The Genetic Association Database

To the editor:

The increasing availability of polymorphism data has allowed more gene association studies to be carried out and the number of published genetic association studies is growing rapidly. Studies done secondarily to successful linkage studies over the last decade have also fueled the increase in published association studies. Although there are single-nucleotide polymorphism and human variation databases[1,2], there is currently no public repository for genetic association data. It is difficult to query association data in a systematic manner or to integrate association data with other molecular databases. OMIM[3], the main repository of genetic information for mendelian disorders, is largely text based and is of a historical narrative design, making it difficult to compare large sets of molecular data. Moreover, OMIM archives mature, high-quality data of high significance, the standard in rare mendelian disorders. Although this data is useful, OMIM does not routinely collect findings of lower significance or negative findings. The study of nonmendelian, common complex disorders is often a struggle to find disease relevance with lower significance values, and often conflicting evidence. Negative data are often not reported or are marginalized into obscure and less accessible scientific journals, resulting in a publication bias favoring positive genetic associations[4]. Here, we describe the development of a genetic association database (GAD; http://geneticassociationdb.nih.gov) that aims to collect, standardize and archive genetic association study data and to make it easily accessible to the scientific community.

There are no standards for designing, implementing, interpreting or reporting association studies (*e.g.*, sample size, replication, significant *P* values), although guidelines have been suggested[4–7]. The literature is filled with alternative, idiosyncratic and arbitrary gene names and gene symbols, as well as a continuum of phenotypic descriptions. Studies using arbitrary nomenclature continue to be published, making cross-comparison and meta-analysis difficult. One goal of GAD is to standardize molecular nomenclature in the archival process by including official HUGO gene symbols. After this assignment, each record is annotated with links to molecular databases (LocusLink, GeneCards, HapMap, etc.) and reference databases (PubMed, CDC), among others. Once they are standardized, integrating association data with other molecular databases, data mining tools, annotation and future sources of molecular data (*e.g.*, gene interactions, quantitative trait loci) can be done systematically. Moreover, cross-comparison and meta-analysis of studies becomes more efficient.

There are three main components of GAD: a web interface, Perl modules and the database, which uses the Oracle RDBMS. The database has three layers; gene and disease data are organized into a large fact table in a middle layer with dimensional views on the top layer. The bottom layer contains the tools for adding, editing, batch loading and downloading data to and from the database.

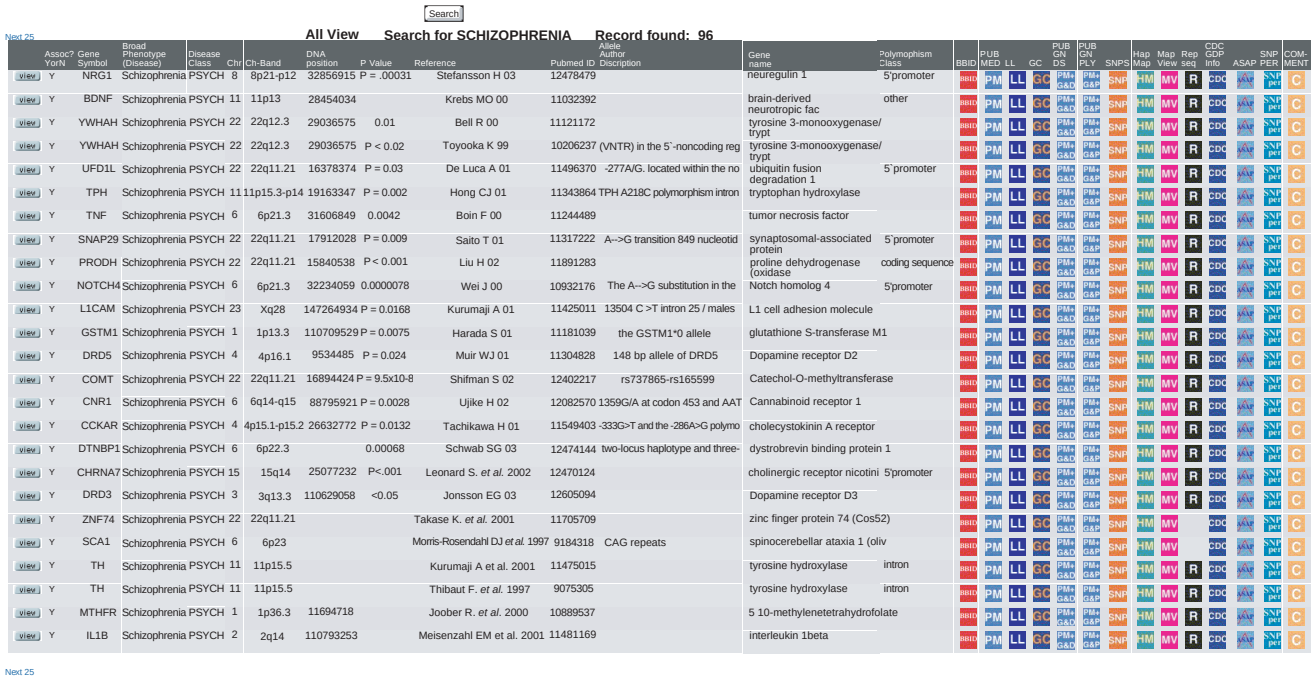
We identify data fields common to genetic association studies, such as disease phenotypes, sample sizes, significance values, population information and allele descriptions. These fields are grouped into five views relevant to disease phenotypes (Disease View), gene-based molecular data (Gene View), chromosomal and mutation information (CH-SNP-Hap View), Reference View and All View. **Table 1** shows a summary of the current contents in the database.

Query tools include key-word-search functions that permit field-specific searches, advanced combinatorial queries and pull-down selections of controlled vocabularies (**Fig. 1**). Batch searches are done against an aggregate table, allowing the user to input a list of genes (300) at once. In this way, batch results from high-throughput assays, such as microarrays, proteomic, cDNA sequencing and SAGE (serial analysis of gene expression), can be rapidly queried in the context of human disease associations.

Of particular interest are phenotypic descriptions captured at multiple levels. A top level 'disease class' is assigned, followed by 'disease' from the original paper. If studies recognize clinical subphenotypes, endophenotypes or intermediate phenotypes, this is noted in 'narrow phenotype'. Moreover, certain alleles have defined molecular characteristics and are noted under 'molecular phenotype'. These molecular and pathway variants may have a closer relationship to a polymorphism than to the end-stage complex phenotype, such as altered transcription due to a promoter polymorphism (*IL6*) or serum levels of ACE. Using this hierarchical phenotypic assignment makes it easier to consider molecular phenotypes in the context of end-stage disease. In some cases, although independent end-stage diseases may not share

Table 1 Current contents of the GAD

Unique categories	Number of records
Current records	5,937
Unique genes	1,647
Population designations	328
Unique studies	4,940
Diseases or clinical traits	1,943
Positive association	3,195
Lack of association	1,367
Unassigned association	1,374
Disease classes	**Number of records**
Immune	1,296
Cardiovascular	619
Metabolic	389
Neurodegenerative	341
Psychiatric	516
Cancer	429
Aging	47
Infection	29
Other	217
Unclassified	2,053


Search

All View Search for SCHIZOPHRENIA Record found: 96

Next 25

	Assoc? Y or N	Gene Symbol	Broad Phenotype (Disease)	Disease Class	Chr	Ch-Band	DNA position	P Value	Reference	Pubmed ID	Allele Author Discription	Gene name	Polymorphism Class
view	Y	NRG1	Schizophrenia	PSYCH	8	8p21-p12	32856915	P = .00031	Stefansson H 03	12478479		neuregulin 1	5'promoter
view	Y	BDNF	Schizophrenia	PSYCH	11	11p13	28454034		Krebs MO 00	11032392		brain-derived neurotropic fac	other
view	Y	YWHAH	Schizophrenia	PSYCH	22	22q12.3	29036575	0.01	Bell R 00	11121172		tyrosine 3-monooxygenase/ trypt	
view	Y	YWHAH	Schizophrenia	PSYCH	22	22q12.3	29036575	P < 0.02	Toyooka K 99	10206237	(VNTR) in the 5'-noncoding reg	tyrosine 3-monooxygenase/ trypt	
view	Y	UFD1L	Schizophrenia	PSYCH	22	22q11.21	16378374	P = 0.03	De Luca A 01	11496370	-277A/G. located within the no	ubiquitin fusion degradation 1	5'promoter
view	Y	TPH	Schizophrenia	PSYCH	11	11p15.3-p14	19163347	P = 0.002	Hong CJ 01	11343864	TPH A218C polymorphism intron	tryptophan hydroxylase	
view	Y	TNF	Schizophrenia	PSYCH	6	6p21.3	31606849	0.0042	Boin F 00	11244489		tumor necrosis factor	
view	Y	SNAP29	Schizophrenia	PSYCH	22	22q11.21	17912028	P = 0.009	Saito T 01	11317222	A-->G transition 849 nucleotid	synaptosomal-associated protein	5'promoter
view	Y	PRODH	Schizophrenia	PSYCH	22	22q11.21	15840538	P< 0.001	Liu H 02	11891283		proline dehydrogenase (oxidase	coding sequence
view	Y	NOTCH4	Schizophrenia	PSYCH	6	6p21.3	32234059	0.0000078	Wei J 00	10932176	The A-->G substitution in the	Notch homolog 4	5'promoter
view	Y	L1CAM	Schizophrenia	PSYCH	23	Xq28	147264934	P = 0.0168	Kurumaji A 01	11425011	13504 C >T intron 25 / males	L1 cell adhesion molecule	
view	Y	GSTM1	Schizophrenia	PSYCH	1	1p13.3	110709529	P = 0.0075	Harada S 01	11181039	the GSTM1*0 allele	glutathione S-transferase M1	
view	Y	DRD5	Schizophrenia	PSYCH	4	4p16.1	9534485	P = 0.024	Muir WJ 01	11304828	148 bp allele of DRD5	Dopamine receptor D2	
view	Y	COMT	Schizophrenia	PSYCH	22	22q11.21	16894424	P = 9.5x10-8	Shifman S 02	12402217	rs737865-rs165599	Catechol-O-methyltransferase	
view	Y	CNR1	Schizophrenia	PSYCH	6	6q14-q15	88795921	P = 0.0028	Ujike H 02	12082570	1359G/A at codon 453 and AAT	Cannabinoid receptor 1	
view	Y	CCKAR	Schizophrenia	PSYCH	4	4p15.1-p15.2	26632772	P = 0.0132	Tachikawa H 01	11549403	-333G>T and the -286A>G polymo	cholecystokinin A receptor	
view	Y	DTNBP1	Schizophrenia	PSYCH	6	6p22.3		0.00068	Schwab SG 03	12474144	two-locus haplotype and three-	dystrobrevin binding protein 1	
view	Y	CHRNA7	Schizophrenia	PSYCH	15	15q14	25077232	P<.001	Leonard S. *et al.* 2002	12470124		cholinergic receptor nicotini	5'promoter
view	Y	DRD3	Schizophrenia	PSYCH	3	3q13.3	110629058	<0.05	Jonsson EG 03	12605094		Dopamine receptor D3	
view	Y	ZNF74	Schizophrenia	PSYCH	22	22q11.21			Takase K. *et al.* 2001	11705709		zinc finger protein 74 (Cos52)	
view	Y	SCA1	Schizophrenia	PSYCH	6	6p23			Morris-Rosendahl DJ *et al.* 1997	9184318	CAG repeats	spinocerebellar ataxia 1 (oliv	
view	Y	TH	Schizophrenia	PSYCH	11	11p15.5			Kurumaji A et al. 2001	11475015		tyrosine hydroxylase	intron
view	Y	TH	Schizophrenia	PSYCH	11	11p15.5			Thibaut F. *et al.* 1997	9075305		tyrosine hydroxylase	intron
view	Y	MTHFR	Schizophrenia	PSYCH	1	1p36.3	11694718		Joober R. *et al.* 2000	10889537		5 10-methylenetetrahydrofolate	
view	Y	IL1B	Schizophrenia	PSYCH	2	2q14	110793253		Meisenzahl EM et al. 2001	11481169		interleukin 1beta	

Next 25

Figure 1 A simple search of positive associations for the disease schizophrenia. Fields in this view include Official Gene Symbol, Disease Phenotype, Disease Class, Chromosome, Chromosome Band, Genomic DNA Position, *P* Value, Reference, PubMed ID and Allele.

overt similarities at a clinical level, the genetic factors that contribute to those diseases may be shared at a molecular level[8,9]. The development of a hierarchy of phenotypes, from broad to specific, may allow classification of diseases, subphenotypes and molecular parameters of disease and their relationship to complex traits.

GAD is an archive of published genetic association studies that provides a comprehensive, public, web-based repository of molecular, clinical and study parameters for >5,000 human genetic association studies at this time. This approach will allow the systematic analysis of complex common human genetic disease in the context of modern high-throughput assay systems and current annotated molecular nomenclature.

Kevin G Becker[1], Kathleen C Barnes[2], Tiffani J Bright[1] & S Alex Wang[3]

[1]Gene Expression and Genomics Unit, 333 Cassell Drive, National Institute on Aging, National Institutes of Health, Baltimore, Maryland 21224, USA. [2]Johns Hopkins Asthma and Allergy Center, Johns Hopkins University, 5501 Hopkins Bayview Circle, Baltimore, Maryland 21224, USA. [3]Division of Computational Bioscience, Center for Information Technology, National Institutes of Health, Bethesda, Maryland 20892, USA. Correspondence should be addressed to K.G.B. (beckerk@grc.nia.nih.gov).

1. Thorisson, G.A. & Stein, L.D. *Nucleic Acids Res.* **31**, 124–127 (2003).
2. Sherry, S.T. *et al. Nucleic Acids Res.* **29**, 308–311 (2001).
3. Hamosh, A. *et al. Nucleic Acids Res.* **30**, 52–55 (2002).
4. Coope, D.N., Nussbaum, R.L. & Krawczak, M. *Hum. Genet.* **110**, 207–208 (2002).
5. Anonymous. *Nat. Genet.* **22**, 1–2 (1999).
6. Dahlman, I. *et al. Nat. Genet.* **30**, 149–150 (2002).
7. Funalot, B., Varenne, O. & Mas, J.L. *Nat. Genet.* **36**, 3 (2004).
8. Mira, M.T. *et al. Nature* **427**, 636–640 (2004).
9. Becker, K.G. *Med. Hypotheses* **62**, 309–317 (2004)

Is mismatch repair really required for ionizing radiation–induced DNA damage signaling?

To the editor:

The MMR system has evolved to increase the fidelity of DNA replication and homologous recombination[1]. MMR is also implicated in the processing of other types of DNA damage, as mammalian cells with defective MMR are tolerant to S_N1 type methylating agents such as N-methyl-N′-nitro-N-nitrosoguanidine and to 6-thioguanine and cisplatin[2].

Reports describing the differential sensitivity of MMR-proficient and -deficient cells to ionizing radiation raised some controversy, as MMR-deficient cells were found to be slightly more resistant to ionizing radiation in some laboratories[3] but either equally[4] or less resistant[5] in others. The survival differences were also questioned, because MMR status was reported to affect the length of the G2-M checkpoint rather than cell viability[6]. A report by Brown *et al.*[7] has reopened this discussion by describing the requirement of a functional MMR system for activating the S-phase checkpoint and signaling of ionizing radiation–induced damage.

The aforementioned studies used matched MMR-proficient and -deficient mouse or human cell lines. Given that the establishment of these lines involved long periods of growth in cell culture, and that the MMR-deficient