Alcoholism: Collaborative Study on the Genetics of Alcoholism (COGA)

John P Rice, *Washington University School of Medicine, St Louis, Missouri, USA*
Nancy L Saccone, *Washington University School of Medicine, St Louis, Missouri, USA*
Tatiana Foroud, *Indiana University School of Medicine, Indianapolis, Indiana, USA*
Howard J Edenberg, *Indiana University School of Medicine, Indianapolis, Indiana, USA*
John I Nurnberger Jr, *Indiana University School of Medicine, Indianapolis, Indiana, USA*
Alison Goate, *Washington University School of Medicine, St Louis, Missouri, USA*
Raymond R Crowe, *University of Iowa College of Medicine, Iowa City, Iowa, USA*
Victor Hesselbrock, *University of Connecticut Health Center, Farmington, Connecticut, USA*
Marc Schuckit, *University of California at San Diego, San Diego, California, USA*
Bernice Porjesz, *SUNY HSC, Brooklyn, New York, USA*
Theodore Reich, *Washington University School of Medicine, St Louis, Missouri, USA*
Henri Begleiter, *SUNY HSC, Brooklyn, New York, USA*

Intermediate article

Article contents

Twin, adoption and family studies provide evidence for a genetic component for the susceptibility to alcoholism. The Collaborative Study on the Genetics of Alcoholism (COGA) is a comprehensive family study aimed at identifying specific genetic risk factors.

Introduction

To more efficiently identify novel genetic loci contributing to susceptibility to alcoholism, recent studies have focused on a genome-wide approach. After the collection of extended pedigrees with multiple members diagnosed with alcoholism (alcohol dependence), genetic analysis techniques can be employed to evaluate the evidence for linkage throughout the genome. Such a strategy was employed by the Collaborative Study on the Genetics of Alcoholism (COGA), a multisite family study designed to include both alcoholic and community control probands and their biological relatives.

Samples

Alcoholic probands were recruited from consecutive admissions to both inpatient and outpatient treatment facilities. Probands were required to meet both American Psychiatric Association DSM-III-R criteria and Feighner diagnostic criteria at the definite level for alcoholism. These 'Stage I' probands and their biological first-degree relatives (over the age of 6 years) completed a formal psychiatric diagnostic interview. Families in which the proband and at least two other first-degree relatives had a diagnosis of alcohol dependence were designated as 'Stage II' families and extended through alcoholic relatives to include their first-degree relatives also. Subjects in Stage II families have, in addition to the comprehensive diagnostic interview (the Semi-Structured Assessment for the Genetics of Alcoholism (SSAGA)), completed a battery of neuropsychological tests and an electrophysiological event-related potential (ERP) assessment employing visual and auditory paradigms. In addition, blood samples were obtained for biochemical analysis and the establishment of lymphoblastoid cell lines.

Control families were identified through dental clinics, drivers' license records and health maintenance organizations. Control families consisted of at least three children over the age of 14 years and their biological parents. These families were evaluated using the Stage II protocol as described above.

Genotyping

A subset of Stage II families was selected to be genotyped in two samples. The initial genome screen was described by Reich *et al.* (1998), and the analysis of the replication sample was described by Foroud *et al.* (2000). The initial sample consisted of 105 families that contained 987 individuals with genotypic data. The replication sample consisted of 157 families with genotypic data for 1295 individuals.

Linkage Results

Alcohol dependence

The initial linkage analyses (Reich *et al.*, 1998) used only one definition of alcoholism: individuals were defined as affected if they fulfilled criteria for alcoholism based on both DSM-III-R and Feighner criteria (termed 'COGA' criteria). Using 382 affected sibling pairs, regions on chromosomes 1, 2 and 7 were identified as harboring genes that predispose an individual to alcoholism (Reich *et al.*, 1998). The most significant finding of the study was on chromosome 7, with a lod (logarithm of the odds) score of 3.5 near the marker D7S1793. On chromosome 1, a peak lod score of 2.9 was found near the marker D1S1588. A second locus on chromosome 1, about 60 cM from the stronger linkage finding, had a lod score of 1.6. A lod score of 1.8 was found on chromosome 2, near the marker D2S1790.

Additional analyses using individuals without a diagnosis of alcoholism provided evidence for a protective locus on chromosome 4 near the alcohol dehydrogenase gene family. Linkage to chromosome 4 was particularly interesting since the COGA sample has few Asian families, but rather consists primarily of non-Hispanic Caucasian and African-American families. This linkage result on chromosome 4 suggests that the protective effects of alcohol dehydrogenase are not limited to the Asian population.

Subsequently, linkage analyses were completed in a replication dataset of 157 pedigrees ascertained and evaluated using criteria identical to those used in the initial sample (Foroud *et al.*, 2000). Genetic analyses of affected sibling pairs supported linkage to chromosome 1 (lod = 1.6) in the replication dataset as well as in a combined analysis of the two samples (lod = 2.6). Evidence of linkage to chromosome 7 decreased in the combined data (lod = 2.9). The lod score on chromosome 2 in the initial dataset increased following genotyping of additional markers; however, combined analyses of the two datasets resulted in overall lower lod scores (lod = 1.8) on chromosome 2. A new finding of linkage to chromosome 3 was identified in the replication dataset (lod = 3.4). Thus, analyses of a second large sample of alcoholic families provided further evidence of genetic susceptibility loci on chromosomes 1 and 7. Genetic analyses also identified susceptibility loci on chromosomes 2 and 3 that may act in only one of the two datasets.

The linkage results are summarized in **Table 1**. It should be noted that the marker density was increased between the analyses reported by Reich *et al.* (1998) and those reported by Foroud *et al.* (2000). **Table 1** summarizes the results in terms of the updated map. For example, the lod score of 3.5 near D7S1793 in Reich *et al.* (1998) in the initial sample was found to be 2.0 when reanalyzed with the dense map. It is this latter value that is given in the table.

Analyses of narrower alcohol phenotypes

A potentially more powerful methodology to identify genes for alcoholism may be to develop novel, narrower phenotypes that define a genetically more homogeneous sample, wherein a smaller number of genes for susceptibility to alcoholism are segregating. To this end, a series of latent class analyses were performed in the COGA dataset to identify a more homogeneous phenotype that included some components of physiological dependence (Foroud *et al.*, 1998). The analyses used items from the interview data thought to reflect more severe alcohol symptoms and nondiagnostic items, such as maximum amount of alcohol consumed. Using 11 items, a four-class solution was computed. The classes may be characterized as follows: an unaffected group (class 1) that contained 47% of the individuals ($n = 419$); a mildly problematic group that contained 23% of the sample ($n = 172$); a moderately affected group (class 3) that included 17% of the individuals ($n = 137$); and a severely affected group (class 4) with 13% of the sample ($n = 104$).

Therefore, by analyzing the sibling pairs consisting of individuals in either class 3 or 4, genes for a more severe type of dependence can be localized. Multipoint linkage analysis provided significant evidence of linkage to a group of markers on chromosome 16, with a maximum lod score of 4.0 near the marker D16S675. The consistent linkage findings on chromosome 16 in the COGA sample using the latent class phenotype and the ICD-10 definition of alcoholism may be attributable to the overlap in the two 'disease' definitions used in the analyses, since 88% of the individuals in latent classes 3 and 4 also fulfilled ICD-10 criteria for alcoholism. More recent analyses on the denser map yield the lod score of 3.2 given in **Table 1**.

Candidate Gene Analyses

A complementary approach to linkage analyses is to evaluate specific candidate genes. As noted in the COGA linkage publications, the signals are quite broad (i.e. they represent genomic regions of 10–30 cM), and the identification of a susceptibility locus is likely to be difficult. We have evaluated two candidate genes suggested in the alcohol literature.

Dopamine receptor D2 (*DRD2*)

The possible association of alleles at the *DRD2* locus, and in particular the *TaqI-A1* allele, with alcoholism

Table 1 Summary of COGA linkage results

		Initial sample		Replication sample		Combined sample	
Chromosome	Phenotype[a]	Lod[b]	cM	Lod	cM	Lod	cM
1	COGA DX or depression	5.1	120	[1.5]		4.7	122
1	COGA DX	2.5	146	[0.7]		2.6	144
2	MAO activity	2.04	131	2.43	110	2.85	126
2	ERP (P300, O2)	3.3	238				
2	COGA DX	3.0	99	[0]		[1.9]	
3	COGA DX	[0.1]		3.4	98	2.4	101
3	ERP (01-P3 unprimed)	3.1	182				
4	EEG (beta)					5.01	56
4	ERP (01-P3 primed)	3.3	219				
4	'max drinks'	2.2	124	[1.5]		3.5	121
5	ERP (T7-P3 unprimed)	3.6	104				
5	ERP (C3-P3 unprimed)	3.5	104				
5	ERP (P300, T8)	2.1	104				
6	ERP (P300, CZ)	3.4	154				
7	COGA DX	2.0	91	[0.2]		2.1	105
7	COGA DX	[1.6]		[1.3]		2.9	129
8	Harm avoidance	3.2	pter				
9	MAO activity	[1.50]		2.40	122	3.27	115
13	ERP (P300, T8)	2.1	49				
16	LCA	3.2	5.2				
21	SRE					4.0	80

[a]'max drinks': maximum number of drinks in a 24-h period; ERP: event-related potential; LCA: latent class analysis; MOA: monoamine oxidase; SRE: self-rating of the effects of alcohol; COGA DX or depression: diagnosis of alcohol dependence or depression; EEG (beta): electroencephalogram (12–29 Hz frequency).
[b]For comparison, lod scores in other samples below 2.0 are given in square brackets if analyzed within that sample.

remains controversial. Numerous studies have been reported in different populations, with both positive and negative results (reviewed in Edenberg *et al.* (1998a)). Populations are known to differ in allele frequencies at this locus, raising a serious caution about potential false positive results because of problems in matching cases and controls. Family-based association studies provide a method that avoids such problems. The possible associations of both a microsatellite marker in intron 2 and the *TaqA* alleles at the *DRD2* locus with alcohol dependence were tested using two family-based association methods: the transmission/disequilibrium test (TDT) and the affected family-based controls (AFBAC) test. The data provide no evidence of linkage or association between the *DRD2* locus and alcohol dependence (Edenberg *et al.*, 1998a). There was also no evidence of linkage to this region in the genome survey conducted by COGA.

Although there was an earlier study suggesting an association between *DRD2* and smoking, a larger, family-based analysis of the COGA data provided no evidence for any association with smoking or with comorbid smoking and alcohol dependence (Bierut *et al.*, 2000).

Solute carrier family 6 (neurotransmitter transporter, serotonin), member 4 (*SLC6A4*)

A population association between a regulatory variation in the promoter of *SLC6A4* and severe alcohol dependence has been reported. This potential association was analyzed in the COGA families. Analyses focused on individuals defined as alcohol dependent by criteria from ICD-10, and on subsets of these individuals reporting withdrawal-related symptoms. Application of the TDT did not provide support for either linkage or association between this functional polymorphism and alcohol dependence; there was no significant bias in the transmission of either allele to the alcohol-dependent offspring (Edenberg *et al.*, 1998b).

Linkage Analyses of Alcohol-related Endophenotypes

One of the strengths of the COGA project's design is that it has included measurement and assessment not only of diagnoses of alcohol dependence (via the

SSAGA, a polydiagnostic instrument) but also of related phenotypes (endophenotypes), including potential biological markers and comorbid disorders. Additional genome-wide linkage screens have been carried out using these related phenotypes.

Maximum number of drinks

Saccone *et al.* (2000) examined the maximum number of drinks consumed in a 24-h period ('max drinks'), as reported by individuals as part of the SSAGA. This measure proved to be closely related to a diagnosis of alcoholism in the COGA dataset: strictly defined unaffected individuals report low levels of consumption, while among the high-consumption classes, affected individuals predominate. One advantage of using 'max drinks' rather than diagnosis is that it provides a quantitative measure to grade all individuals who have ever drunk alcohol, leading to a larger and more powerful sample than affecteds only or strict unaffecteds only. Prior to linkage analyses, the maximum drinks score was log-transformed to reduce skewness due to individuals who report extremely high consumption. Multipoint sib-pair linkage analyses of the initial sample, the replication sample and the combined sample led to consistent linkage signals on chromosome 4 in the region of the alcohol dehydrogenase gene cluster, with a maximal lod score in the combined sample of 3.5 near D4S2407. The two-allele system at alcohol dehydrogenase 1C (class I), gamma polypeptide (*ADH1C*) has been genotyped in the COGA sample. However, the maximum drinks phenotype does not significantly vary by *ADH1C* genotype in the COGA data, indicating that alcohol dehydrogenase subunits other than *ADH1C* may be involved. The linkage to the 'max drinks' phenotype on chromosome 4 is consistent with prior linkage results from the analysis of strictly defined unaffecteds in COGA (Reich *et al.*, 1998), and with two-point linkage signals for alcohol dependence in a population of Native Americans (Long *et al.*, 1998).

Monoamine oxidase activity

Genome-wide linkage screens have been performed for platelet monoamine oxidase (MAO) activity (Saccone *et al.*, 1999, 2002). In the past, MAO activity has been cited as a possible biological marker for alcohol dependence; however, more recent work has indicated that the associations between low MAO activity and alcohol dependence are likely to be due to direct effects of cigarette smoking on MAO activity. Gender and smoking have significant effects on platelet MAO activity in the COGA data (Daw *et al.*, 2001); hence, linear regression was used to correct the MAO activity values for gender, smoking status, and the COGA center at which the subject's blood samples were drawn. Linkage analysis (Haseman–Elston regression using sib pairs) of the initial COGA sample indicated modest evidence of linkage to regions of chromosome 2 and chromosome 6 (Saccone *et al.*, 1999). Follow-up analyses (Saccone *et al.*, 2002) have now applied both Haseman–Elston sib-pair methods and full-pedigree variance components methods to corrected MAO activity in the initial sample, the replication sample, and the combined sample. Regions on chromosomes 2, 9 and 12 indicated consistent evidence for linkage across the two distinct datasets by at least one analysis method. In particular, the strongest evidence appeared on chromosome 9, with a lod score in the combined sample of 3.3 at 115 cM near D9S261 from the variance components analysis. This signal is also supported by the second highest lod score (1.62) obtained by sib-pair regression with independent pairs. On chromosome 2, sib-pair regression gave a lod of 2.85 in the combined data near D2S436; this finding is weakly supported by variance components analysis.

Neurophysiology

COGA has focused a great deal of attention on the collection of ERPs in order to improve the power to detect quantitative trait loci (QTLs) influencing susceptibility to alcoholism. ERPs are averages of the electroencephalographic activity recorded from the scalp with noninvasive electrodes following a stimulus event (e.g. light, sound). These electrodes record the electrical activity of the brain on a millisecond basis while the individual responds to the stimuli. Although these recordings are made at the scalp, they record overlapping activities emanating from various brain circuits along pathways from sensory reception to higher cognitive processes.

A great deal of attention has been focused on the P3 component of the ERP, a positive-going voltage change of scalp-recorded electroencephalographic activity that occurs between 300 ms and 500 ms after stimulus onset. It is elicited when a stimulus is perceived, memory operations are engaged, and additional resources are allocated toward its processing. It has been repeatedly observed that the P3 component is of significantly smaller amplitude in abstinent alcoholics as well as in the offspring of alcoholics (see Porjesz and Begleiter, 1998). Results suggest that rather than being a consequence of years of heavy drinking, low P3 amplitudes antecede the development of alcoholism, and that P3 amplitude has utility as a potential phenotypic marker for alcoholism (Porjesz *et al.*, 1998).

Linkage analysis has been performed on the age-regressed visual P3 (VP3) target (607 individuals in 103 families) data set with SOLAR (sequential oligogenic linkage analysis routines), a multivariate,

multipoint quantitative linkage package using variance components. This analysis found significant linkage for the P3 amplitude at the O2 electrode on chromosomes 2 (D2S434, 218 cM) (lod = 3.28; $p < 0.0299$) and 6 (D6S495, 213 cM) (lod = 3.41; $p < 0.0219$) for the Cz electrode (Begleiter *et al.*, 1998). Lod scores greater than 2.0, suggestive of linkage, were also found for the T8 electrode on chromosomes 5 (D5S1501, 76 cM) and 13 (D13S321, 45 cM) (Begleiter *et al.*, 1998). Recent findings on a larger sample of subjects from the COGA project have corroborated the initial findings on chromosomes 2, 5, 6 and 13.

Given the strong evidence for several QTLs influencing VP3 amplitude, we wished to determine whether these loci also influence the risk of alcoholism. Bivariate quantitative genetic and linkage analyses on these traits were performed with SOLAR, which allows for joint consideration of both the disease and quantitative precursors/correlates in pedigrees of arbitrary size and complexity (Williams *et al.*, 1999). For the qualitative disease outcome, a continuous underlying liability distribution is assumed from which disease is determined by a threshold process. This procedure can assess whether correlations between P3 amplitude and alcoholism stem from shared genetic influences. We performed bivariate linkage analysis of the genome screen data using three alcoholism diagnoses (ICD-10, DSM-IV and COGA) with P3 amplitude at Cz. The pattern of results was similar between diagnoses, but the strongest evidence for linkage was obtained with DSM-IV. Joint consideration of the DSM-IV diagnosis of alcoholism and the amplitude of the P3 component of the Cz ERP significantly increased the evidence for linkage of these traits to a chromosome 4 region near the alcohol dehydrogenase gene cluster. A likelihood-ratio test for complete pleiotropy was significant, suggesting that the same QTL influences both risk of alcoholism and the amplitude of the P3 component.

A semantic priming paradigm was used to elicit the N4 component, a negative component occurring approximately 400–600 ms after an incongruent (unprimed) word among contextually related (primed) words. While N4 is obtained to unprimed but not primed words in normal subjects, alcoholics manifest N4's to both primed and unprimed words. A genome-wide linkage screen was performed on the amplitude of the N4 and P3 components of the ERP, measured at 19 scalp locations in response to a semantic priming task for 604 individuals in 100 pedigrees ascertained as part of COGA (Almasy *et al.*, 2001). N4 and P3 amplitudes in response to three stimuli (nonwords, primed words (i.e. antonyms) and unprimed words) all showed significant heritability estimates, the highest being 0.54. Both N4 and P3 showed significant genetic correlations across stimulus type at a given lead and across leads within a stimulus, indicating shared genetic influences among the traits. There were also substantial genetic correlations between the N4 and P3 amplitudes for a given lead, even across stimulus type. N4 amplitudes showed suggestive evidence of linkage in several chromosomal regions, and P3 amplitudes showed significant evidence of linkage to chromosome 5 (D5S1501, 90 cM) and suggestive evidence of linkage to chromosome 4 (D4S2374, 172 cM).

A recent analysis found significant linkage (lod = 5.01) and linkage disequilibrium between the beta frequency of the electroencephalogram (EEG) and the γ-aminobutyric acid A receptor gene cluster ($GABA_A$ receptor-associated protein(s)) on chromosome 4 (Porjesz *et al.*, 2002).

Other phenotypes

In addition, linkage analyses have been reported for dimensions of personality (Cloninger *et al.*, 1998), the Self-Rating of the Effects of Alcohol (SRE) questionnaire (Schuckit *et al.*, 2001), and a phenotype consisting of a diagnosis of alcoholism or depression (Nurnberger *et al.*, 2001). A very strong linkage finding for the broad phenotype alcoholism or depression was found in the same region of chromosome 1 as the linkage for alcoholism. The findings are indicated in **Table 1**.

Summary

Linkage analyses of the clinical phenotype of alcohol dependence suggest heterogeneity with no clearcut interpretation. Although promising, many of the lod scores could be false positive results. The initial linkage findings on chromosome 4 (unfortunately there were not enough unaffected individuals genotyped in the replication sample), together with the results for the maximum number of drinks in a 24-h period, suggest a protective locus near the alcohol dehydrogenase complex. The high lod score for the phenotype consisting of alcohol dependence or depression also merits further investigation.

An alternative strategy to analysis of the clinical phenotype is the analysis of related quantitative traits (endophenotypes) to identify susceptibility genes. Our data suggest that the low MAO activity in alcoholics, reported in many studies, is a direct result of cigarette smoking and thus a spurious association. Results for the ERP (and EEG) remain promising and offer an important strategy for understanding the genetics of alcohol dependence.

Acknowledgments

The Collaborative Study on the Genetics of Alcoholism (COGA) (H. Begleiter, SUNY HSCB, Principal Investigator; T. Reich, Washington University, Co-Principal Investigator) includes nine different centers where data collection, analysis and/or storage takes place. The nine sites and Principal Investigators and Co-Investigators are: Indiana University (T.-K. Li, J. Nurnberger Jr, P. M. Conneally, H. J. Edenberg); University of Iowa (R. Crowe, S. Kuperman); University of California at San Diego and Scripps Institute (M. Schuckit); University of Connecticut (V. Hesselbrock); State University of New York, Health Sciences Center at Brooklyn (B. Porjesz, H. Begleiter); Washington University in St Louis (T. Reich, C. R. Cloninger, J. Rice, A. Goate); Howard University (R. Taylor); Rutgers University (J. Tischfield); and Southwest Foundation (L. Almasy). This national collaborative study was supported by NIH Grant U10AA08403 from the National Institute on Alcohol Abuse and Alcoholism (NIAAA). This research was also supported by AA00285 (TF) and MH37685.

See also

Alcoholism and Drug Addictions

References

Almasy L, Porjesz B, Blangero J, *et al.* (2001) Genetics of event-related brain potentials in response to a semantic priming paradigm in families with a history of alcoholism. *American Journal of Human Genetics* **68**: 128–135.

Begleiter H, Porjesz B, Reich T, *et al.* (1998) Quantitative trait loci analysis of human event-related brain potentials: P3 voltage. *Electroencephalography and Clinical Neurophysiology* **108**: 244–250.

Bierut LJ, Rice J, Edenberg HJ, *et al.* (2000) A family-based study of the association of the dopamine D2 receptor gene (DRD2) with habitual smoking. *American Journal of Medical Genetics* **90**: 299–302.

Cloninger CR, Van Eerdewegh P, Goate A, *et al.* (1998) Anxiety proneness linked to epistatic loci in a genomic scan of human personality traits. *American Journal of Medical Genetics (Neuropsychiatric Genetics)* **81**: 313–317.

Daw EW, Rice JP, Anthenelli RM, *et al.* (2001) A bootstrapped commingling analysis of platelet monoamine oxidase activity levels corrected for cigarette smoking. *Psychiatric Genetics* **11**: 177–185.

Edenberg HJ, Foroud T, Koller DL, *et al.* (1998a) A family-based analysis of the association of the dopamine D2 receptor (*DRD2*) with alcoholism. *Alcoholism: Clinical and Experimental Research* **22**: 505–512.

Edenberg HJ, Reynolds J, Koller DL, *et al.* (1998b) A family-based analysis of whether the functional promoter alleles of the serotonin transporter gene HTT affect the risk for alcohol dependence. *Alcoholism: Clinical and Experimental Research* **22**: 1080–1085.

Foroud T, Bucholz KK, Edenberg HJ, *et al.* (1998) Linkage of an alcoholism-related severity phenotype to chromosome 16. *Alcoholism: Clinical and Experimental Research* **22**: 2035–2042.

Foroud T, Edenberg HJ, Goate A, *et al.* (2000) Alcoholism susceptibility loci: confirmation studies in a replicate sample and further mapping. *Alcoholism: Clinical and Experimental Research* **24**(7): 933–945.

Long JC, Knowler WC, Hanson RL, *et al.* (1998) Evidence for genetic linkage to alcohol dependence on chromosomes 4 and 11 from an autosome-wide scan in an American Indian population. *American Journal of Medical Genetics (Neuropsychiatric Genetics)* **81**: 216–221.

Nurnberger Jr JI , Foroud T, Flury L, *et al.* (2001) Evidence for a locus on chromosome 1 that influences vulnerability to alcoholism and affective disorder. *American Journal of Psychiatry* **158**: 718–724

Porjesz B, Almasy L, Edenberg HJ, *et al.* (2002) Linkage disequilibrium between the beta frequency of the human EEG and a GABAA receptor gene locus. *Proceedings of the National Academy of Sciences of the United States of America* **99**: 3729–3733.

Porjesz B and Begleiter H (1998) Genetic basis of the event-related potentials and their relationship to alcoholism and alcohol use. *Journal of Clinical Neurophysiology* **15**(1): 44–57.

Porjesz B, Begleiter H, Reich T, *et al.* (1998) Amplitude of visual P3 event-related potential as a phenotypic marker for a predisposition to alcoholism: preliminary results from the COGA project. *Alcoholism: Clinical and Experimental Research* **22**: 1317–1323.

Reich T, Edenberg HJ, Goate A, *et al.* (1998) Genome-wide search for genes affecting the risk for alcohol dependence. *American Journal of Medical Genetics (Neuropsychiatric Genetics)* **81**: 207–215.

Saccone NL, Kwon JM, Corbett J, *et al.* (2000) A genome screen of maximum number of drinks as an alcoholism phenotype. *American Journal of Medical Genetics (Neuropsychiatric Genetics)* **96**: 632–637.

Saccone NL, Rice JP, Rochberg N, *et al.* (1999) Genome screen for platelet monoamine oxidase (MAO) activity. *American Journal of Medical Genetics* **88**: 517–521.

Saccone NL, Rice JP, Rochberg N, *et al.* (2002) Linkage for platelet monoamine oxidase (MAO) activity: results from a replication sample. *Alcoholism: Clinical and Experimental Research* **26**(5): 603–609.

Schuckit MA, Edenberg HJ, Kalmijn J, *et al.* (2001) Genome-wide search for genes that relate to a low level of response to alcohol. *Alcoholism: Clinical and Experimental Research* **25**: 323–329.

Williams JT, Begleiter H, Porjesz B, *et al.* (1999) Joint multipoint linkage analysis of multivariate qualitative and quantitative traits. II. Alcoholism and event related potentials. *American Journal of Human Genetics* **65**: 1148–1160.

Agarwal DP (2002) Alcoholism. In: King RA, Rotter JI and Motulsky AG (eds.) *The Genetics Basis of Common Diseases*, pp. 876–913. New York, NY: Oxford University Press.

McGuffin P, Owen MJ and Gottesman II (2002) *Psychiatric Genetics and Genomics*. New York, NY: Oxford University Press.

Web Links

Dopamine receptor D2 (*DRD2*); Locus ID: 1813. LocusLink: http://www.ncbi.nlm.nih.gov/LocusLink/LocRpt.cgi?l=1813

Solute carrier family 6 (neurotransmitter transporter, serotonin), member 4 (*SLC6A4*); Locus ID: 6532. LocusLink: http://www.ncbi.nlm.nih.gov/LocusLink/LocRpt.cgi?l=6532

Alcohol dehydrogenase 1C (class I), gamma polypeptide (*ADH1C*); Locus ID 126. LocusLink: http://www.ncbi.nlm.nih.gov/LocusLink/LocRpt.cgi?l=126
Dopamine receptor D2 (*DRD2*); MIM number: 126450. OMIM: http://www.ncbi.nlm.nih.gov/htbin-post/Omim/dispmim?126450
Solute carrier family 6 (neurotransmitter transporter, serotonin), member 4 (*SLC6A4*); MIM number: 182138. OMIM: http://www.ncbi.nlm.nih.gov/htbin-post/Omim/dispmim?182138
Alcohol dehydrogenase 1C (class I), gamma polypeptide (*ADH1C*); MIM number: 103730. OMIM: http://www3.ncbi.nlm.nih.gov/htbin-post/Omim/dispmim?103730

Alignment: Statistical Significance

Richard Mott, *University of Oxford, Oxford, UK*

Intermediate article

Computation of the probability that an observed alignment between two protein or DNA sequences could have arisen by chance is useful for identifying genuinely related sequences.

Article contents

Introduction

The availability of large volumes of deoxyribonucleic acid (DNA) and protein data has inevitably led to the problem of how one should determine which sequences are genuinely related, that is they share a common ancestral sequence or a common structure or function, and which are similar by chance. Because sequence alignments are quantified by a score that reflects the degree of similarity between the sequences, the problem can be thought of as how to determine a threshold value, T, such that if two sequences' alignment score is below the threshold they are considered to be unrelated. The threshold T depends on the probability distribution of alignment scores between random, unrelated sequences, and is influenced by several factors:

1. Sequence alignment method. The Smith–Waterman algorithm is the most sensitive method for comparing sequences, and we will not consider other algorithms here; this restriction is reasonable because most popular software for database searching, such as BLAST (see Web Links) and FASTA, are essentially sophisticated implementations of the Smith–Waterman algorithm that run quickly enough to permit the search of large databanks. Furthermore, the statistical behavior of many other biologically important scores, such as those of alignments between sequences and profiles, is qualitatively similar to Smith–Waterman scores. (***See*** BLAST Algorithm; FASTA Algorithm; Smith–Waterman Algorithm.)

2. Scoring scheme, that is, the substitution matrix and gap penalty. A substitution matrix specifies the score for aligning a given pair of letters: substitutions that are favored have positive scores while unlikely ones are negative. In this article we will assume that on average the substitution score is negative. Changing the scoring scheme has a marked effect on the score distribution between random unrelated sequences, and hence T. (***See*** Substitution Matrices.)

3. Sequence length. Long sequences have a higher score thresholds than shorter ones because the expected similarity score between two random sequences increases with sequence length. In fact, if the sequence lengths are m, n, then the expected score is proportional to $\log(mn)$ approximately. We are usually interested in assigning significance in the context of a databank search, in which case the total length of the databank is a critical factor.

4. Sequence compositions. Sequences with similar letter frequencies will tend to have higher alignment scores, even when they are unrelated. For example the amino acid cysteine is usually rare, and cysteine–cysteine matches are assigned large positive scores in most substitution matrices. Consequently, cysteine-rich protein sequences tend to have unusually high alignment scores, and will be interpreted as being related unless these unusual amino-acid frequencies are taken into account.

The observed sequence alignment score between two real sequences must be compared to the distribution of scores observed in comparisons, using the same algorithm and scoring scheme, between sequences sampled from populations of random unrelated sequences, tailored to take account of the lengths and compositions of the actual sequences being compared.