Genetic Epidemiology 17(Suppl 1): S325–S330 (1999)

Structural Decomposition of Genetic Diversity in Families with Alcoholism

Hans H. Stassen, Henri Begleiter, Bernice Porjesz, John Rice, Christian Scharfetter, and Theodore Reich

Research Department (H.H.S., C.S.), Psychiatric University Hospital, Zurich, Switzerland; and Neurodynamics Laboratory (H.B., B.P.), SUNY Health Science Center, Brooklyn, New York; and School of Medicine (T.R., J.R.), Washington University, St. Louis, Missouri

Using genotypes of 280 marker loci on the 22 autosomes of 105 alcohol-dependent probands, their affected and unaffected sibs, as well as their parents, we iteratively constructed a genetic similarity function that enabled us to quantify the interindividual genetic distances $d(x_i, x_j)$ between feature vectors x_i, x_j made up by the allelic patterns of individuals i, j with respect to loci $l_1, l_2, ..., l_n$. Based on this similarity function, we investigated the sib-sib similarities that are expected to deviate from "0.5" in affected sib pairs if the region of interest contains markers close to disease-causing genes. The reference value "0.5" was derived from the parents-offspring similarities, because these are independent of the affection status. The question of population admixture was addressed by means of multivariate structural analyses. These analyses led to four "natural" groups whose validity was tested through the father-mother similarities. Additionally, we determined the eigenvectors that optimally represented the genetic variation and found several marker configurations on chromosomes 1, 3, 7, 15, and 17 that reproducibly discriminated ($p \leq 0.01$) affected probands/sibs from unaffected sibs, while no such differences were found between affected probands and affected sibs.

Key words: alcohol dependence, molecular genetics, population admixture

INTRODUCTION

In the case of genetically complex disorders, like alcohol dependence, the standard phenotype-to-genotype research strategy may not readily lead to the detection of "signals" if the contributions of single loci are small, and if there exist significant interactions

Address reprint requests to Dr. H.H. Stassen, Psychiatric University Hospital, Research Department, P.O. Box 68, 8029 Zurich, Switzerland.

between loci. We therefore propose a genotype-to-phenotype strategy that has its main focus on oligogenic, interacting models and evaluates the within-family similarities of high-dimensional genetic feature vectors with respect to deviations from expected values. The multidimensional variation ("genetic diversity") inherent in a given set of genetic feature vectors allows one to directly assess the genetic heterogeneity that is caused by population admixture ("ethnicity"). Thus, standard multivariate methods, such as principal component analysis, cluster analysis and metric/nonmetric multidimensional scaling, can be applied to structurally decompose a sample into genetically more homogeneous subgroups. In what follows, we will demonstrate how the notion of genetic diversity is quantifiable by means of a genetic similarity function that simultaneously evaluates the allelic information available at several genetic loci, how validity and performance of a genetic similarity function can be tested empirically, and how this similarity function can be used to analyse the structural properties of a population's ethnic background.

Specifically, we tested the following hypotheses: (1) There exists a genetic similarity function that reproducibly assesses the genetic similarity of "0.5" between first-degree relatives for all autosomes, and that discriminates the distribution of the corresponding similarity coefficients from that of the genetic similarities between unrelated individuals; (2) There exist significantly different, reproducibly assessable, "natural" ethnic subgroups; (3) There exist marker configurations for which the between-sib genetic similarity deviates in affected sib pairs significantly from the parents-offspring genetic similarity which is always "0.5" irrespective of the affection status in parents and offspring.

METHODS

Genetic Similarity

Central to the genotype-to-phenotype approach is the similarity function that allows one to quantify the genetic distances $d(x_i,x_j)$ between high-dimensional feature vectors x_i, x_j made up by the allelic patterns of individuals i, j with respect to loci $l_1, l_2, ..., l_n$. We rely on a set-theoretical similarity function $s(x_i,x_j)$ that has been designed to assess the current genetic distances between individuals rather than to model genetic distance in terms of evolutionary history [Goldstein et al., 1995; Zhivotovsky and Feldman, 1995; Kimmel et al., 1996; Di Rienzo et al., 1998]. The specific properties of this similarity function are given elsewhere [Tversky, 1977; Stassen, 1985]. It is defined as:

$$s(\vec{x}_i,\vec{x}_j)=\frac{\sum_k w_k\left[X_{ik}\cap X_{jk}\right]}{\sum_k w_k\left[X_{ik}\cup X_{jk}\right]}$$

where w_k designates the weight of the feature vector's k-th component, and $X_{.k}$ the area spanned by the 2 alleles A_{k1}, A_{k2} of the k-th component. Its performance can easily be verified for a given set of weights through a computerized recognition of person (CRP) test on the basis of a sufficiently large and representative sample of unrelated individuals. For appropriately chosen genetic feature vectors the rates of false-positive and false-

or the within-pair similarity of sibs are found to be "0.5" after standardization. The similarity function may be optimized by incorporating the allele frequencies of the population as weights, because concordance in a frequent allele may have less weight than concordance in a rare allele. The "minimal interindividual" together with the "maximal intra-individual" similarity is a well-suited optimization criterion under the assumption of genotype errors ≤ p%, where p implicitely defines the "noise" level. To circumvent the problem of local maxima during optimization, independent "learn" and "test" samples are recommended. Once the similarity function has been constructed, it must be calibrated on the basis of (1) the distribution of parents-offspring similarities ("0.5"), (2) the distribution of sib-sib similarities which are expected to deviate from "0.5" in affected sib pairs if the region of interest contains markers close to disease-causing or protecting genes, and (3) the distribution of interindividual similarities of unrelated individuals ("0").

Statistical Power

The statistical power to detect deviations from the genetic similarity "0.5" in affected sib pairs depends on (1) the number of families, (2) the number of loci, (3) the number and frequencies of alleles at each locus, and (4) the number of trait loci. We conducted a power analysis by means of computer simulations on the basis of 60, 100, and 200 families with two affected and two unaffected offspring. The number of loci varied from 20 to 30 with an average number of four to10 alleles, whereby five randomly selected loci were chosen as "affected" in terms of a 10% increase in concordance. Specifically, we determined the distributions of genetic similarities with respect to between-subject comparisons of unrelated individuals, parents-offspring comparisons, and between-sib comparisons of affected sib pairs. The respective means and standard deviations suggested a statistical power > 90% ($\alpha = 0.01$) to detect deviations from the genetic similarity "0.5" in affected sib pairs.

Data

Using genotypes of 280 marker loci on the 22 autosomes of 105 alcohol-dependent probands, their affected and unaffected sibs, as well as their parents, we reconstructed missing alleles where possible, or replaced them randomly from the "family pool" or, as the least favorable option, with respect to population frequencies. The marker data were then combined into feature vectors in a chromosome-wise manner for all 22 autosomes, thereby excluding the X-chromosome because of the methodological problems that arise from the sex-specific differences in allelic information. The affection status was determined on the basis of DSM-III-R and definite Feighner criteria. A detailed description of the data material is given elsewhere [Begleiter et al., 1995; Reich et al., 1998a].

RESULTS

When probands, their sibs, and parents were combined into one sample with missing alleles treated as described in the previous paragraph, the CRP test (based on systematic inter- and intra-individual comparisons) yielded false positive/false negative classification errors < 2% across the 22 autosomes. Under the assumption of a 10% uncertainty in the genotypes, the rates of the two classification errors increased, but remained < 5%. A further validation of the similarity function was achieved by comparing the parents-offspring similarities (n = 860) with the corresponding interindividual similarities

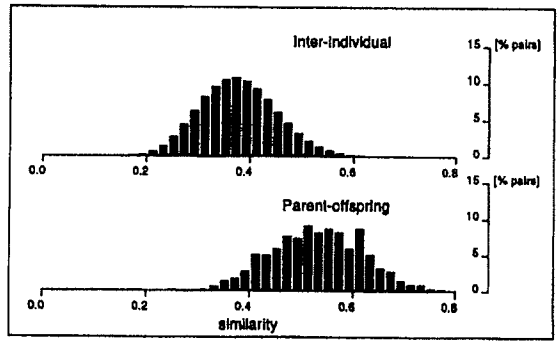

Fig. 1. Interindividual genetic similarities derived from systematic comparisons between unrelated individuals (n = 626) versus parents-offspring similarities (n = 860). The underlying feature vector comprised 20 markers on chromosome 1.

computed from all possible combinations of unrelated individuals (n = 626). The respective distributions of similarity coefficients derived from chromosome 1 were found to be approximately normal (Figure 1), and to exhibit highly significant differences ($p < 0.0001$). This picture of normally distributed similarity coefficients and significant differences between the two distributions under comparison was essentially the same for all autosomes, except for chromosome 5, where a bimodal distribution of the parents-offspring similarities indicated insufficient genotype data.

In the next step, we focused our interest on the question of ethnic heterogeneity and determined all chromosomes for which the father-mother similarity was significantly higher than the corresponding interindividual similarity between unrelated individuals. We found 10 chromosomes with significant differences ($p < 0.05$). This finding suggested that there existed distinct ethnic groups within the sample, and that marriages occurred in

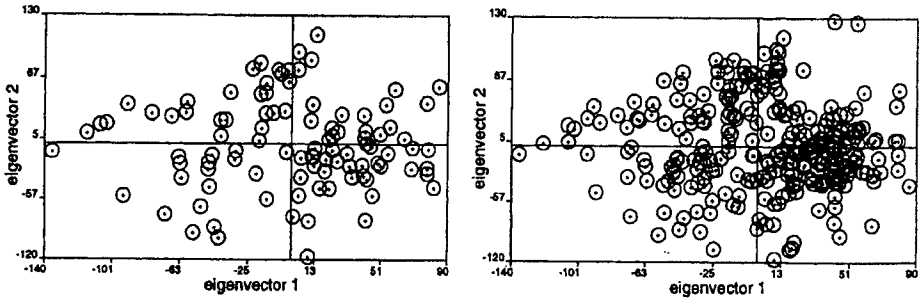

Fig. 2. Projection of the genotype feature vectors of 105 alcohol dependent probands (left panel) and 268 alcohol dependend sibs (right panel) onto the plane defined by the 2 "largest" eigenvectors of the proband sample. The scales of the two axes measure genetic distance and are based on arbitrary units.

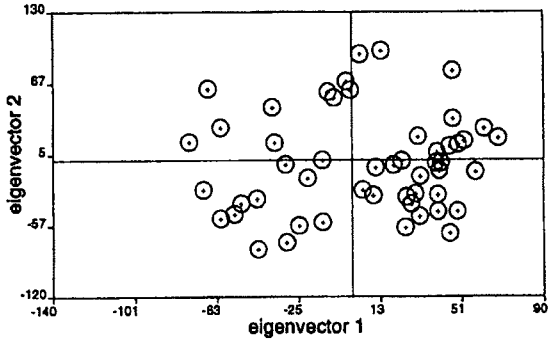

Fig. 3. Projection of the genotype feature vectors of 51 unaffected sibs onto the plane defined by the two eigenvectors of the proband sample. The axes are scaled as in Figure 2.

their majority within these groups. Subsequent cluster analyses led to a partitioning of the total sample into four "natural" ethnic subgroups whose validity was supported by the fact that the father-mother similarities, computed separately for each subgroup, did not exhibit deviations from the corresponding interindividual values. On the other hand, the accordance between "natural" and "reported" ethnicity, as stated by the probands, turned out to be low.

The affected sib-pair analysis yielded several "signals," i.e., marker configurations for which the between-sib similarity in affected sib pairs deviated significantly ($p \leq 0.01$) from the reference value "0.5." These signals were verified by means of a principal component analysis that led to a representation of the original feature vectors through a set of orthogonal eigenvectors (uncorrelated coordinates). We found that (1) the first three to four "largest" eigenvectors typically explained 60% of the genetic variation associated with the 8-20 markers of each autosome, and (2) the marker configurations on chromosomes 1, 3, 7, 15, and 17 reproducibly discriminated ($p \leq 0.01$) affected probands/sibs from unaffected sibs, while no such differences were found between affected probands and affected sibs. The statistical comparisons were based on the means and standard deviations of the individuals' components on the two "largest" eigenvectors (42-58% of the genetic variation). Differences were regarded as "reproducible" if they reached significance in the two "affected-unaffected" comparisons. Examination of the variable weights, as provided by the principal component analysis, revealed that typically one to two markers contributed > 60% of the discrimination, thus suggesting that the respective markers might be not too far from disease-causing genes. However, our results must be regarded as preliminary, because (1) the average missing data rate per family was > 10% of genotypes, within individuals as well as within markers, (2) the wide-meshed marker grid with an average intermarker distance of 13 cM was not optimal for signal detection (fine-meshed grids with intermarker distances ≤ 4 cM would be desirable), (3) the sample of unaffected sibs ($n = 51$) was too small to test the reproducibility of our findings by means of random splitting techniques, and (4) ethnicity-specific signal detection remains to be carried out.

Figure 2 gives a visual impression of the genetic diversity as assessed through 20 markers on chromosome 1. The left panel shows the relative positions of the 105 alcohol dependent probands in the two-dimensional plane spanned up by the two "largest" eigenvectors of the proband sample, and the right panel shows the 268 alcohol dependent

sibs with respect to the same eigenvectors. The genetic structure, as reflected by the interindividual genetic distances, is obviously reproducible across the two populations of affected probands and affected sibs. Yet interestingly, the unaffected sibs displayed a somewhat reduced genetic variability on these two eigenvectors, which could in part be due to there being fewer observations (Figure 3).

DISCUSSION

Using a similarity approach to modeling genetic diversity, we investigated the problem of signal detection in the case of complex disorders, where single loci are, by themselves, neither necessary nor sufficient for developing the phenotype, and where genetic and environmental factors underlying the same phenotype may vary among ethnically different subgroups. Our results suggested that the genotype-to-phenotype strategy may become a valuable extension of the standard phenotype-to-genotype approaches, although the continued evidence for linkage on chromosomes 4 and 16 [Reich et al., 1998b; Foroud et al., 1998] could not yet be replicated by the method. However, the new COGA dense-map data and ethnicity-specific signal detection may elucidate these inconsistencies.

ACKNOWLEDGMENTS

This work was supported in part by the Swiss National Science Foundation grant number SNF 32-46782.96.

REFERENCES

Begleiter H, Reich T, Hesselbrock V, Porjesz B, Li TK, Schuckit MA, Edenberg HJ, Rice JP (1995): The Collaborative Study on the Genetics of Alcoholism. The genetics of alcoholism. Alcohol Health Res World 19:228-236.

Di Rienzo A, Donnelly P, Toomajian C, Sisk B, Hill A, Petzl-Erler ML, Haines GK, Barch DH (1998): Heterogeneity of microsatellite mutations within and between loci, and implications for human demographic histories. Genetics 148:1269-1284.

Foroud T, Bucholz KK, Edenberg HJ, Goate A, Neuman RJ, Porjesz B, Koller DL, Rice J, Reich T, Bierut LJ, Cloninger CR, Nurnberger JI Jr, Li TK, Conneally PM, Tischfield JA, Crowe R, Hesselbrock V, Schuckit M, Begleiter H (1998): Linkage of an alcoholism-related severity phenotype to chromosome 16. Am J Med Genet (Neuropsychiatr Genet) 81:479 (Abstract).

Goldstein DB, Ruiz Linares A, Cavalli-Sforza LL, Feldman LW (1995): An evaluation of genetic distances for use with microsatellite loci. Genetics 139:463-471.

Kimmel M, Chakraborty R, Stivers DN, Deka R (1996): Dynamics of repeat polymorphisms under a forward-backward mutation model: within- and between-population variability at microsatellite loci. Genetics 143:549-555.

Reich T, Begleiter H, Willig C, Crose C, Carr K, Shears S, Wu W, Cloninger CR, Crowe RR, Tischfield JA, Nurnberger JI Jr, Conneally PM, Li TK, Projesz B, Bucholz K, Schuckit MA, Hesselbrock V, Foroud T, Van Eerdewegh P, Rice JP, Williams JT, Goate A, Edenberg HJ (1998a): Genome-wide search for genes affecting the risk for alcohol dependence. Am J Med Genet (Neuropsychiatr Genet) 81:207-215.

Reich T, Goate A, Edenberg H, Rice J, Foroud T, Hesselbrock V, Schuckit M, Porjesz B, Nurnberger JI Jr, Crowe R, Begleiter H (1998b): Replicating genetic linkage in the Collaborative Study on the Genetics of Alcoholism (COGA). Am J Med Genet (Neuropsychiatr Genet) 81:478.

Stassen HH (1985): The similarity approach to EEG analysis. Meth Inform Med 24:200-212.

Tversky A (1977): Features of similarity. Psychol Rev 84:327-352.

Zhivotovsky LA, Feldman MW (1995): Microsatellite variability and genetic distances. Proc Natl Acad Sci USA 92:11549-1152.