology of common heritable disorders is unclear at this point, as is the importance of this distinction in the mapping of such traits and disorders (Carrasquillo *et al*. 2002; Cox *et al*. 1999; Holmans 2002; Tempeton 2000). Computational pathway predictions together with Gene Ontology annotations, gene regulatory information and other molecular interaction data should inform the characterization of additive vs. epistatic gene–gene interactions in ways that complement genetic studies. Thus, it is hoped that computational and bioinformatic approaches will lead to the identification of 'candidate gene networks' that encompass a significant fraction of a given disease's heritable component.

Table 3 summarizes the candidate gene predictions based upon six bioinformatic methods. With the exception of *LIFR* and *EIF4E*, all candidate genes detected by two or more of the automated search strategies were likewise detected by manual searches, suggesting that convergent findings from automated strategies are more reliable. The identified candidate genes are biased in favor of neurobiological disease etiology due to our search strategies. However, the etiology of autism may depend on susceptibility to environmental insults, rather than primary neurological deficits. For example, four candidates with known immunological function (*IL6ST*, *LIFR*, *CD44 and IL8*) were only detected by predicted pathway relationships, reflecting the lack of bias inherent in these pathway prediction approaches (Table 3).

In the present study we identified several genes using multiple bioinformatic approaches. Most notably the serotonin transporter (*SLC6A4* a.k.a. *5-HTT*) was identified by all but one of our search strategies (Table 3) including allele specific association studies (Table 1). Of the 408 microsatellite markers genotyped for linkage to ASD in the study by Yonan *et al*. (in press), the single most significant linkage was detected by a marker that maps less than one megabase distal to *SLC6A4*. It is also noteworthy that *SLC6A4* is located in the only linkage region identified by Yonan *et al*. (in press) that overlaps with the findings of another linkage study (Table 2). Other studies have indicated that autism patients and their unaffected first degree relatives have elevated blood serotonin levels and there is evidence that drugs that selectively target the serotonin transporter can ameliorate some autism related symptoms (Cook & Leventhal 1996; Gingrich & Hen 2001). Thus *SLC6A4* appears to be a particularly promising candidate gene for ASD, although it is not clear that the new data substantially bolster pre-existing data. Both pathway prediction programs predict relationships between glutamate receptor 6 (*GLUR6*; which has been positively associated with autism; Table 1) and positional candidates glial cell derived neurotrophic factor (*GDNF*) and *SLC6A4*, though not obviously via common pathways. Finally, the prolactin receptor (*PRLR*), and zinc-fingers and homeoboxes 1 (*ZHX1*), were identified by the transcriptional pathways prediction method (Table 4) as well as by the manual and GO strategies (Table 3).

Piccolo (*PCLO*) was identified by the transcriptional pathways prediction method as being coexpressed with the positional candidate, heterogeneous nuclear ribonucleoprotein D-like (*HNRPDL*) (Table 4). While *PCLO* itself is not located in our linkage region, and thus is not a positional candidate, this finding raises the possibility that *HNRPDL* may be upstream of *PCLO* in a transcriptional cascade. Based on its function, *PCLO* has been suggested as a possible candidate gene for autism (Fenster & Garner 2002), although this is not substantiated by allelic association (Table 1) (Nabi *et al*. 2003). The coexpression data suggest an alternative possibility: a polymorphism within *HNRPDL* may lead to differential expression of downstream targets that include *PCLO*. Thus the present results suggest that *HNRPDL* might be a viable candidate for autism, a conclusion that would not have been reached by any of the strategies whose results are reflected in Table 3.

The current study was designed to explore emerging bioinformatic technologies for the purpose of parsing large sets of genetically mapped (positional) candidate genes in search of disease related genetic variation. Using a large family study of autism and ASD we show that sophisticated bioinformatics approaches can be applied to this task and that convergent approaches might be used to offset inherent biases in any given approach to ultimately identify a subset of genes that are enriched for disease related genetic variation in the study sample, thus providing testable hypotheses. We further note with optimism that the genesis of integrative databases, powerful whole genome computational data-mining approaches, and high-throughput experimental paradigms to evaluate molecular interactions and pathway associations, bode well for the merger of bioinformatic and gene mapping approaches in the future.

Supplementary material

The following material is available from: http//www.blackwellpublishing.com/products/journals/suppmat/GBB/GBB041/GBB041sm.htm

The supplementary material contains the complete list of all 383 positional candidate genes in the top five regions from Yonan *et al*. (2003). There is a detailed explanation of the positions in cM and Mb for each candidate region, as well as their LOD scores in each region. Also, each chromosomal region is shown separately with only the genes that reside in that region listed for ease of comparison.

References

Abel, T., Nguyen, P.V., Barad, M., Deuel, T.A., Kandel, E.R. & Bourtchouladze, R. (1997) Genetic demonstration of a role for PKA in the late phase of LTP and in hippocampus-based long-term memory. *Cell* **88,** 615–626.

Altmuller, J., Palmer, L.J., Fischer, G., Scherb, H. & Wjst, M. (2001) Genomewide scans of complex human diseases: true linkage is hard to find. *Am J Hum Genet* **69,** 936–950.

Auranen, M., Vanhala, R., Varilo, T., Ayers, K., Kempas, E., Ylisaukko-Oja, T., Sinsheimer, J.S., Peltonen, L. & Jarvela, I. (2002) A genomewide screen for autism-spectrum disorders: evidence for a major susceptibility locus on chromosome 3q25–27. *Am J Hum Genet* **71,** 777–790.

Bader, G.D., Betel, D. & Hogue, C.W. (2003) BIND: the Biomolecular Interaction Network Database. *Nucl Acids Res* **31,** 248–250.

Bailey, A., Le Couteur, A., Gottesman, I., Bolton, P., Simonoff, E., Yuzda, E. & Rutter, M. (1995) Autism as a strongly genetic disorder: evidence from a British twin study. *Psychol Med* **25**, 63–77.

Bard, J. (2003) Ontologies: Formalising biological knowledge for bioinformatics. *Bioessays* **25,** 501–506.

Baron, M. (2002) Manic-depression genes and the new millennium: poised for discovery. *Mol Psychiatry* **7,** 342–358.

Barstead, R. (2001) Genome-wide RNAi. *Current Opinion Chem Biol* **5,** 63–66.

Betancur, C., Corbex, M., Spielewoy, C., Philippe, A., Laplanche, J.L., Launay, J.M., Gillberg, C., Mouren-Simeoni, M.C., Hamon, M., Giros, B., Nosten-Bertrand, M. & Leboyer, M. (2002) Serotonin transporter gene polymorphisms and hyperserotonemia in autistic disorder. *Mol Psychiatry* **7,** 67–71.

Beyer, K.S., Klauck, S.M., Benner, A., Poustka, F. & Poustka, A. (2002) Association studies of the HOPA dodecamer duplication variant in different subtypes of autism. *Am J Med Genet* **114,** 110–115.

Bonora, E., Bacchelli, E., Levy, E.R., Blasi, F., Marlow, A., Monaco, A.P. & Maestrini, E. (2002) International Molecular Genetic Study of Autism Consortium (IMGSAC) Mutation screening and imprinting analysis of four candidate genes for autism in the 7q32 region. *Mol Psychiatry* **7,** 289–301.

Botstein, D. & Risch, N. (2003) Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat Genet* **33,** 228–237.

Bottini, N., De Luca, D., Saccucci, P., Fiumara, A., Elia, M., Porfirio, M.C., Lucarelli, P. & Curatolo, P. (2001) Autism: evidence of association with adenosine deaminase genetic polymorphism. *Neurogenetics* **3,** 111–113.

Bussemaker, H.J., Li, H. & Siggia, E.D. (2001) Regulatory element detection using correlation with expression. *Nat Genet* **27,** 167–171.

Buxbaum, J.D., Silverman, J.M., Smith, C.J., Greenberg, D.A., Kilifarski, M., Reichert, J., Cook, E.H. Jr, Fang, Y., Song, C.Y. & Vitale, R. (2002) Association between a GABRB3 polymorphism and autism. *Mol Psychiatry* **7,** 311–316.

Buxbaum, J.D., Silverman, J.M., Smith, C.J., Kilifarski, M., Reichert, J., Hollander, E., Lawlor, B.A., Fitzgerald, M., Greenberg, D.A. & Davis, K.L. (2001) Evidence for a susceptibility gene for autism on chromosome 2 and for genetic heterogeneity. *Am J Hum Genet* **68,** 1514–1520.

Carrasquillo, M.M., McCallion, A.S., Puffenberger, E.G., Kashuk, C.S., Nouri, N. & Chakravarti, A. (2002) Genome–wide association study and mouse model identify interaction between RET and EDNRB pathways in Hirschsprung disease. *Nat Genet* **32,** 237–244.

Chakrabarti, S. & Fombonne, E. (2001) Pervasive developmental disorders in preschool children. *JAMA* **285,** 3093–3099.

Charman, T. (2002) The prevalence of autism spectrum disorders Recent evidence and future challenges. *Eur Child Adolesc Psychiatry* **11,** 249–256.

Chen, A., Muzzio, I.A., Malleret, G., Bartsch, D., Verbitsky, M., Pavlidis, P., Yonan, A.L., Vronskaya, S., Grody, M.B., Cepeda, I., Gilliam, T.C. & Kandel, E.R. (2003) Inducible Enhancement of Memory Storage and Synaptic Plasticity in Transgenic Mice Expressing an Inhibitor of ATF4 (CREB-2) and C/EBP Proteins. *Neuron* **39,** 655–669.

CLSA (Collaborative Linkage Study of Autism) (1999) An autosomal genomic screen for autism. *Am J Med Genet* **88,** 609–615.

Cook, E.H. & Leventhal, B.L. (1996) The serotonin system in autism. *Curr Opin Pediatr* **8,** 348–354.

Cox, N.J., Frigge, M., Nicolae, D.L., Concannon, P., Hanis, C.L., Bell, G.I. & Kong, A. (1999) Loci on chromosomes 2 (NIDDM1) and 15 interact to increase susceptibility to diabetes in Mexican Americans. *Nat Genet* **21,** 213–215.

Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* **95,** 14863–14868.

Eisenberg, D., Marcotte, E.M., Xenarios, I. & Yeates, T.O. (2000) Protein function in the post-genomic era. *Nature* **405,** 823–826.

Elbashir, S.M., Harborth, J., Lendeckel, W., Yalcin, A., Weber, K. & Tuschl, T. (2001) Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature* **411,** 494–498.

Fenster, S.D. & Garner, C.C. (2002) Gene structure and genetic localization of the PCLO gene encoding the presynaptic active zone protein Piccolo. *Int J Dev Neurosci* **20,** 161–171.

Folstein, S. & Rutter, M. (1977b) Infantile autism: a genetic study of 21 twin pairs. *J Child Psychol Psychiatry* **18,** 297–321.

Fombonne, E. (2003) The prevalence of autism. *JAMA* **289,** 87–89.

Friedman, C., Kra, P., Yu, H., Krauthammer, M. & Rzhetsky, A. (2001) GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics* **17,** S74–S82.

Ge, H., Liu, Z., Church, G.M. & Vidal, M. (2001) Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. *Nat Genet* **29,** 482–486.

Gillberg, C. & Wing, L. (1999) Autism: not an extremely rare disorder. *Acta Psychiatr Scand* **99,** 399–406.

Gingrich, J.A. & Hen, R. (2001) Dissecting the role of the serotonin system in neuropsychiatric disorders using knockout mice. *Psychopharmacology* **155,** 1–10.

Hatzivassiloglou, V., Duboue, P.A. & Rzhetsky, A. (2001) Disambiguating proteins, genes, and RNA in text: a machine learning approach. *Bioinformatics* **17,** S97–S106.

Heidary, G., Hampton, L.L., Schanen, N.C., Rivkin, M.J., Darras, B.T., Battey, J. & Francke, U. (1998) Exclusion of the gastrin-releasing peptide receptor (GRPR) locus as a candidate gene for Rett syndrome. *Am J Med Genet* **78,** 173–175.

Herault, J., Petit, E., Martineau, J., Perrot, A., Lenoir, P., Cherpi, C., Barthelemy, C., Sauvage, D., Mallet, J. & Muh, J.P. (1995) Autism and genetics. clinical approach and association study with two markers of HRAS gene. *Am J Med Genet* **60,** 276–281.

Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L. & Adams, S.L., *et al.* (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature **415,** 180–183.

Hoggart, C.J., Parra, E.J., Shriver, M.D., Bonilla, C., Kittles, R.A., Clayton, D.G. & McKeigue, P.M. (2003) Control of confounding of genetic associations in stratified populations. *Am J Hum Genet* **72,** 1492–1504.

Holmans, P. (2002) Detecting gene–gene interactions using affected sib pair analysis with covariates. *Hum Hered* **53,** 92–102.

Horikawa, Y., Oda, N., Cox, N.J., Li, X., Orho-Melander, M. & Hara, M., *et al.* (2000) Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus. Nat Genet **26,** 163–175.

Hyttinen, V., Kaprio, J., Kinnunen, L., Koskenvuo, M., Tuomilehto, J. (2003) Genetic liability of type 1 diabetes and the onset age among 22,650 young Finnish twin pairs: a nationwide follow-up study. *Diabetes* **52,** 1052–1055.

Ideker, T., Thorsson, V., Ranish, J.A., Christmas, R., Buhler, J., Eng, J.K., Bumgarner, R., Goodlett, D.R., Aebersold, R. & Hood, L. (2001) Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* **292,** 929–934.

IMGSAC (International Molecular Genetic Study of Autism Consortium) (1998) A full genome screen for autism with evidence for linkage to a region on chromosome 7q. *Hum Mol Genet* **7,** 571–578.

IMGSAC (International Molecular Genetic Study of Autism Consortium) (2001a) A genomewide screen for autism: strong evidence for linkage to chromosomes 2q, 7q, and 16p. *Am J Hum Genet* **69,** 570–581.

IMGSAC (International Molecular Genetic Study of Autism Consortium) (2001b) Further characterization of the autism susceptibility locus AUTS1 on chromosome 7q. *Hum Mol Genet* **10,** 973–982.

Jamain, S., Betancur, C., Quach, H., Philippe, A., Fellous, M., Giros, B., Gillberg, C., Leboyer, M. & Bourgeron, T.: Paris Autism Research International Sibpair (PARIS) Study (2002) Linkage and association of the glutamate receptor 6 gene with autism. *Mol Psychiatry* **7,** 302–310.

Ju, W., Wang, J., Li, B. & Li, Z. (2000) An epidemiology and molecular genetic study on breast cancer susceptibility. *Chin Med Sci J* **15,** 231–237.

Kim, S.J., Cox, N., Courchesne, R., Lord, C., Corsello, C., Akshoomoff, N., Guter, S., Leventhal, B.L., Courchesne, E. & Cook, E.H. Jr (2002a) Transmission disequilibrium mapping at the serotonin transporter gene (SLC6A4) region in autistic disorder. *Mol Psychiatry* **7,** 278–288.

Kim, S.J., Herzing, L.B., Veenstra-VanderWeele, J., Lord, C., Courchesne, R., Leventhal, B.L., Ledbetter, D.H., Courchesne, E. & Cook, E.H. Jr (2002b) Mutation screening and transmission disequilibrium study of ATP10C in autism. *Am J Med Genet* **114,** 137–143.

Kim, S.J., Young, L.J., Gonen, D., Veenstra-VanderWeele, J., Courchesne, R., Courchesne, E., Lord, C., Leventhal, B.L., Cook, E.H. Jr & Insel, T.R. (2002c) Transmission disequilibrium testing of arginine vasopressin receptor 1A (AVPR1A) polymorphisms in autism. *Mol Psychiatry* **7,** 503–507.

Klauck, S.M., Munstermann, E., Bieber-Martig, B., Ruhl, D., Lisch, S., Schmotzer, G., Poustka, A. & Poustka, F. (1997) Molecular genetic analysis of the FMR-1 gene in a large collection of autistic patients. *Hum Genet* **100,** 224–229.

Krauthammer, M., Kra, P., Iossifov, I., Gomez, S.M., Hripcsak, G., Hatzivassiloglou, V., Friedman, C. & Rzhetsky, A. (2002) Of truth and pathways: chasing bits of information through myriads of articles. *Bioinformatics* **18,** S249–S257.

Krauthammer, M., Rzhetsky, A., Morozov, P. & Friedman, C. (2000) Using BLAST for identifying gene and protein names in journal articles. *Gene* **259,** 245–252.

Krebs, M.O., Betancur, C., Leroy, S., Bourdel, M.C. & Gillberg, C., Leboyer, M.: Paris Autism Research International Sibpair (PARIS) Study (2002) Absence of association between a polymorphic GGC repeat in the 5′ untranslated region of the reelin gene and autism. *Mol Psychiatry* **7,** 801–804.

Lamb, J.A., Moore, J., Bailey, A. & Monaco, A.P. (2000) Autism: recent molecular genetic advances. *Hum Mol Genet* **9,** 861–868.

Lander, E. & Kruglyak, L. (1995) Genetic dissection of complex traits. guidelines for interpreting and reporting linkage results. *Nat Genet* **11,** 241–247.

Lander, E.S. & Schork, N.J. (1994) Genetic dissection of complex traits. *Science* **265,** 2037–2048.

Lauritsen, M. & Ewald, H. (2001) The genetics of autism. *Acta Psychiatr Scand* **103,** 411–427.

Liu, J., Nyholt, D.R., Magnussen, P., Parano, E., Pavone, P., Geschwind, D., Lord, C., Iversen, P., Hoh, J., Ott, J. & Gilliam, T.C.: The Autism Genetic Resource Exchange Consortium (2001) A genomewide screen for autism susceptibility loci. *Am J Hum Genet* **69,** 327–340.

McCaffrey, A.P., Meuse, L., Pham, T.T., Conklin, D.S., Hannon, G.J. & Kay, M.A. (2002) RNA interference in adult mice. *Nature* **418,** 38–39.

McCoy, P.A., Shao, Y., Wolpert, C.M., Donnelly, S.L., Ashley-Koch, A., Abel, H.L., Ravan, S.A., Abramson, R.K., Wright, H.H., DeLong, G.R., Cuccaro, M.L., Gilbert, J.R. & Pericak-Vance, M.A. (2002) No association between the WNT2 gene and autistic disorder. *Am J Med Genet* **114,** 106–109.

Menold, M.M., Shao, Y., Wolpert, C.M., Donnelly, S.L., Raiford, K.L., Martin, E.R., Ravan, S.A., Abramson, R.K., Wright, H.H., Delong, G.R., Cuccaro, M.L. & Pericak-Vance, M.A., Gilbert, J.R. (2001) Association analysis of chromosome 15 gabaa receptor subunit genes in autistic disorder. *J Neurogenet* **15,** 245–259.

von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S. & Bork, P. (2002) Comparative assessment of large-scale data sets of protein–protein interactions. *Nature* **417,** 399–403.

Nabi, R., Zhong, H., Serajee, F.J. & Huq, A.H. (2003) No association between single nucleotide polymorphisms in DLX6 and Piccolo genes at 7q21-q22 and autism. *Am J Med Genet* **119B,** 98–101.

Nurmi, E.L., Bradford, Y., Chen, Y., Hall, J., Arnone, B., Gardiner, M.B., Hutcheson, H.B., Gilbert, J.R., Pericak-Vance, M.A., Copeland-Yates, S.A., Michaelis, R.C., Wassink, T.H., Santangelo, S.L., Sheffield, V.C., Piven, J., Folstein, S.E. & Haines, J.L., Sutcliffe, J.S. (2001) Linkage disequilibrium at the Angelman syndrome gene UBE3A in autism families. *Genomics* **77,** 105–113.

Nyholt, D.R. (2000) All LODs are not created equal. *Am J Hum Genet* **67,** 282–288.

O'Brien, E.K., Zhang, X., Nishimura, C., Tomblin, J.B. & Murray, J.C. (2003) Association of Specific Language Impairment (SLI) to the region of 7q31. *Am J Hum Genet* **72,** 1536–1543.

Ogura, Y., Bonen, D.K., Inohara, N., Nicolae, D.L., Chen, F.F., Ramos, R., Britton, H., Moran, T., Karaliuskas, R., Duerr, R.H., Achkar, J.P., Brant, S.R., Bayless, T.M., Kirschner, B.S., Hanauer, S.B., Nunez, G. & Cho, J.H. (2001) A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. *Nature* **411,** 603–606.

Persico, A.M., Militerni, R., Bravaccio, C., Schneider, C., Melmed, R., Conciatori, M., Damiani, V., Baldi, A. & Keller, F. (2000a) Lack of association between serotonin transporter gene promoter variants and autistic disorder in two ethnically distinct samples. *Am J Med Genet* **96,** 123–127.

Persico, A.M., Militerni, R., Bravaccio, C., Schneider, C., Melmed, R., Trillo, S., Montecchi, F., Palermo, M.T., Pascucci, T., Puglisi-Allegra, S., Reichelt, K.L., Conciatori, M., Baldi, A. & Keller, F. (2000b) Adenosine deaminase alleles and autistic

disorder. case-control and family–based association studies. *Am J Med Genet* **96,** 784–790.

Persico, A.M., Militerni, R., Bravaccio, C., Schneider, C., Melmed, R., Trillo, S., Montecchi, F., Palermo, M., Pascucci, T., Puglisi-Allegra, S., Reichelt, K.L., Conciatori, M. & Keller, F. (2001) No association between the 4g/5G polymorphism of the plasminogen activator inhibitor-1 gene promoter and autistic disorder. *Psychiatr Genet* **11,** 99–103.

Persico, A.M., Pascucci, T., Puglisi-Allegra, S., Militerni, R., Bravaccio, C., Schneider, C., Melmed, R., Trillo, S., Montecchi, F., Palermo, M., Rabinowitz, D., Reichelt, K.L., Conciatori, M., Marino, R. & Keller, F. (2002) Serotonin transporter gene promoter variants do not explain the hyperserotoninemia in autistic children. *Mol Psychiatry* **7,** 795–800.

Petit, E., Herault, J., Raynaud, M., Cherpi, C., Perrot, A., Barthelemy, C., Lelord, G. & Muh, J.P. (1996) X chromosome and infantile autism. *Biol Psychiatry* **40,** 457–464.

Philippe, A., Guilloud-Bataille, M., Martinez, M., Gillberg, C., Rastam, M., Sponheim, E., Coleman, M., Zappella, M., Aschauer, H., Penet, C., Feingold, J., Brice, A. & Leboyer, M.: Paris Autism Research International Sibpair Study (2002) Analysis of ten candidate genes in autism by association and linkage. *Am J Med Genet* **114,** 125–128.

Philippe, A., Martinez, M., Guilloud-Bataille, M., Gillberg, C., Rastam, M., Sponheim, E., Coleman, M., Zappella, M., Aschauer, H., Van Maldergem, L., Penet, C., Feingold, J., Brice, A., Leboyer, M. & van Malldergerme, L. (1999) Genome-wide scan for autism susceptibility genes. Paris Autism Res Int Sibpair Study. *Hum Mol Genet* **8,** 805–812.

Plank, S.M., Copeland-Yates, S.A., Sossey-Alaoui, K., Bell, J.M., Schroer, R.J., Skinner, C. & Michaelis, R.C. (2001) Lack of association of the (AAAT) 6 allele of the GXAlu tetranucleotide repeat in intron 27b of the NF1 gene with autism. *Am J Med Genet* **105,** 404–405.

Prior, M. (2003) Is there an increase in the prevalence of autism spectrum disorders? *J Paediatr Child Health* **39,** 81–82.

Riordan, J.R., Rommens, J.M., Kerem, B., Alon, N., Rozmahel, R., Grzelczak, Z., Zielenski, J., Lok, S., Plavsic, N., Chou, J.L, Drumm, M.L., Iannuzzi, M.C., Collins, F.S. & Tsui, L.C. (1989) Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science* **245,** 1066–1073.

Risch, N., Spiker, D., Lotspeich, L., Nouri, N., Hinds, D. & Hallmayer, J, *et al.* (1999) A genomic screen of autism: evidence for a multilocus etiology. *Am J Hum Genet* **65,** 493–507.

Rommens, J.M., Iannuzzi, M.C., Kerem, B., Drumm, M.L., Melmer, G., Dean, M., Rozmahel, R., Cole, J.L., Kennedy, D., Riordan, J.R., Tsui, L.C. & Collins, F.S. (1989) Identification of the cystic fibrosis gene: chromosome walking and jumping. *Science* **245,** 1059–1065.

Rutter, M. (2000) Genetic studies of autism: from the 1970s into the millennium. *J Abnorm Child Psychol* **28,** 3–14.

Rzhetsky, A., Koike, T., Kalachikov, S., Gomez, S.M., Krauthammer, M., Kaplan, S.H., Kra, P., Russo, J.J. & Friedman, C. (2000) A knowledge model for analysis and simulation of regulatory networks. *Bioinformatics* **16,** 1120–1128.

Sandberg, R., Yasuda, R., Pankratz, D.G., Carter, T.A., Del Rio, J.A., Wodicka, L., Mayford, M., Lockhart, D.J. & Barlow, C. (2000) Regional and strain-specific gene expression mapping in the adult mouse brain. *Proc Natl Acad Sci USA* **97,** 11038–11043.

Serajee, F.J., Zhong, H., Nabi, R. & Huq, A.H. (2003) The metabotropic glutamate receptor 8 gene at 7q31: partial duplication and possible association with autism. *J Medical Genet* **40**, e42.

Shao, Y., Wolpert, C.M., Raiford, K.L., Menold, M.M., Donnelly, S.L., Ravan, S.A., Bass, M.P., McClain, C., von Wendt, L., Vance, J.M., Abramson, R.H., Wright, H.H., Ashley-Koch, A., Gilbert, J.R., DeLong, R.G., Cuccaro, M.L. & Pericak-Vance, M.A. (2002) Genomic screen and follow-up analysis for autistic disorder. *Am J Med Genet* **114,** 99–105.

Smalley, S.L. (1997) Genetic influences in childhood-onset psychiatric disorders: autism and attention-deficit/hyperactivity disorder. *Am J Hum Genet* **60,** 1276–1282.

Smalley, S.L., Asarnow, R.F. & Spence, M.A. (1988) Autism and genetics. A decade of research. *Arch General Psychiatry* **45,** 953–961.

Stubbs, E.G. & Magenis, R.E. (1980) HLA and autism. *J Autism Dev Disord* **10,** 15–19.

Szatmari, P., Jones, M.B., Zwaigenbaum, L. & MacLean, J.E. (1998) Genetics of autism: overview and new directions. *J Autism Dev Disord* **28,** 351–368.

Talebizadeh, Z., Bittel, D.C., Miles, J.H., Takahashi, N., Wang, C.H., Kibiryeva, N. & Butler, M.G. (2002) No association between HOXA1 and HOXB1 genes and autism spectrum disorders (ASD). *J Medical Genet* **39**, e70.

Tempeton, A.R. (2000) Epistasis and complex traits. In Wolf, J.B., Brodie, E.D., III & Wade, M.J. (eds), *Epistasis and the Evolutionary Process*. Oxford University Press, Oxford, pp. 41–57.

Tong, A.H., Evangelista, M., Parsons, A.B., Xu, H., Bader, G.D., Page, N., Robinson, M., Raghibizadeh, S., Hogue, C.W., Bussey, H., Andrews, B., Tyers, M. & Boone, C. (2001) Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* **294,** 2364–2368.

Vourc'h, P., Martin, I., Marouillat, S., Adrien, J.L., Barthelemy, C., Moraine, C., Muh, J.P. & Andres, C. (2003) Molecular analysis of the oligodendrocyte myelin glycoprotein gene in autistic disorder. *Neurosci Lett* **338,** 115–118.

Warren, R.P., Odell, J.D., Warren, W.L., Burger, R.A., Maciulis, A., Daniels, W.W. & Torres, A.R. (1996) Strong association of the third hypervariable region of HLA-DR beta 1 with autism. *J Neuroimmunol* **67,** 97–102.

Wassink, T.H., Piven, J., Vieland, V.J., Huang, J., Swiderski, R.E., Pietila, J., Braun, T., Beck, G., Folstein, S.E., Haines, J.L. & Sheffield, V.C. (2001) Evidence supporting WNT2 as an autism susceptibility gene. *Am J Med Genet* **105,** 406–413.

Weiss, K.M. & Terwilliger, J.D. (2000) How many diseases does it take to map a gene with SNPs? *Nat Genet* **26,** 151–157.

Yeargin-Allsopp, M., Rice, C., Karapurkar, T., Doernberg, N., Boyle, C. & Murphy, C. (2003) Prevalence of autism in a US metropolitan area. *JAMA* **289,** 49–55.

Yonan, A.L., Alarcón, M., Cheng, R., Magnusson, P.K., Spence, S.J., Palmer, A.A., Grunn, A., Hark Juo, S.H., Terwilliger, J.D., Liu, J., Cantor, R.M., Geschwind, D.H. & Gilliam, T.C. (2003) A genomewide screen of 345 families for autism susceptibility loci. *Am J Hum Genet* **73,** 886–897.

Zar, J.H. (1999) *Biostatistical Analysis*. Prentice Hall, Upper Saddle River, NJ.

Zhang, H., Liu, X., Zhang, C., Mundo, E., Macciardi, F., Grayson, D.R., Guidotti, A.R. & Holden, J.J. (2002) Reelin gene alleles and susceptibility to autism spectrum disorders. *Mol Psychiatry* **7,** 1012–1017.

Zhao, X., Lein, E.S., He, A., Smith, S.C., Aston, C. & Gage, F.H. (2001) Transcriptional profiling reveals strict boundaries between hippocampal subregions. *J Comp Neurol* **441,** 187–196.

Zhong, H., Serajee, F.J., Nabi, R. & Huq, A.H. (2003) No association between the EN2 gene and autistic disorder. *J Med Genet* **40**, e4.

Acknowledgments

We gratefully acknowledge the Autism Genome Resource Exchange (AGRE) families who made this study possible and the Cure Autism Now Foundation, which founded and continues to support AGRE. This research was funded by MH64547 (TCG) and a generous donation from Judith P. Sulzberger, MD. We are grateful to the AMDeC Bioinformatics Core Facility for assistance with all bioinformatic studies. Finally, we would like to thank Adina Grunn for technical assistance in determining the genotypes that lead to this analysis.