D1016–D1022 *Nucleic Acids Research, 2012, Vol. 40, Database issue* *Published online 1 December 2011*
doi:10.1093/nar/gkr1145

AutismKB: an evidence-based knowledgebase of autism genetics

Li-Ming Xu, Jia-Rui Li, Yue Huang, Min Zhao, Xing Tang and Liping Wei*

Center for Bioinformatics, State Key Laboratory of Protein and Plant Gene Research, College of Life Sciences, Peking University, Beijing, 100871, P.R. China

Received August 14, 2011; Revised October 4, 2011; Accepted November 9, 2011

ABSTRACT

Autism spectrum disorder (ASD) is a heterogeneous neurodevelopmental disorder with a prevalence of 0.9–2.6%. Twin studies showed a heritability of 38–90%, indicating strong genetic contributions. Yet it is unclear how many genes have been associated with ASD and how strong the evidence is. A comprehensive review and analysis of literature and data may bring a clearer big picture of autism genetics. We show that as many as 2193 genes, 2806 SNPs/VNTRs, 4544 copy number variations (CNVs) and 158 linkage regions have been associated with ASD by GWAS, genome-wide CNV studies, linkage analyses, low-scale genetic association studies, expression profiling and other low-scale experimental studies. To evaluate the evidence, we collected metadata about each study including clinical and demographic features, experimental design and statistical significance, and used a scoring and ranking approach to select a core data set of 434 high-confidence genes. The genes mapped to pathways including neuroactive ligand–receptor interaction, synapse transmission and axon guidance. To better understand the genes we parsed over 30 databases to retrieve extensive data about expression patterns, protein interactions, animal models and pharmacogenetics. We constructed a MySQL-based online database and share it with the broader autism research community at http://autismkb.cbi.pku.edu.cn, supporting sophisticated browsing and searching functionalities.

INTRODUCTION

Autism spectrum disorder (ASD) is a heterogeneous neurodevelopmental disorder characterized by impairments in reciprocal social interaction and communication and presence of restricted, repetitive and stereotyped patterns of behavior, interests and activities (1). ASD is an umbrella term for Autistic Disorder, Asperger Syndrome and Pervasive Developmental Disorder Not Otherwise Specified (PDD-NOS) (1). With an early onset prior to age 3 and a prevalence as high as 0.9–2.6% (2,3), ASD is one of the leading causes of childhood disability and inflicts serious suffering and burden for the family and society (4).

Understanding the causes of ASD is critical for developing better treatment. Twin studies have shown that the heritability of ASD is as high as 38–90%, indicating strong contributions by genetic factors as well as environmental factors (5,6). The search for environmental factors has not yet led to convincing major candidates whereas the search for genes associated with autism, although far from complete or conclusive, has been more fruitful. The genes discovered so far can be roughly grouped into two categories: 'syndromic autism related genes' or causal genes underlying genetic disorders that cause autistic symptoms such as Fragile X Syndrome, Rett Syndrome, Tuberous Sclerosis Complex and dozens of other disorders (7,8), and 'non-syndromic autism related genes' most of which are susceptibility genes (9). Many experimental methods have been used to identify associated genes, including the earlier linkage analyses and low-scale candidate gene association or experimental studies as well as the more recent genome-wide association studies (GWAS), genome-wide CNV studies and expression profiling.

With hundreds of studies published, especially the recent genome-wide studies, and with the next-generation sequencing technologies providing even more power for further gene discoveries (10), a new challenge has emerged: it has become more and more difficult for an autism researcher to answer with confidence how many genes have been associated with ASD, how strong the evidence is, what features the genes have and what pathways they involve. The amount of available literature and data and the intrinsic complexity of autism genetics demand bioinformatic data management and analysis.

*To whom correspondence should be addressed. Tel/Fax: +86 10 6275 5206; Email: weilp@mail.cbi.pku.edu.cn

Three efforts have been made so far by different groups to collect genes and variations associated with ASD: AutDB (also known as SAFRI Gene) collected 219 genes (11,12), Autism genetic database (AGD) collected 226 genes and 743 CNVs (13) and Autism Chromosome Rearrangement Database (ACRD) collected 372 breakpoints and other genomic features (14). However, they are far from a comprehensive survey of autism genetics. To bring a clearer big picture of autism genetics, we performed a comprehensive review and analysis of published literature and data, described below, resulting in a total of 2193 genes, 2806 SNPs/VNTRs, 4544 CNVs and 158 linkage regions. We provide the results as an online resource for the broader autism research community at http://autismkb.cbi.pku.edu.cn/ with extensive evidence and annotations, supporting sophisticated browsing and searching functionalities.

DATA COLLECTION

Literature search

We searched the PubMed database for publications related to autism genetics, using the query term 'autism AND associat*' for association studies, 'autism AND (gene OR microarray OR proteomics)' for expression profiling studies and the other low-scale experimental studies, and 'autism AND (CNV OR copy number variation OR microarray* OR microdel* OR microdup* OR rearrange* OR (genome-wide AND (linkage OR associa* OR scan)))' for CNV and linkage studies. The abstracts of the 4000+ articles retrieved were reviewed to remove irrelevant papers, resulting in a final set of 579 articles, reporting a total of 11 GWAS, 242 low-scale candidate gene association studies, 13 expression profiling studies, 95 genome-wide CNV studies, 23 genome-wide linkage analyses and 236 other low-scale experimental studies.

For syndromic autism-related genes, we first collected the autism-related disorders and their causal genes from a recently published comprehensive review (7). We then searched OMIM to get the official disease names and linked all the disorders to OMIM, and searched PubMed for additional citations using the query '(OMIM disease name) AND autism' for each disease. All citations were double-checked manually. Finally, 99 genes for 94 autism-related disorders supported by 250 references were included in our data set of 'Syndromic Autism Related Genes'.

In total, we collected as many as 2135 non-syndromic autism-related genes, 99 syndromic autism-related genes, 4544 CNVs and 158 linkage regions. The genes located in the CNV and linkage regions were then retrieved by the UCSC Genome Browser (15).

Evidence collection

To establish the strength of evidence, we collected metadata about each study and result. Supplementary Table S1–S7 list the evidence collected for each type of experimental methods. In summary, for each study of non-syndromic autism, we collected the clinical and demographic features of the samples including ancestral background, country of origin, inclusion and exclusion criteria, number of cases and controls with gender ratio, age at examination and diagnosis criteria. We collected metadata about the experimental design including platform, experimental methods, statistical methods and statistical significance.

For each gene, we estimated how much evidence supports its role in autism by each type of experimental methods and calculated a weighted sum, following a multi-dimensional evidence-based candidate gene prioritization approach (16). First, we assigned initial scores to the genes for each type of experimental methods (Supplementary Table S8). Score 0 is given if there is no positive evidence of the type. Table 1 lists the distribution of the scores for each type. Next, we used a benchmark data set consisting of 21 non-syndromic autism-related genes considered high confidence from six autism reviews (8,9,17–20) (Supplementary Table S9) to calculate the weights. We followed a gene prioritization approach (16) to generate a candidate weight matrix pool consisting of $d^N = 7^6$ weight vectors, where N represents the number of experimental methods and $d = N+1$ represents possible different weights, 1–7 in the weight vectors. A combined score for each gene was then calculated by summing up the products of the scores and corresponding weights from the six experimental methods (16). All the 2135 candidate genes including 21 benchmark genes were sorted by their combined scores. We selected the weight matrix that gave the benchmark genes the highest rank as the optimal weight matrix (Supplementary Table S10). About 95% benchmark genes were ranked among the top 98% of all candidate genes. We chose the lowest combined score, 9, from the benchmark data set as the cutoff of high-confident genes, resulting in a core data set of 383 non-syndromic autism-related genes. Because the definition of 'optimal weight matrix' is always debatable, we provide an online ranking tool to allow users to re-rank the genes interactively by inputting customized weights based on their own experiences and preferences.

Table 1. Score distribution of genes discovered by each experiment method

Experimental methods	Scores	Number of Genes
Genome-wide association studies	1	81
	2	46
	3	5
Expression profiling	1	1320
	2	285
	3	50
Genome-wide CNV studies	1	1086
	2	34
	3	19
Linkage analyses	1	535
	2	43
	3	0
Low scale genetic association studies	1	128
	2	23
	3	12
Other low-scale experimental studies	1	241
	2	37
	3	30

For syndromic autism, we assigned four levels to the autism-related disorders: Level 1 disorders have one reported case with autistic symptoms, Level 2 have two to three cases in a single family, Level 3 have cases in more than one family and Level 4 are reported in multiple review papers (8). Causal genes of Level 3 and 4 disorders were considered high-confident genes in the core dataset.

Functional annotations

To better understand the function of the genes associated with autism, we collected extensive functional information and data, including crosslinks to NCBI Entrez gene (21), OMIM (21), Uniprot (http://www.uniprot.org/) and Ensembl (http://www.ensembl.org/), functional groups based on Gene Ontology (http://www.geneontology.org/), protein–protein interactions from database BioGRID (22), BIND (23) and HPRD (24), and genomic variants from the Database of Genomic Variants (DGV) (25). We linked the genes to three psychiatric disease databases, AlzGene (26), SzGene (27) and PDGene (http://www.pdgene.org/), when the gene is common between these diseases and ASD. Information about homologues of the genes were retrieved from Mouse Genome Informatics (MGI) (28), Zebrafish Model Organism Database (ZFIN) (29) and FlyBase (30). We collected comprehensive mRNA expression profiling data, including ESTs from NCBI Unigene Profiles (21), microarray expression profiles from BioGPS (31) and Allen Brain Atlas (32), and RNA-Seq (33–38). Protein expression evidence at peptide level was retrieved from PRIDE (39) and Peptide Atlas (40). We also collected transcription factor binding sites in the upstream regions of the genes from in-house collection of ChIP-Chip and ChIP-Seq data, miRNAs that may target the genes from miRWalk (41) and TarBase (42), and natural antisense transcripts that may regulate the genes from NATsDB (43). Possible post-translation modifications were retrieved from UniProt and dbPTM (44). We used KOBAS 2.0 (45) to retrieve the pathways that the genes are involved in from BioCyc (46), KEGG Pathway (47), PID (48), PID Reactome (48), PANTHER (49) and Reactome (50) and possible association with other diseases from Disease databases include KEGG Disease (51), FunDO (52,53), GAD (54), NHGRI GWAS Catalog (55) and OMIM (21). Pharmaco-genetics and drug information was collected from Comparative Toxicogenomics Database (CTD) (56), Pharmacogenomics Knowledge Base (57) and DrugBank (58). Supplementary Table S11 summarizes the gene coverage from each source database. The overlap between the genes discovered by expression profiling and those by the other genetic technologies is shown in Supplementary Table S12.

Enriched functional pathways were identified by KOBAS 2.0 (45) and enriched GO terms were identified by DAVID (59). Pathways such as neuroactive ligand–receptor interaction, synapse transmission, and axon guidance were statistically significantly enriched in the core data set (Table 2). In addition to synapse transmission, GO terms such as transmission of nerve impulse, neuron differentiation were also found to be statistically significant (Table 3). The result is consistent with recent findings that synapse development, axon targeting and neuron motility are related to autism etiology (60,61).

DATABASE INTERFACE

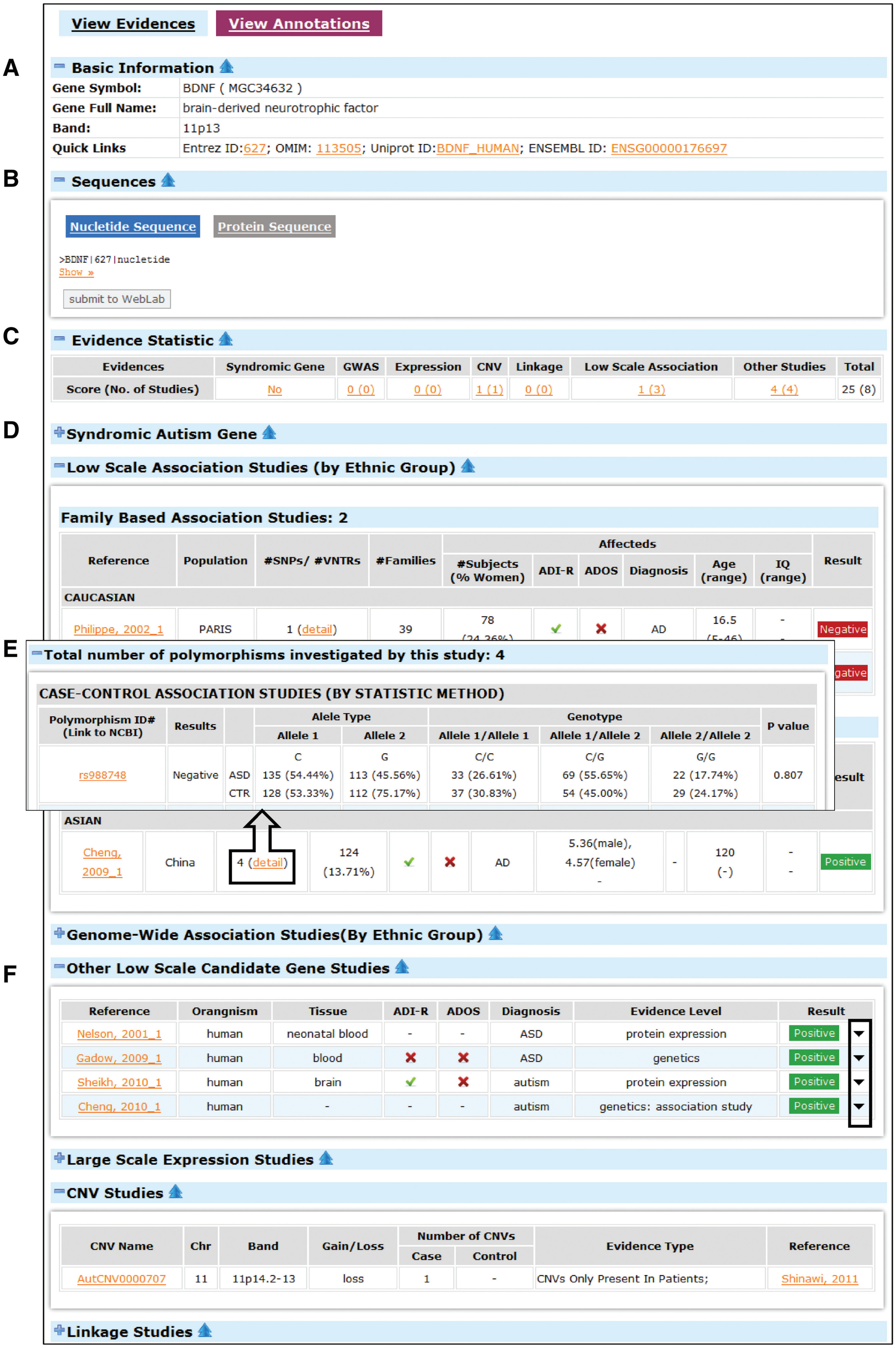
We set up a MySQL relational database to store all the data. A user-friendly web interface for browsing and searching was implemented by PHP and JavaScript, powered by JQuery framework.

Table 2. Top five enriched pathway of the genes in the high-confident core dataset, using KOBAS2.0

Term	Database	ID	P Value	Q Value
Neuroactive ligand-receptor interaction	KEGG PATHWAY	hsa04080	1.03E-11	1.65E-09
Synaptic Transmission	Reactome	REACT:13685	7.50E-10	9.06E-08
Axon guidance	Reactome	REACT:18266	1.29E-08	1.24E-06
Calcium signaling pathway	KEGG PATHWAY	hsa04020	2.25E-08	1.97E-06
Long-term potentiation	KEGG PATHWAY	hsa04720	1.76E-07	9.98E-06

Table 3. Top 10 enriched GO terms of the genes in the high-confident core dataset

GO ID	GO Term	P Value	Q Value
GO:0019226	transmission of nerve impulse	5.44E-29	9.73E-26
GO:0007268	synaptic transmission	4.59E-28	8.21E-25
GO:0007610	synapse	1.05E-23	1.45E-20
GO:0045202	behavior	4.53E-23	8.10E-20
GO:0044057	synapse part	7.21E-22	9.94E-19
GO:0007267	regulation of system process	4.12E-21	7.38E-18
GO:0044456	cell-cell signaling	4.17E-21	7.46E-18
GO:0030182	neuron differentiation	8.21E-19	1.47E-15
GO:0031644	regulation of neurological system process	1.53E-18	2.74E-15
GO:0051969	regulation of transmission of nerve impulse	1.74E-18	3.11E-15



Basic Information

Gene Symbol:	BDNF (MGC34632)
Gene Full Name:	brain-derived neurotrophic factor
Band:	11p13
Quick Links	Entrez ID:627; OMIM: 113505; Uniprot ID:BDNF_HUMAN; ENSEMBL ID: ENSG00000176697

Sequences

>BDNF|627|nucletide

Evidence Statistic

Evidences	Syndromic Gene	GWAS	Expression	CNV	Linkage	Low Scale Association	Other Studies	Total
Score (No. of Studies)	No	0 (0)	0 (0)	1 (1)	0 (0)	1 (3)	4 (4)	25 (8)

Syndromic Autism Gene

Low Scale Association Studies (by Ethnic Group)

Family Based Association Studies: 2

Reference	Population	#SNPs/ #VNTRs	#Families	Affecteds						Result
				#Subjects (% Women)	ADI-R	ADOS	Diagnosis	Age (range)	IQ (range)	
CAUCASIAN										
Philippe, 2002_1	PARIS	1 (detail)	39	78 (24.36%)	✓	✗	AD	16.5 (5-46)	- -	Negative
										gative

Total number of polymorphisms investigated by this study: 4

CASE-CONTROL ASSOCIATION STUDIES (BY STATISTIC METHOD)

Polymorphism ID# (Link to NCBI)	Results		Alele Type		Genotype			P value
			Allele 1	Allele 2	Allele 1/Allele 1	Allele 1/Allele 2	Allele 2/Allele 2	
rs988748	Negative		C	G	C/C	C/G	G/G	0.807
		ASD	135 (54.44%)	113 (45.56%)	33 (26.61%)	69 (55.65%)	22 (17.74%)	
		CTR	128 (53.33%)	112 (75.17%)	37 (30.83%)	54 (45.00%)	29 (24.17%)	

											esult
ASIAN											
Cheng, 2009_1	China	4 (detail)	124 (13.71%)	✓	✗	AD	5.36(male), 4.57(female) -	-	120 (-)	- -	Positive

Genome-Wide Association Studies(By Ethnic Group)

Other Low Scale Candidate Gene Studies

Reference	Orangnism	Tissue	ADI-R	ADOS	Diagnosis	Evidence Level	Result
Nelson, 2001_1	human	neonatal blood	-	-	ASD	protein expression	Positive
Gadow, 2009_1	human	blood	✗	✗	ASD	genetics	Positive
Sheikh, 2010_1	human	brain	✓	✗	autism	protein expression	Positive
Cheng, 2010_1	human	-	-	-	autism	genetics: association study	Positive

Large Scale Expression Studies

CNV Studies

CNV Name	Chr	Band	Gain/Loss	Number of CNVs		Evidence Type	Reference
				Case	Control		
AutCNV0000707	11	11p14.2-13	loss	1	-	CNVs Only Present In Patients;	Shinawi, 2011

Linkage Studies

Figure 1. A typical gene entry in AutismKB. (**A**) Basic information and quick links, (**B**) nucleotide and protein sequences, (**C**) evidence statistics and links to different data sources, (**D**) example of default collapsed data source, (**E**) link and example of polymorphism information and (**F**) example of expanded data source with hidden information.

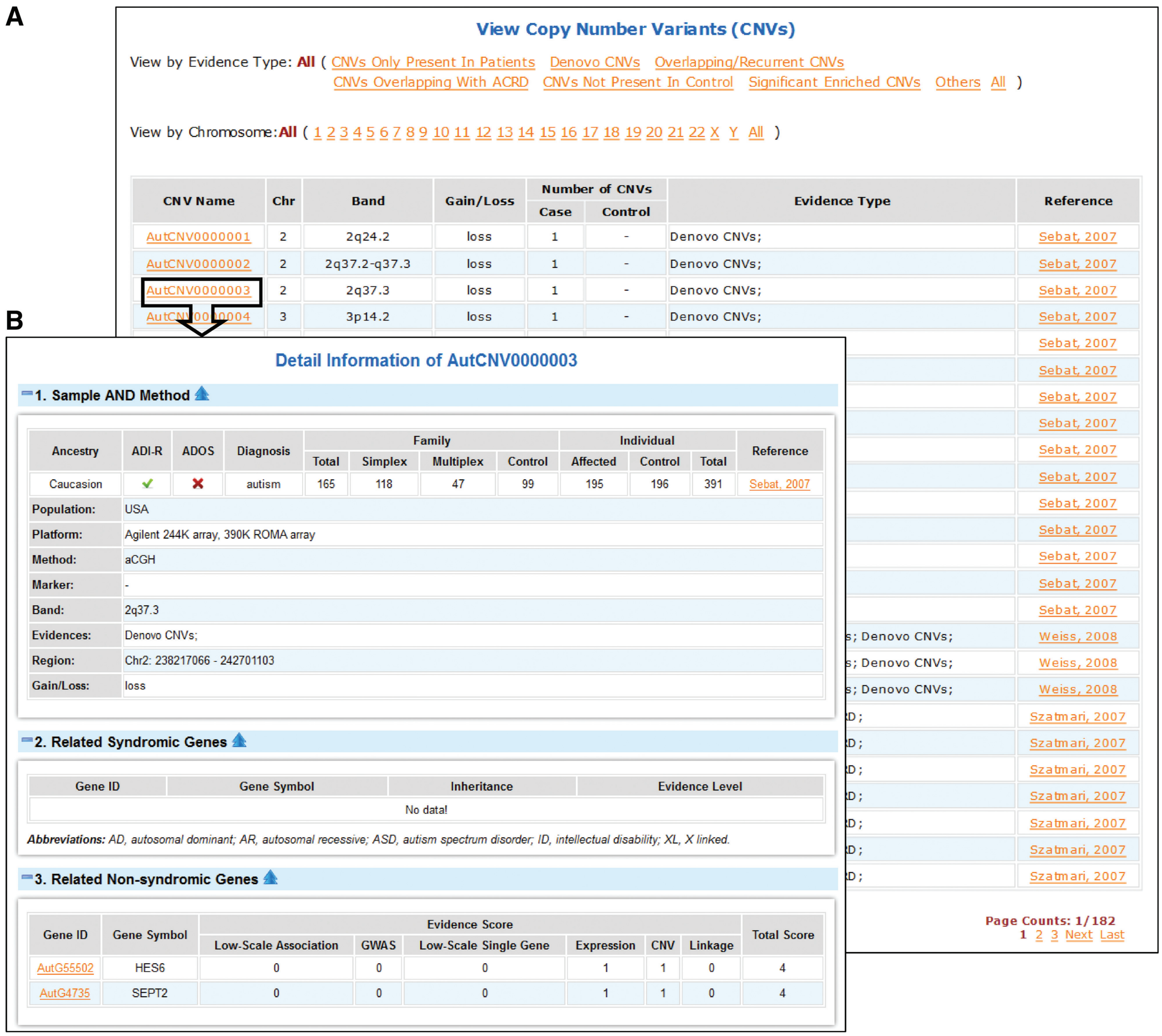
Browsing

Users can browse the data in AutismKB in a variety of ways, including by data sets, experimental methods or chromosome. The gene lists include a summary of information about the genes, hyperlinked to detailed gene evidence and annotation pages. Figure 1 shows a typical AutismKB gene entry. Basic information such as gene symbol, gene name, cytoband and cross links are provided (Figure 1A). Nucleotide sequences and protein sequences can be sent to WebLab (62) for further analysis (Figure 1B). Summaries of supporting evidence and category-specific scores are provided (Figure 1C). Users can click on the hyperlinks of the category-specific score to view different category of evidences. The categories without any evidence are hidden by default (Figure 1D). Users can click on '+' to expand or '–' to collapse different categories. Detailed information of polymorphisms for low scale association studies and GWAS can be found by clicking on 'detail' in the tables (Figure 1E). When exploring other low-scale studies and large-scale expression studies, users can click the down arrow in the right of the table to obtain more information (Figure 1F). Annotations of each gene can be obtained by clicking the label 'view annotation' in the top left.

CNVs are provided by a tabular view with name, cytoband, gain or loss, number, evidence types and reference. Users can use evidence type and chromosome to filter the table (Figure 2A). Clicking on the name can bring the detail information of each CNV including the samples and methods of the study, CNV region, and any syndromic and non-syndromic autism genes in the region (Figure 2B). Users can use chromosome to filter the linkage regions and click on linkage name to view detailed information.

Searching

AutismKB supports both text-based search and sequence-based search. Users can find a quick search box on the top right of each page to search by gene



View Copy Number Variants (CNVs)

View by Evidence Type: All (CNVs Only Present In Patients Denovo CNVs Overlapping/Recurrent CNVs CNVs Overlapping With ACRD CNVs Not Present In Control Significant Enriched CNVs Others All)

View by Chromosome: All (1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y All)

CNV Name	Chr	Band	Gain/Loss	Number of CNVs		Evidence Type	Reference
				Case	Control		
AutCNV0000001	2	2q24.2	loss	1	-	Denovo CNVs;	Sebat, 2007
AutCNV0000002	2	2q37.2-q37.3	loss	1	-	Denovo CNVs;	Sebat, 2007
AutCNV0000003	2	2q37.3	loss	1	-	Denovo CNVs;	Sebat, 2007
AutCNV0000004	3	3p14.2	loss	1	-	Denovo CNVs;	Sebat, 2007
							Sebat, 2007
							Sebat, 2007
							Sebat, 2007
							Sebat, 2007
							Sebat, 2007
							Sebat, 2007
							Sebat, 2007
							Sebat, 2007
							Sebat, 2007
							Sebat, 2007
							Sebat, 2007
						s; Denovo CNVs;	Weiss, 2008
						s; Denovo CNVs;	Weiss, 2008
						s; Denovo CNVs;	Weiss, 2008
						D ;	Szatmari, 2007
						D ;	Szatmari, 2007
						D ;	Szatmari, 2007
						D ;	Szatmari, 2007
						D ;	Szatmari, 2007
						D ;	Szatmari, 2007
						D ;	Szatmari, 2007

Page Counts: 1/182
1 2 3 Next Last

Detail Information of AutCNV0000003

1. Sample AND Method

Ancestry	ADI-R	ADOS	Diagnosis	Family				Individual			Reference
				Total	Simplex	Multiplex	Control	Affected	Control	Total	
Caucasion	✓	✗	autism	165	118	47	99	195	196	391	Sebat, 2007

Population:	USA
Platform:	Agilent 244K array, 390K ROMA array
Method:	aCGH
Marker:	-
Band:	2q37.3
Evidences:	Denovo CNVs;
Region:	Chr2: 238217066 - 242701103
Gain/Loss:	loss

2. Related Syndromic Genes

Gene ID	Gene Symbol	Inheritance	Evidence Level
No data!			

Abbreviations: AD, autosomal dominant; AR, autosomal recessive; ASD, autism spectrum disorder; ID, intellectual disability; XL, X linked.

3. Related Non-syndromic Genes

Gene ID	Gene Symbol	Evidence Score						Total Score
		Low-Scale Association	GWAS	Low-Scale Single Gene	Expression	CNV	Linkage	
AutG55502	HES6	0	0	0	1	1	0	4
AutG4735	SEPT2	0	0	0	1	1	0	4

Figure 2. CNV list and a typical CNV entry in AutismKB. (**A**) the CNV list in AutismKB and (**B**) a typical CNV entry.

symbol. Advanced search was provided to search genes, CNVs, linkage regions by gene name, gene symbol, NCBI Entrez id, Ensemble id, GO terms, UniProt ID, location, score, method and PubMed ID. Finally, a BLAST search against the nucleotide or protein sequences of all AutismKB genes is also available.

CONCLUSION

AutismKB is a comprehensive knowledgebase of autism-related genes, CNVs and linkage regions with extensive evidence and annotations. AutismKB will be updated periodically. We hope that it can be a valuable resource for the autism research community.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1-12.

ACKNOWLEDGEMENTS

We thank Ge Gao, Chuan-Yun Li, Yong-Xin Ye and Ying-Fu Zhong for useful comments on the web interface.

FUNDING

Funding for open access charge: National Outstanding Young Investigator Award from Natural Science Foundation of China (grant number: 31025014); 973 Basic Research Program (grant number: 2011CBA01102); scholarships from Merck and Johnson and Johnson.

Conflict of interest statement. None declared.

REFERENCES

1. American Psychiatric Association. (2000) *Diagnostic and Statistical Manual of Mental Disorders: DSM-IV-TR*. American Psychiatric Publishing, Inc, Arlington, VA.
2. Kogan,M.D., Blumberg,S.J., Schieve,L.A., Boyle,C.A., Perrin,J.M., Ghandour,R.M., Singh,G.K., Strickland,B.B., Trevathan,E. and van Dyck,P.C. (2009) prevalence of parent-reported diagnosis of autism spectrum disorder among children in the US, 2007. *Pediatrics*, **124**, 1395–1403.
3. Kim,Y.S., Leventhal,B.L., Koh,Y.J., Fombonne,E., Laska,E., Lim,E.C., Cheon,K.A., Kim,S.J., Kim,Y.K., Lee,H. *et al.* (2011) Prevalence of autism spectrum disorders in a total population sample. *Am. J. Psychiatry*, **168**, 904–912.
4. Ganz,M.L. (2006) The Costs of Autism. In: Moldin,S.O. and Rubenstein,J.L.R. (eds), *Understanding Autism: from Basic Neuroscience to Treatment*. CRC Press, Boca Raton, FL, pp. 476–498.
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. Bailey,A., Le Couteur,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.
7. Betancur,C. (2011) Etiological heterogeneity in autism spectrum disorders: more than 100 genetic and genomic disorders and still counting. *Brain Res.*, **1380**, 42–77.
8. Abrahams,B.S. and Geschwind,D.H. (2008) Advances in autism genetics: on the threshold of a new neurobiology. *Nat. Rev. Genet.*, **9**, 341–355.
9. State,M.W. (2010) The genetics of child psychiatric disorders: focus on autism and Tourette syndrome. *Neuron*, **68**, 254–269.
10. Lander,E.S. (2011) Initial impact of the sequencing of the human genome. *Nature*, **470**, 187–197.
11. Banerjee-Basu,S. and Packer,A. (2010) SFARI Gene: an evolving database for the autism research community. *Dis. Model Mech.*, **3**, 133–135.
12. Basu,S.N., Kollu,R. and Banerjee-Basu,S. (2009) AutDB: a gene reference resource for autism research. *Nucleic Acids Res.*, **37**, D832–D836.
13. Matuszek,G. and Talebizadeh,Z. (2009) Autism Genetic Database (AGD): a comprehensive database including autism susceptibility gene-CNVs integrated with known noncoding RNAs and fragile sites. *BMC Med. Genet.*, **10**, 102.
14. 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.
15. Fujita,P.A., Rhead,B., Zweig,A.S., Hinrichs,A.S., Karolchik,D., Cline,M.S., Goldman,M., Barber,G.P., Clawson,H., Coelho,A. *et al.* (2011) The UCSC Genome Browser database: update 2011. *Nucleic Acids Res.*, **39**, D876–D882.
16. Sun,J., Jia,P., Fanous,A.H., Webb,B.T., van den Oord,E.J., Chen,X., Bukszar,J., Kendler,K.S. and Zhao,Z. (2009) A multi-dimensional evidence-based candidate gene prioritization approach for complex diseases-schizophrenia as a case. *Bioinformatics*, **25**, 2595–6602.
17. Freitag,C.M. (2007) The genetics of autistic disorders and its clinical relevance: a review of the literature. *Mol. Psychiatry*, **12**, 2–22.
18. Klauck,S.M. (2006) Genetics of autism spectrum disorder. *Eur. J. Hum. Genet.*, **14**, 714–720.
19. Losh,M., Sullivan,P.F., Trembath,D. and Piven,J. (2008) Current developments in the genetics of autism: from phenome to genome. *J. Neuropathol. Exp. Neurol.*, **67**, 829–837.
20. Muhle,R., Trentacoste,S.V. and Rapin,I. (2004) The genetics of autism. *Pediatrics*, **113**, e472–e486.
21. Sayers,E.W., Barrett,T., Benson,D.A., Bolton,E., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Federhen,S. *et al.* (2011) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **39**, D38–51.
22. Stark,C., Breitkreutz,B.J., Chatr-Aryamontri,A., Boucher,L., Oughtred,R., Livstone,M.S., Nixon,J., Van Auken,K., Wang,X., Shi,X. *et al.* (2011) The BioGRID Interaction Database: 2011 update. *Nucleic Acids Res.*, **39**, D698–704.
23. Gilbert,D. (2005) Biomolecular interaction network database. *Brief Bioinform.*, **6**, 194–198.
24. Keshava Prasad,T.S., Goel,R., Kandasamy,K., Keerthikumar,S., Kumar,S., Mathivanan,S., Telikicherla,D., Raju,R., Shafreen,B., Venugopal,A. *et al.* (2009) Human Protein Reference Database—2009 update. *Nucleic Acids Res.*, **37**, D767–D772.
25. Zhang,J., Feuk,L., Duggan,G.E., Khaja,R. and Scherer,S.W. (2006) Development of bioinformatics resources for display and analysis of copy number and other structural variants in the human genome. *Cytogenet Genome Res.*, **115**, 205–214.
26. Bertram,L., McQueen,M.B., Mullin,K., Blacker,D. and Tanzi,R.E. (2007) Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. *Nat. Genet.*, **39**, 17–23.
27. Allen,N.C., Bagade,S., McQueen,M.B., Ioannidis,J.P., Kavvoura,F.K., Khoury,M.J., Tanzi,R.E. and Bertram,L. (2008) Systematic meta-analyses and field synopsis of genetic association studies in schizophrenia: the SzGene database. *Nat. Genet.*, **40**, 827–834.
28. Blake,J.A., Bult,C.J., Kadin,J.A., Richardson,J.E. and Eppig,J.T. (2011) The Mouse Genome Database (MGD): premier model organism resource for mammalian genomics and genetics. *Nucleic Acids Res.*, **39**, D842–D848.
29. Bradford,Y., Conlin,T., Dunn,N., Fashena,D., Frazer,K., Howe,D.G., Knight,J., Mani,P., Martin,R., Moxon,S.A. *et al.*

(2011) ZFIN: enhancements and updates to the Zebrafish Model Organism Database. *Nucleic Acids Res.*, **39**, D822–D829.

30. Tweedie,S., Ashburner,M., Falls,K., Leyland,P., McQuilton,P., Marygold,S., Millburn,G., Osumi-Sutherland,D., Schroeder,A., Seal,R. *et al.* (2009) FlyBase: enhancing Drosophila Gene Ontology annotations. *Nucleic Acids Res.*, **37**, D555–D559.
31. Su,A.I., Wiltshire,T., Batalov,S., Lapp,H., Ching,K.A., Block,D., Zhang,J., Soden,R., Hayakawa,M., Kreiman,G. *et al.* (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl Acad. Sci. USA*, **101**, 6062–6067.
32. Jones,A.R., Overly,C.C. and Sunkin,S.M. (2009) The Allen Brain Atlas: 5 years and beyond. *Nat. Rev. Neurosci.*, **10**, 821–828.
33. Wang,E.T., Sandberg,R., Luo,S., Khrebtukova,I., Zhang,L., Mayr,C., Kingsmore,S.F., Schroth,G.P. and Burge,C.B. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**, 470–476.
34. Wu,J.Q., Habegger,L., Noisa,P., Szekely,A., Qiu,C., Hutchison,S., Raha,D., Egholm,M., Lin,H., Weissman,S. *et al.* (2010) Dynamic transcriptomes during neural differentiation of human embryonic stem cells revealed by short, long, and paired-end sequencing. *Proc. Natl Acad. Sci. USA*, **107**, 5254–5259.
35. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
36. Mortazavi,A., Williams,B.A., McCue,K., Schaeffer,L. and Wold,B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621–628.
37. Trapnell,C., Pachter,L. and Salzberg,S.L. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–1111.
38. Trapnell,C., Williams,B.A., Pertea,G., Mortazavi,A., Kwan,G., van Baren,M.J., Salzberg,S.L., Wold,B.J. and Pachter,L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, **28**, 511–515.
39. Vizcaino,J.A., Cote,R., Reisinger,F., Foster,J.M., Mueller,M., Rameseder,J., Hermjakob,H. and Martens,L. (2009) A guide to the Proteomics Identifications Database proteomics data repository. *Proteomics*, **9**, 4276–4283.
40. Farrah,T., Deutsch,E.W., Omenn,G.S., Campbell,D.S., Sun,Z., Bletz,J.A., Mallick,P., Katz,J.E., Malmstrom,J., Ossola,R. *et al.* (2011) A high-confidence human plasma proteome reference set with estimated concentrations in PeptideAtlas. *Mol. Cell Proteomics*, **10**, M110 006353.
41. Dweep,H., Sticht,C., Pandey,P. and Gretz,N. (2011) miRWalk - Database: Prediction of possible miRNA binding sites by 'walking' the genes of three genomes. *J. Biomed Inform*, **44**, 839–847.
42. Papadopoulos,G.L., Reczko,M., Simossis,V.A., Sethupathy,P. and Hatzigeorgiou,A.G. (2009) The database of experimentally supported targets: a functional update of TarBase. *Nucleic Acids Res.*, **37**, D155–D158.
43. Zhang,Y., Liu,X.S., Liu,Q.R. and Wei,L. (2006) Genome-wide in silico identification and analysis of cis natural antisense transcripts (cis-NATs) in ten species. *Nucleic Acids Res.*, **34**, 3465–3475.
44. Lee,T.Y., Hsu,J.B., Chang,W.C., Wang,T.Y., Hsu,P.C. and Huang,H.D. (2009) A comprehensive resource for integrating and displaying protein post-translational modifications. *BMC Res. Notes*, **2**, 111.
45. Xie,C., Mao,X., Huang,J., Ding,Y., Wu,J., Dong,S., Kong,L., Gao,G., Li,C.Y. and Wei,L. (2011) KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. *Nucleic Acids Res.*, **39**, W316–W322.
46. Karp,P.D., Ouzounis,C.A., Moore-Kochlacs,C., Goldovsky,L., Kaipa,P., Ahren,D., Tsoka,S., Darzentas,N., Kunin,V. and Lopez-Bigas,N. (2005) Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res.*, **33**, 6083–6089.
47. Kanehisa,M., Araki,M., Goto,S., Hattori,M., Hirakawa,M., Itoh,M., Katayama,T., Kawashima,S., Okuda,S., Tokimatsu,T. *et al.* (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, **36**, D480–D484.
48. Schaefer,C.F., Anthony,K., Krupa,S., Buchoff,J., Day,M., Hannay,T. and Buetow,K.H. (2009) PID: the Pathway Interaction Database. *Nucleic Acids Res.*, **37**, D674–D679.
49. Thomas,P.D., Campbell,M.J., Kejariwal,A., Mi,H., Karlak,B., Daverman,R., Diemer,K., Muruganujan,A. and Narechania,A. (2003) PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res.*, **13**, 2129–2141.
50. Croft,D., O'Kelly,G., Wu,G., Haw,R., Gillespie,M., Matthews,L., Caudy,M., Garapati,P., Gopinath,G., Jassal,B. *et al.* (2011) Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.*, **39**, D691–D697.
51. Kanehisa,M., Goto,S., Furumichi,M., Tanabe,M. and Hirakawa,M. (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.*, **38**, D355–D360.
52. Osborne,J.D., Flatow,J., Holko,M., Lin,S.M., Kibbe,W.A., Zhu,L.J., Danila,M.I., Feng,G. and Chisholm,R.L. (2009) Annotating the human genome with Disease Ontology. *BMC Genomics*, **10(Suppl. 1)**, S6.
53. Du,P., Feng,G., Flatow,J., Song,J., Holko,M., Kibbe,W.A. and Lin,S.M. (2009) From disease ontology to disease-ontology lite: statistical methods to adapt a general-purpose ontology for the test of gene-ontology associations. *Bioinformatics*, **25**, i63–i68.
54. Becker,K.G., Barnes,K.C., Bright,T.J. and Wang,S.A. (2004) The genetic association database. *Nat. Genet.*, **36**, 431–432.
55. Hindorff,L.A., Sethupathy,P., Junkins,H.A., Ramos,E.M., Mehta,J.P., Collins,F.S. and Manolio,T.A. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci USA*, **106**, 9362–9367.
56. Davis,A.P., King,B.L., Mockus,S., Murphy,C.G., Saraceni-Richards,C., Rosenstein,M., Wiegers,T. and Mattingly,C.J. (2011) The Comparative Toxicogenomics Database: update 2011. *Nucleic Acids Res.*, **39**, D1067–D1072.
57. Relling,M.V., Gardner,E.E., Sandborn,W.J., Schmiegelow,K., Pui,C.H., Yee,S.W., Stein,C.M., Carrillo,M., Evans,W.E. and Klein,T.E. (2011) Clinical Pharmacogenetics Implementation Consortium guidelines for thiopurine methyltransferase genotype and thiopurine dosing. *Clin. Pharmacol. Ther.*, **89**, 387–391.
58. Knox,C., Law,V., Jewison,T., Liu,P., Ly,S., Frolkis,A., Pon,A., Banco,K., Mak,C., Neveu,V. *et al.* (2011) DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res.*, **39**, D1035–D1041.
59. Huang da,W., Sherman,B.T. and Lempicki,R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.
60. Gilman,S.R., Iossifov,I., Levy,D., Ronemus,M., Wigler,M. and Vitkup,D. (2011) Rare de novo variants associated with autism implicate a large functional network of genes involved in formation and function of synapses. *Neuron*, **70**, 898–907.
61. Voineagu,I., Wang,X., Johnston,P., Lowe,J.K., Tian,Y., Horvath,S., Mill,J., Cantor,R.M., Blencowe,B.J. and Geschwind,D.H. (2011) Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature*, **474**, 380–384.
62. Liu,X., Wu,J., Wang,J., Zhao,S., Li,Z., Kong,L., Gu,X., Luo,J. and Gao,G. (2009) WebLab: a data-centric, knowledge-sharing bioinformatic platform. *Nucleic Acids Res.*, **37**, W33–W39.