Human Molecular Genetics, 2000, Vol. 9, No. 19 **2895–2908**

Sequence variability and candidate gene analysis in complex disease: association of μ opioid receptor gene variation with substance dependence

Margret R. Hoehe[1,2,5,+], Karla Köpke[1,2], Birgit Wendel[1], Klaus Rohde[1], Christina Flachmeier[1], Kenneth K. Kidd[3], Wade H. Berrettini[4] and George M. Church[5]

[1]Genome Research, Max-Delbrück-Center for Molecular Medicine, Robert-Rössle-Strasse 10, D-13092 Berlin, [2]GenProfile AG, Robert-Rössle-Strasse 10, D-13125 Berlin, Germany, [3]Department of Genetics, Yale University School of Medicine, New Haven, CT 06520, USA, [4]Center for Neurobiology and Behavior, University of Pennsylvania, Philadelphia, PA 19104, USA and [5]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA

Received 14 August 2000; Revised and Accepted 27 September 2000

To analyze candidate genes and establish complex genotype–phenotype relationships against a background of high natural genome sequence variability, we have developed approaches to (i) compare candidate gene sequence information in multiple individuals; (ii) predict haplotypes from numerous variants; and (iii) classify haplotypes and identify specific sequence variants, or combinations of variants (pattern), associated with the phenotype. Using the human μ opioid receptor gene (*OPRM1*) as a model system, we have combined these approaches to test a potential role of *OPRM1* in substance (heroin/cocaine) dependence. All known functionally relevant regions of this prime candidate gene were analyzed by multiplex sequence comparison in 250 cases and controls; 43 variants were identified and 52 different haplotypes predicted in the subgroup of 172 African-Americans. These haplotypes were classified by similarity clustering into two functionally related categories, one of which was significantly more frequent in substance-dependent individuals. Common to this category was a characteristic pattern of sequence variants [–1793T→A, –1699Tins, –1320A→G, –111C→T, +17C→T (A6V)], which was associated with substance dependence. This study provides an example of approaches that have been successfully applied to the establishment of complex genotype–phenotype relationships in the presence of abundant DNA sequence variation.

INTRODUCTION

The analysis of candidate genes is a key step in strategies for disease gene identification. Candidate genes may, on the one hand, be identified based on functional information about the disease and/or on genetic map information obtained by linkage or linkage disequilibrium (1,2). On the other hand, whole genome candidate gene approaches have been proposed as the future in the genetics of complex disease (2–4). In addition, as the human genome reference sequence comes to completion, the analysis of genetic variation is becoming increasingly important (2,5). Those few studies that have systematically compared individual candidate gene sequences, either parts of a gene (6–9) or whole gene sequences (10,11), suggest that genes and the human genome may be much more variable than previously thought. Allelic complexity in candidate genes will be large and such complexity will make the analysis of genotype–phenotype relationships difficult, particularly in the situation of complex disease (12–16). The spectrum of polymorphic profiles may include any variant, or combinations of variants (patterns), that may interact to determine those functional variations that are involved in phenotypic variation. Since it is the entire gene and its encoded protein that act as the units of function which potentially affect a phenotype (and ultimately allow first conclusions on disease mechanisms), we must analyze the entire sequences of the individual genes including their regulatory and critical intronic regions. It is therefore essential in diploid organisms to determine the specific combinations of given gene sequence variants for each of the chromosomes defined here as haplotypes. Given the high level of sequence variation now evident, the number of haplotypes may be unfeasibly high for association studies (6,15). This represents a particular challenge in this approach to candidate gene analysis, leading to the following major questions: how to analyze genotype–phenotype relationships in the situation of abundant sequence variation; how to identify those variants, or combinations of variants, that are of importance for the phenotype, given that the functionally relevant variants represent only a subset of the naturally occurring sequence variation. We have developed approaches to (i) compare candidate gene sequence information in multiple individuals; (ii) predict haplotypes from numerous variants; and (iii) classify haplotypes into functionally related categories to allow identification of those specific sequence variants, or combinations of variants (pattern), associated with the disease phenotype. Using the

[+]To whom correspondence should be addressed at: Genome Research, Max-Delbrück-Center for Molecular Medicine, Robert-Rössle-Strasse 10, D-13092 Berlin, Germany. Tel: +49 30 9489 2167; Fax: +49 30 9489 2166; Email: mhoehe@mdc-berlin.de

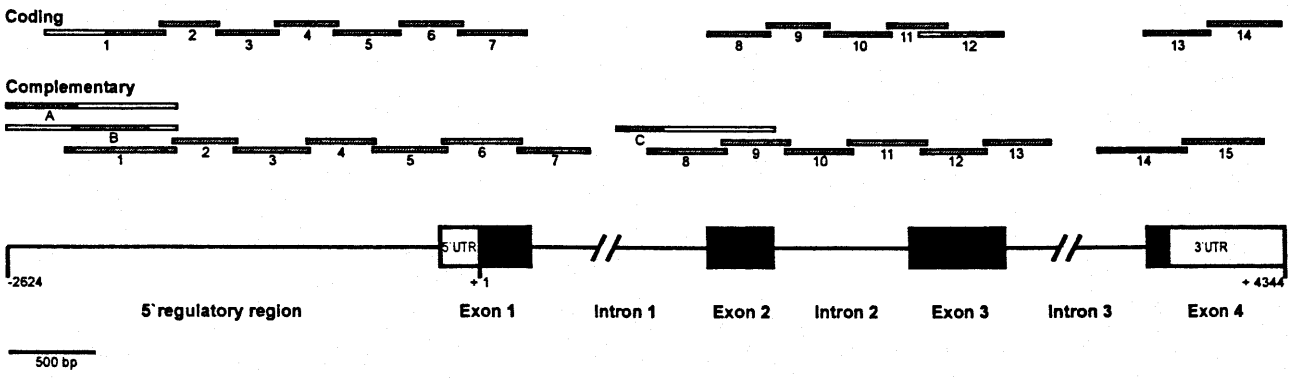

Figure 1. Dissection of *OPRM1* sequence into target DNA segments. The *OPRM1* genomic reference sequence (6968 bp), which is presented as baseline, integrates cloned 5′ regulatory, exonic and intronic sequences (24,64) (GenBank accession nos AJ000341, AJ000371–AJ000375; L25119). Base pair coordinates are given relative to the translation start site; sequences are drawn to scale. The dissection of the gene, its coding and complementary strands, into target DNA segments according to a 'staggered design' is shown at the top. DNA segments of each strand are numbered as they appear in the plex pool; all 14 DNA segments covering the coding strand of each individual *OPRM1* gene plus one HLA-DQα segment as quality control (65) were PCR amplified and pooled to sequence 15 segments simultaneously in one process; the same applies to the 15 segments covering the reverse strand. Segments covered in a second pass are marked by upper case letters. Filled bars indicate the sequences analyzed, open bars amplified template. DNA segments/PCR fragments are defined and analyzed by oligonucleotides that are used as PCR and sequencing primers, or as probes (M.R. Hoehe *et al.*, in preparation). Forward and reverse primers utilized to amplify adjacent fragments are reverse complementary to minimize sequence gaps between these fragments (~20 bp), which are covered by complementary strand fragments; usually, PCR primers are used as sequencing primers. Probes are reverse complementary to genomic sequence and placed adjacently to each sequencing primer ('nested detection'). Specific information on the oligonucleotides functioning as PCR, sequencing primers and probes will be made available on request. Overall, 80% of the sequence were covered by analysis of both strands; those DNA regions covered once were resequenced in case evidence for variation was obtained.

human μ opioid receptor gene (*OPRM1*) as a model system, we have combined these approaches to the analysis of genotype–phenotype relationships in complex genetic disease.

The human μ opioid receptor gene (*OPRM1*) is a prime candidate for substance dependence by both biology and genetic map position. Its gene product is the principal molecular target of morphine, mediating reward, tolerance and dependence (17,18). A marked interindividual variation in euphoric responsiveness to μ opioid receptor (MOR) agonists has been observed in opioid-naïve human subjects (19). Evidence from pharmacological, clinical and animal studies suggests that MORs may also be involved in the reinforcing actions of non-opioid drugs, such as cocaine and alcohol (20,21). Thus, the MOR could be a common component of addiction crossing pharmacological boundaries and may play a role in the mediation of drug reward. This hypothesis is supported by a reward-based animal model of substance dependence, morphine preference in B6 mice which have a genetic predisposition to experience euphoria-producing drugs such as morphine, but also cocaine and alcohol, in an intense and compelling manner (22). This represents on one hand an intriguing parallel to the polysubstance dependence of humans and on the other hand points to a genetic co-determination of polysubstance dependence. Genetic mapping of the morphine self-administration behavior revealed a locus of major effect harboring the *OPRM1* gene (22). Thus, analysis of genetic variation in this gene in relation to substance dependence in humans may contribute important clues to the pathophysiology of addiction. In our study, we have focussed on those individuals more likely to have a significant genetic component to etiology of their disease, to increase the power of the sample to detect the relevant gene(s) (23).

In this study, we have—for the case of the *OPRM1* gene—systematically analyzed candidate gene variation in all known functionally relevant regions of the gene including 6.7 kb regulatory, exonic and critical intronic sequences in a total of 250 substance-dependent individuals and controls. Critical intronic sequences include splice junctions, branch-point sequences, polypyrimidine tracts and in particular (A/T)GGG repeats (24). We have applied multiplex sequence comparison (MSC) to generate a substantial body of sequence data on multiple individuals for the same gene, equivalent to the analysis of ~1.7 Mb. Evidence for remarkable DNA sequence variation was obtained, as demonstrated by a total of 43 variants and 52 different haplotypes predicted for the subgroup of 172 African-American substance-dependent individuals and controls. Moreover, in our study, we classified both haplotypes and genotypes by similarity clustering, which allowed establishment of associations between (functionally) related groups of haplotypes and genotypes, respectively, and the disease phenotype. By this approach, we were able to extract a characteristic pattern of five sequence variants in the *OPRM1* gene, which was observed significantly more frequently in several forms of substance dependence, one of the most devastating complex diseases known.

RESULTS

Analysis of *OPRM1* gene sequence variation

High throughput analysis of DNA sequence variation was performed by application of a novel comparative sequencing approach, MSC, which is based on the principle of multiplex DNA sequencing (25). MSC has been designed to screen for DNA sequence differences over multiple candidate DNA segments in many individuals in parallel. This approach, which scales up conventional gel-based sequencing technologies by one order of magnitude, will be described in more detail elsewhere (M.R. Hoehe *et al.*, in preparation). A short description

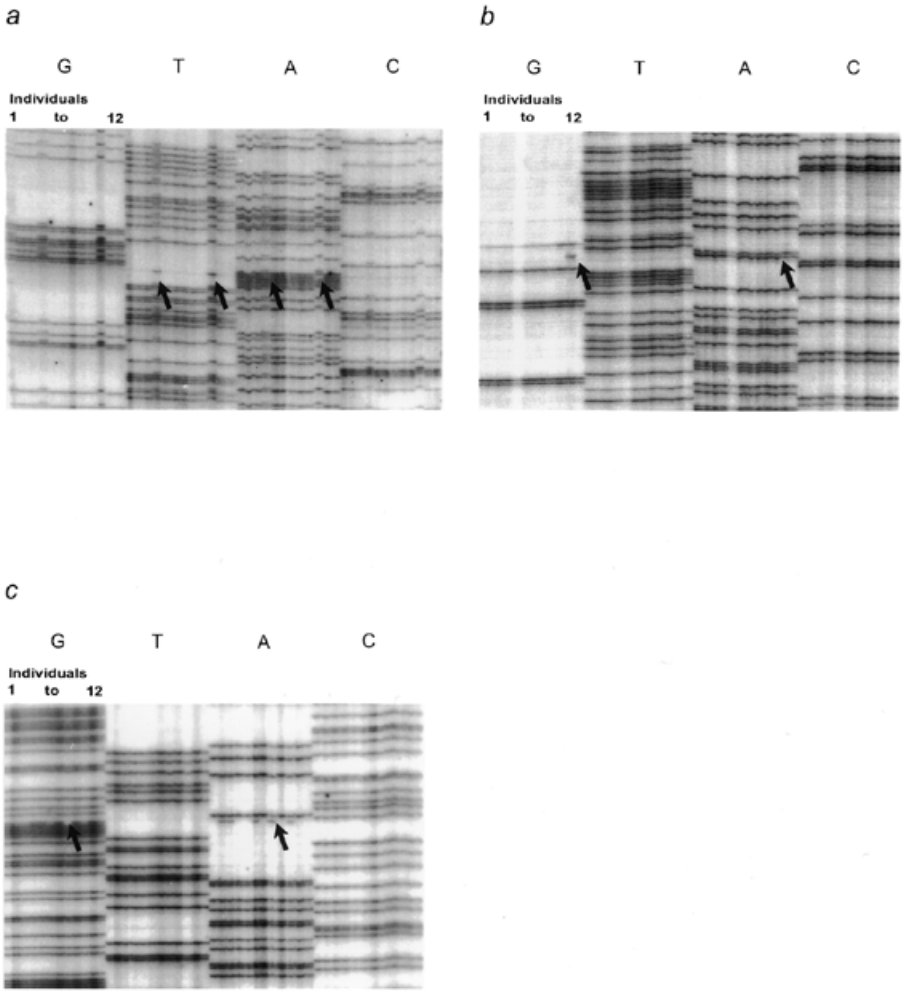

Figure 2. Images decoded during 15 sequential probing cycles. Images reveal complementary strand sequence. Base-specific termination reactions G, T, A and C are grouped in blocks of 12 individuals each, corresponding to lanes 1–12. This loading format allows identification of sequence differences between individual genes immediately as additional and/or missing bands, as indicated by arrows ('skilled pattern analysis'). In total, 96 lanes were loaded on one gel, equivalent to the comparison of 24 individual gene sequences defining one target DNA segment. (**A**) Image revealing two A→T substitutions in segment 1 (Fig. 1), one of which was heterozygous in individual 4 (both a weaker A and a T band visible) and one homozygous in individual 10 (lack of an A band, strong T band). In addition, a single nucleotide insertion originating 5′ to the target segment sequence can be observed in the same individuals as heterozygous and homozygous variants, with consequences in all four base runs. (**B**) Image revealing a heterozygous A→G substitution in individual 11, segment 2. The opposite A is expected to be less clear since it is changing from two copies (in all other individuals) to one (in this variant) whereas the G is changing from zero copies to one. (**C**) Image revealing more frequent G→A substitutions in segment 6 in five individuals, four of whom were heterozygous (nos 2, 3, 6 and 10) and one homozygous (no. 8). Decodings refer to probing cycles 14, 7 and 1.

of the principles and the protocols is given in Materials and Methods. This study describes the first application of MSC to the analysis of candidate gene variation on a megabase scale. The *OPRM1* gene including its 5′ regulatory, exonic and critical intronic regions (24) was dissected into target DNA segments (Fig. 1). For each individual, 15 PCR products covering either coding or complementary strands were pooled and simultaneously sequenced by MSC. Two hundred and fifty individual *OPRM1* genes, both strands, were analyzed. Given ~6.7 kb per gene, this amounts to ~1.7 Mb of 'comparative' sequence data. Examples of typical data generated by this approach, demonstrating both heterozygous and homozygous sequence variations in individuals, are shown in Figure 2.

Regarding nature and distribution of sequence variation in *OPRM1* in two major populations, 172 African-Americans and 66 European-Americans, a total of 43 biallelic variants were identified by sequencing (Fig. 3), 40 of which were single nucleotide polymorphisms (SNPs) (93%). Among those, transition substitutions were more prevalent (25 of 40, 62.5%) than transversions (15 of 40, 37.5%). Three insertion/deletion variants (7%) were found. Among these 43 variants, 15 were observed only once. Twenty-four variants were present in the 5′ regulatory region (1 every 99 bp), 4 in the 5′-untranslated region (UTR) (1 every 73 bp), 1 in the 3′-UTR (1 every 710 bp), 6 in the coding region (1 every 267 bp), 5 of which clearly affect the encoded protein, and 8 variants in the intronic regions (1 every 261 bp). The critical regions of intron 2 appeared more polymorphic (1 every 110 bp) than those of introns 1 and 3 (1 variant in ~1580 bp). Clearly, the relative number of variable sites in non-coding regions (1 every

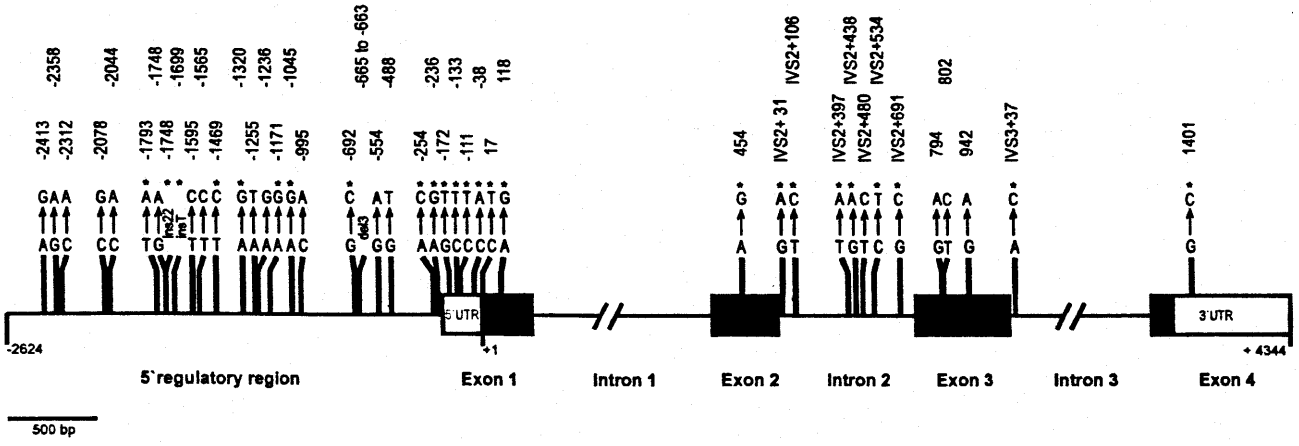

Figure 3. Polymorphic spectrum of the *OPRM1* gene. The 6968 bp genomic reference sequence (Fig. 1) is presented as baseline; base pair coordinates relative to the translation start site are given. Sequences are drawn to scale, which is indicated. All gene variants are specified by position numbers and nucleotide variations (substitutions, insertions and deletions) according to the mutation nomenclature (66–68). Those sites marked by an asterisk have been included in haplotype analysis. The other variants given were observed only once with the exception of 942, a silent mutation, which was excluded from haplotype analysis, as were variants –2078 and –2358, which were discovered during the additional pass of sequencing. All these variants were observed maximally at a frequency of ~3%.

147 bp) was higher than in coding regions. In the sample of African-Americans, 4 variants, –172, 17, IVS2+31 and IVS2+691, exhibited a frequency of 10% or more in the group of cases ($n = 137$) and/or controls ($n = 35$) (Table 1). Of the variants identified in the second pass of sequencing, variants –2358 and –2078 were seen in 5 and 2 of the African-Americans, respectively. Nine variants exceeded a heterozygosity of 10% in the group of cases and/or controls, 4 of these a heterozygosity of 20%, the mean observed heterozygosity for the 43 variants in this ethnic group being <0.065. Among European-Americans, 5 variants (–1748, –1320, –1045, –172 and –38) were present with frequencies between 1 and 10% in the groups of cases and/or controls and 3 variants (118, IVS2+31 and IVS2+691) exceeded a frequency of 10%. Significant ethnic differences were found for variant 17, which was observed in African-Americans only ($P < 0.005$), and variant 118, which was more frequent in European-Americans ($P < 0.05$). Variant –1236 was identified in one European-American only. All observed genotype frequencies in all subgroups were in Hardy–Weinberg equilibrium.

Variants of interest in the 5′ regulatory region were those potentially affecting transcription regulatory motifs (24): a T→A exchange in a YY1 transcription factor binding motif at nucleotide position –1793 neighboring a CRE motif and a T insertion in an AP-1 binding site at nucleotide position –1699 (Fig. 3). In the coding region, two amino acid substitutions (A6V at bp 17 and N40D at bp 118) were found in the N-terminus (26–28). Additional substitutions of highly conserved amino acids were observed in the third transmembrane domain, N152D at bp 454, and in the third cytoplasmic loop, R265H at bp 794 and S268P at bp 802, which destroyed a CaM-kinase II-dependent phosphorylation site (29). The N152D mutation was found in two African-American patients, the R265H and the S268P mutations were found in one African-American patient each. The intronic variant IVS2+31G→A disrupts one of the (A/T)GGG repeats, which have been found to regulate alternative splicing (30).

Haplotype (genotype) analysis

As the next step, we applied statistical techniques to predict the specific combinations of variants, haplotypes, on each of the two chromosomes of each individual. We estimate that current experimental methods (15,31) to determine haplotypes to be 10 times more costly than the combination of algorithm and diploid (mixed) sequencing applied here. The ethnic groups were analyzed separately; in the following, we concentrate on the notably larger group, the African-Americans. As described the heterozygosity (32) in our sample is not high, the mean observed heterozygosity for the 25 variants included in haplotype analysis being <0.1 (Table 1). We observed 25 of 172 individuals being homozygous at all variable positions and 33 individuals being heterozygous at only one single site, each providing unambiguous information about the haplotypes. The average genotype was heterozygous at three sites. Therefore, a reliable statistical prediction of the genetic haplotypes can be expected. Nevertheless, potentially important results obtained by statistical haplotype prediction will need to be confirmed by molecular genetic techniques. To process numerous variants, the computer program MULTIHAP has been developed by us (see Materials and Methods); MULTIHAP has been specifically designed to predict the most likely haplotype pair for each genotype in a sample. We included in the analysis those 25 variants of the total of 43 that occurred more than once, with the exception of variant 942, a third-base exchange not involving an amino acid exchange (Fig. 3; Table 1). In the combined group of 172 substance-dependent individuals and controls, 52 different haplotypes could be distinguished (Table 2). The five most frequent haplotypes common to both groups were haplotypes no. 43 (39 and 36% in cases and controls, respectively), no. 14, which was completely identical to the reference sequence (12 and 19%), no. 4 (7% in both groups) and nos 24 (4 and 6%) and 7 (4 and 6%). These five haplotypes constituted 66% of the haplotypes in the group of substance-dependent individuals and 74% of haplotypes in the controls. In the smaller European-American sample, 13 different haplotypes

Table 1. Relative frequencies and heterozygosities of variants

Position		Frequencies		Heterozygosities	
Ref. seq.	Haplotype	Controls	Cases	Controls	Cases
–1793	1	0.014	0.088	0.028	0.160
–1748	2	0.029	0.022	0.056	0.043
–1699	3	0.014	0.099	0.028	0.178
–1469	4	0.043	0.015	0.082	0.029
–1320	5	0.014	0.091	0.028	0.166
–1171	6	0.000	0.051	0.000	0.097
–1045	7	0.014	0.015	0.028	0.029
–692	8	0.043	0.055	0.082	0.103
–254	9	0.000	0.022	0.000	0.043
–236	10	0.000	0.007	0.000	0.014
–172	11	0.114	0.102	0.202	0.194
–133	12	0.000	0.022	0.000	0.043
–111	13	0.014	0.091	0.028	0.166
–38	14	0.014	0.018	0.028	0.035
17	15	0.186	0.197	0.303	0.316
118	16	0.043	0.040	0.082	0.077
454	17	0.014	0.004	0.028	0.007
IVS2+31	18	0.143	0.095	0.245	0.171
IVS2+106	19	0.014	0.004	0.028	0.007
IVS2+397	20	0.014	0.022	0.028	0.043
IVS2+438	21	0.043	0.033	0.082	0.064
IVS2+534	22	0.000	0.044	0.000	0.083
IVS2+691	23	0.429	0.500	0.490	0.500
IVS3+37	24	0.000	0.007	0.000	0.014
1401	25	0.029	0.026	0.056	0.050

Relative frequencies and heterozygosities of the variants found in African-Americans with a frequency of >1%, with the exception of variant 942, a silent mutation. These correspond to those included in haplotype predictions (Fig. 3). Positions are specified according to the reference sequence (Ref. seq.) and to the haplotype.

were distinguished. The most frequent ones corresponded to those in the African-American sample.

Classification of haplotypes (genotypes) and risk pattern identification

Obviously, the number of different haplotypes is unfeasibly large, so that the power is not sufficient to detect an association with any single haplotype. Previous approaches to perform association analysis, many different haplotypes given, were based on evolutionary relationships (33–35). Another approach to cope with the multiplicity of haplotypes could be the classification of haplotypes into functionally related (ideally functionally equivalent) ones based on sequence–structure–function similarity. Once a classification has been derived, the haplotype frequencies of cases and controls in the different classes can be compared. Because it is not known how many different classes may exist (if they exist), a stepwise classification process such as a hierarchical cluster procedure appears suitable. This procedure is based on an algorithm which starts with each haplotype in a separate cluster (step 0) and merges step by step the two most similar clusters until one final cluster is left (for our sample step 21). The results of the cluster analysis are illustrated by a dendrogram. The classification procedure is performed independently of the phenotype. The existence of functionally different classes would be likely, if at least one class included haplotypes from cases significantly more (or less) frequently than haplotypes from controls. In that case, the haplotypes in the different classes are inspected for consensus patterns. The pattern(s) observed more frequently among individuals with disease could be interpreted as susceptibility pattern(s) whereas pattern(s) more frequent among controls could be considered protective.

The results of the hierarchical cluster analysis of *OPRM1* haplotypes are presented in Figure 4. This cluster analysis provided the basis for comparing the haplotype frequencies of cases and controls between the clusters, which have been derived at each step of the clustering process; the most significant *P*-value of 0.017 was reached at step 20, where two clusters were left. Haplotypes in one of these clusters corresponded to a mixture of cases and controls, whereas the other cluster contained with one exception haplotypes of substance-dependent individuals only. Analyzing the haplotype profiles in this second cluster, haplotype numbers 3, 9, 11, 17, 25, 27, 31, 37, 45, 49, 51, 16 and 39 were found combined (Table 2). The first 11 haplotypes featured a unique constellation of five polymorphic sites (pattern) located in the 5′ regulatory region and exon 1. The first two sites at nucleotide positions –1793 and –1699 were those affecting sequence motifs potentially critical for transcriptional regulation (24), the third site involved an A→G exchange at nucleotide position –1320, the two additional sites were located in exon 1, at nucleotide positions –111 and 17 (A6V). Two more haplotypes were found in this cluster, which contained three of these sites (haplotype nos 16 and 39) and in addition a variation 5′ upstream, at nucleotide position –2078, affecting a potential CRE/AP-1 motif (24). This additional variation was observed exclusively in these two individuals. Comparing these haplotypes, one might assume that the second, third and fourth of the polymorphic sites common to all of them may be those related to substance dependence. On the other hand, 92% of the haplotypes in this cluster contain the complete combination of these five sites. This clearly stresses this constellation; the observation that the combination of three of these polymorphic sites in haplotypes 16 and 39 is complemented 5′ upstream by a variation at a potentially relevant CRE motif might indicate a second haplotype pattern related to the disease. Since the importance of these polymorphic sites and their combinations is yet unknown, we have conservatively used the class containing the haplotypes with the complete pattern as the basis for association analysis and included haplotypes nos 16 and 39 in a second pass. The remaining cluster, in contrast, did not yield any evidence for a distinctive pattern.

Classifying the corresponding genotypes (Table 3) by similarity clustering resulted in the distinction of three categories. The first cluster contained exclusively substance-dependent individuals, who were homozygotes for the described haplotype pattern. The second included—with one exception—substance-dependent individuals, who were heterozygotes, and the third included again a mixture of substance-dependent

Table 2. Haplotypes of African-American substance-dependent individuals and controls

No.	Haplotypes	Counts		No.	Haplotypes	Counts	
		Cases	Controls			Cases	Controls
1	1111111211111111111111211	12	1	27	2121221111112121111121112	2	0
2	1111111111121111121111111	1	0	28	1111111111121112111111211	1	0
3	2121211111112121111211111	1	1	29	1211111121211121111111111	1	0
4	1111111112111111111111111	20	5	30	1211111111111111111211111	1	0
5	1111111111111111112111111	1	0	31	2121211111112121111111111	3	0
6	1111112111111111111112211	2	0	32	1112111111111121121121112	0	1
7	1111111111111121111111111	10	4	33	1111111111111111112111111	1	1
8	1111111211111121111111111	2	2	34	1111112111211111111111111	1	0
9	2121211111112121111112111	1	0	35	1111111111111111121111121	1	0
10	1111111111111112111111211	8	3	36	1111111111111121121111111	6	2
11	2121212111112121111111111	1	0	37	2121211111122121111112111	1	0
12	1211111111211111111111111	0	1	38	1111111111211211111111211	1	0
13	1111112111211121111111111	0	1	39	1121211111112111111111111	1	0
14	1111111111111111111111111	34	13	40	1111111111111121111121112	3	1
15	1111111111211211121111111	0	1	41	1121111211111121111121111	1	0
16	1121211111112111121111111	1	0	42	1211111111211121111121112	1	0
17	2121211112112121111211111	2	0	43	1111111111111111111111211	108	25
18	1111111111111121111111121	1	0	44	1112111111111111121111111	4	1
19	1111111111111111211111111	0	1	45	2121221111112121111111111	1	0
20	1211111111111111111111211	0	1	46	1111111121111121111111111	5	0
21	1111111111111111121211111	1	0	47	1211111211111111111111211	1	0
22	1112111111111121121121111	0	1	48	1111111111121111111111211	1	0
23	1211111111111112111111211	2	0	49	2121211111112121111111211	1	0
24	1111111111111111121111111	11	4	50	1111111111121121111121112	1	0
25	2121221111112121111121111	1	0	51	2121221111112121111112111	10	0
26	1111111112111112111111111	1	0	52	1111111111211211111111111	4	0

1, identical with the reference sequence (for GenBank accession nos see legend to Fig. 1); 2, different from the reference sequence (for newly identified variants see Fig. 3).
The haplotypes are numbered (left column); at the right, the absolute frequencies of cases and controls carrying the haplotypes are given. For polymorphic sites specified by positions 1–25 see Table 1.

individuals and controls without any characteristic features. It is interesting to note that genotypes 55 and 56 (Table 3) containing haplotypes 16 and 39 (Table 2) were found included in the second cluster (results of the hierarchical cluster analysis are not shown). These results confirm the haplotype-based analysis and one can assume a dominant mode of action. Importantly, the data suggest that the predicted 'statistical' haplotypes may reflect the genetic haplotypes. In fact, subsequent allele-specific PCR experiments utilizing as template a gene segment, which included all five polymorphic sites of this pattern, clearly confirmed this for each carrier of the pattern. Therefore, the genetic haplotypes characterized by the pattern and not the statistically predicted haplotypes were utilized in the final phase of statistical analysis.

The frequencies of haplotypes containing this pattern were compared in a final sample of 51 African-American controls and 158 African-American substance-dependent individuals. These included those individuals who had been genotyped by MSC (Table 3) (35 controls, 137 cases) and those who had been recruited in addition (see Materials and Methods: Subjects; 16 controls, 21 cases) and genotyped by means of allele-specific PCR for absence or presence of the pattern. Both groups, the 51 controls and 158 cases, were significantly different (P = 0.002); including haplotypes 16 and 39, a P-value of 0.001 was obtained. Also, the subgroups of cocaine- (n = 125) and opiate- (n = 33) dependent individuals differed significantly from controls (P = 0.004 and P = 0.006, respectively). Including haplotypes 16 and 39 in the subgroup of cocaine-dependent individuals, a P-value of 0.002 was reached. These results were confirmed by association analysis based on genotypes, combining individuals, who were homo- or heterozygous for this pattern. Again, significant differences

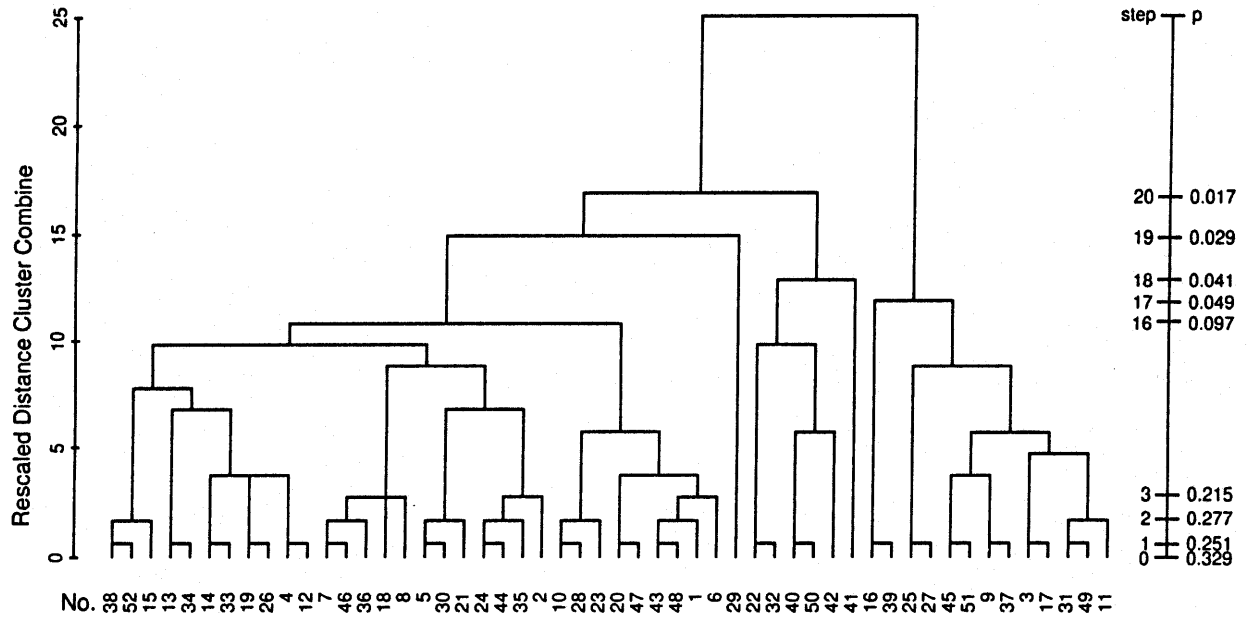

Figure 4. Dendrogram reflecting the hierarchical clustering process of haplotypes according to similarity. The actual distances at each step are described by distance coefficients, that represent average distances between clusters and have been rescaled to an interval from 0 to 25 (left). The haplotype numbers given at the bottom correspond to those given in Table 2. At the right side, the iterative steps in the clustering process and selected *P*-values are indicated (asymptotic values and from step 17 exact-test values, are presented). *P*-values indicate significance levels resulting from the comparison of haplotype frequencies of cases and controls between the clusters created at each given step. Some of the rescaled distance differences are so small that they are not visible in the dendrogram.

were obtained in all corresponding tests, including genotypes 55 and 56 in a second pass, with *P*-values ranging between 0.002 and 0.006. Estimated odds ratios for substance dependence (pooling heterozygous and homozygous genotypes carrying the pattern of five variants) were 9.85 (51 controls versus 158 cases; CI 1.3–74.5) and 10.8 (including genotypes 55 and 56; CI 1.4–81.3) (36). These results suggest that distinct allelic categories may exist, one of which may be related to substance dependence.

A parallel approach to perform association analysis in this situation of haplotype multiplicity was based on established cladistic approaches that estimate the 'evolutionary relatedness' of haplotypes. An evolutionary tree was constructed by means of the distance-based neighbor-joining method (37) (see Materials and Methods). One defined region of the cladogram was found associated with a significant phenotypic effect ($P = 0.021$). Specifically, one clade in this region included precisely those haplotypes which were characterized by the unique combination of five variants described above. Clearly, these results were in excellent agreement with the results obtained by means of the hierarchical cluster analysis.

Accordingly, haplotype calculations and hierarchical cluster analysis were carried out in the European-American sample. The final sample includes 49 controls and 50 cases ($n = 99$); of these, 38 controls and 23 cases ($n = 61$) had been genotyped by MSC (data not shown), the additional individuals (11 controls, 27 cases) had been genotyped by allele-specific PCR for absence or presence of the pattern. Haplotype analysis and subsequent haplotype classification were performed. However, no evidence of any haplotype pattern, or any common variant, in this smaller sample was given.

DISCUSSION

In our analysis we were forced to address a problem which we expect to be of increasing importance in the future: the enormous sequence variability in human genes, making it extremely difficult to analyze complex genotype–phenotype relationships. Systematic analysis of variation in entire candidate gene sequences results in numerous individually different forms of the gene identified as haplotypes. This requires a logical approach to reduce the multiplicity of haplotypes. A reasonable approach to this problem seems to classify those haplotypes that are similar in function. If a relationship between candidate gene haplotypes and a disease phenotype exists, then significant frequency differences between cases and controls should be observed between the haplotype classes. Consequently, the sequence variants common to such a functionally related class of haplotypes must include those variants that determine (or serve as markers for) differences in gene function. The critical question now arises of how to derive a classification that may reflect functionally different haplotype classes.

The hierarchical cluster analysis used to classify haplotype information is a simple and robust method to describe and analyze haplotype or genotype similarities. Given that different clustering methods may produce different solutions based on the same input data, the question naturally arises of whether the derived clusters are valid. Certain criteria, such as the stability of results or their consistency with the data (38), decide which of two different clustering results may be better. In order to check the stability of our haplotype clustering results, we have in addition applied several combinations of clustering measures ('Squared Euclidean distance', 'Euclidean distance', 'Simple matching') and special methods ('Between-

Table 3. Genotypes of African-American substance-dependent individuals and control

No.	Genotypes	Counts		No.	Genotypes	Counts	
		Cases	Controls			Cases	Controls
1	111111111111111111111111	2	1	40	111111121111111121111211	3	0
2	111111111111111111111211	14	3	41	111111121111121111111211	2	2
3	111111111111111111111311	16	2	42	111111121111121121111211	1	0
4	111111111111111111211211	1	0	43	111111121121111111111211	1	0
5	111111111111111112111211	1	1	44	111111211111111111111211	1	0
6	111111111111111121111111	2	0	49	111211111111111131111111	0	1
7	111111111111111121111211	4	2	50	111211111111112121111211	1	0
8	111111111111111121111221	1	0	51	111211111111121121121111	0	1
9	111111111111111121211111	1	0	52	111211111111121121121112	0	1
10	111111111111111211111211	0	1	53	111211111121111121111111	1	0
11	111111111111112111111211	1	0	54	112111121111121111121211	1	0
12	111111111111112111111311	2	2	55	112121111111211111111211	1	0
13	111111111111121111111211	4	4	56	112121111111211121111111	1	0
14	111111111111121111111221	1	0	57	121111111111111111111211	0	1
15	111111111111121111121212	3	1	58	121111111111111111211111	1	0
16	111111111111121121111111	1	2	59	121111111111112111111311	2	0
17	111111111111121121111211	1	0	60	121111111121111111111211	0	1
18	111111111111122111111211	3	0	61	121111111121121111121112	1	0
19	111111111111131131111111	1	0	62	121111121211121111111111	1	0
20	111111111121111121111111	1	0	63	121111121111111111111311	1	0
21	111111111121111121111211	1	0	64	212121111111212111111211	2	0
22	111111111121112111111311	1	0	65	212121111111212111111311	1	0
23	111111111121121111121212	1	0	66	212121111111212111112211	1	0
24	111111111211111111111111	2	1	67	212121111111212111211211	1	1
25	111111111211111111111211	7	2	68	212121111111212121111111	1	0
26	111111111211111121111111	1	1	69	212121111112212111112211	1	0
27	111111111211111211111211	1	0	70	212121111211212111211211	2	0
28	111111111211112111111211	0	1	71	212121211111212111111211	1	0
29	111111111211121111111111	1	0	72	212122111111212111112111	1	0
30	111111111211121121111111	1	0	73	212122111111212111112211	2	0
31	111111111211211111111211	4	0	74	212122111111212111121212	1	0
32	111111111211211111111311	1	0	75	212122111111212221111112211	1	0
33	111111111211211121111111	0	1	76	212122111111212311111112111	1	0
34	111111111131111111111111	2	0	77	212122111112212111111112211	1	0
35	111111112111112111111211	3	0	78	212122121111212111111211211	1	0
36	111111112111113111111111	1	0	79	313133111111313111111112111	1	0
37	111111112121112111111111	1	0	80	313133111111313111111113111	1	0
38	111111121111111111111211	2	1	81	313133111111313111111122112	1	0
39	111111121111111111111311	3	0				

1, homozygous for nucleotides/sequences identical with the reference sequence; 2, heterozygous; 3, homozygous for nucleotides/sequences different from the reference sequence.

The genotypes are sorted here 'alphabetically', in order to provide a good visual impression of the different genotypes, and numbered (left column); at the right, the absolute frequencies of cases and controls carrying the genotypes are given. For polymorphic sites defined by positions 1–25 see Table 1.

groups linkage', 'Centroid clustering', 'Nearest neighbor'). All applications resulted in a similar clustering and, importantly, the same two classes were consistently obtained in the last classification step, which determined here the result. Additionally, applying the non-hierarchical 'K-Means cluster analysis' with the initial cluster centers method under the assumption of two groups, exactly the same haplotypes were found combined in two clusters. At present, we are deriving similarity measures for classification procedures such that these measures may reflect properties that determine sequence–structure–function similarities, for instance physicochemical properties.

Another approach to combine information from different haplotypes (e.g. multiple restriction site haplotypes) uses the historical information in haplotypes to construct a cladogram that estimates how the different haplotypes are evolutionarily related (33,39). This allows localization of functional mutational changes in the haplotype network by the identification of phenotypic contrasts between sister clades. The use of this evolutionary tree as a statistical design may become difficult when various forces may have influenced the evolutionary history of a population, such as a high rate of recombination (40), multiple mutations to high susceptibility alleles, rapid population growth, migrations, strong selection, or co-evolutionary pressures like changes in environment. Undoubtedly evolution is an essential force that influences present genome sequence variability and the range of given functional variation. Thus, our approach that relies on sequence–function relations is not incompatible with cladistic approaches. However, reconstruction of the specific evolutionary process is neither feasible in most complex disease studies nor even essential for the analysis of genotype–phenotype relationships. It should be sufficient to focus on the 'here and now', the given DNA sequence and its individual variation as it determines structure and function.

Such an approach seems more generally applicable. It may also be more suitable to cope with the various scenarios that may underlie functional variation in complex disease. The spectrum of polymorphic profiles may include single significant mutations as well as mutations in different haplotype frameworks, for instance, or any combinations (patterns) of variants that may cause different or similar pathogenetic constellations. After all, the historically grown concept of 'single causative mutations' found to underlie discrete protein defects in Mendelian diseases may be poorly compatible with those graded functional differences active in complex disease. It is unproven and possibly fallacious in the situation of complex disease. It should be added that, in this analysis we have focussed on one candidate gene—a relatively simple level of complexity in view of the general model that both gene–gene and gene–environment interactions play a key role in the genetics of complex disease (13,16,41).

In our study, in the *OPRM1* gene, we have identified a combination of variants, which is predominantly composed of a specific constellation of changes in putative transcription regulatory motifs and exists significantly more frequently in African-American substance-dependent individuals likely to have a significant genetic predisposition to their substance dependence. Given that we have strong prior arguments for the *OPRM1* candidate gene hypothesis, *P*-values in the range 0.001–0.006 for both haplotype and genotype analysis appear sufficient to support the hypothesis of an association (42,43). Of course, we recognize that these nominal significance levels for an association are difficult to interpret and cannot be accepted as proof of an etiologic relationship. Additional support for an etiologic relationship must come from examining this hypothesis in new groups of patients and controls, as well as in family-based analysis. Our results suggest that individuals carrying this characteristic pattern may be at increased risk of developing substance dependence under certain additional environmental and genetic circumstances. Both genotypes, the heterozygous and the homozygous ones, characterized by one or two of these putative susceptibility patterns, are apparently associated with substance dependence. It is interesting to note that variation in promoter sequences fits quite well with this (dominant) mode of action, often observed in cases of overexpression or inappropriate expression patterns.

In theory, population stratification would be an alternative explanation for the haplotype–phenotype association described here. Available documentation described in Materials and Methods does not suggest the presence of a different subpopulation among the patients. Additional data, examining candidate genes that are particularly sensitive to population differences such as *DRD2* (44,45) and the serotonin transporter gene (46) (B. Wendel *et al.*, in preparation) in these same African-American individuals, show no evidence of such stratification. We therefore conclude that the most likely explanation of our data is a population-specific genetic disposition to substance dependence, which can be observed in ~16% of the affected. In a global survey conducted by our group, including eight major populations of different geographic origins, this pattern could be observed in three Biaka-Africans of ten individuals tested. Thus, this observation suggests that this characteristic pattern originated in Africa. Moreover, the conservation of this combination of polymorphisms in relatively old, sub-Saharan populations (47), including the African ancestors of the younger African-American population, might point to a possible functional significance of this gene segment.

What could this association result imply about the genetic mechanism potentially involved in vulnerability to substance dependence? There are two possibilities: this pattern may serve as a marker, which indicates the location of genetic variation conferring risk to the disease and resides in linkage disequilibrium. This may refer to any neighboring genes, remote regulatory elements or any regulatory features within the large introns 1 and 3. No strong evidence for other candidate genes is presently given on either the human or the homologous murine chromosomal region. On the other hand, there is suggestive evidence that this polymorphic pattern, either one or several polymorphisms in combination, may be causative. The fact that the pattern of polymorphisms concentrates on the 5′ regulatory region and does not extend further upstream, as demonstrated by additional sequence analysis, may support this hypothesis. Moreover, this pattern reflects a specific constellation of putative transcription regulatory sequence motifs. Given the finding that mice lacking MORs do not experience the reinforcing properties and reward mediated by the MOR agonist morphine and do not develop tolerance and dependence (17), we suggest that this polymorphic pattern may confer regulatory properties that result in an increase of receptor expression on exposure to drugs. Specifically, *OPRM1* may be overexpressed as the result of an altered feedback mechanism, or a disturbance of regulatory properties after prolonged agonist exposure; on the other hand, *OPRM1* may

have altered regulatory properties in specific brain tissues. The existence of individual differences in *OPRM1* expression levels in mouse strains and humans has recently been proposed (48).

Importantly, this pattern was found in individuals, who—even though diagnosed with a primary dependence—satisfied criteria for several forms of dependence including cocaine, opiate dependence and alcoholism and were characterized by a potential strong hereditary base for polysubstance dependence. This represents a striking parallel to a reward-based animal model of substance dependence, morphine-preferring B6 mice, which excessively self-administer opioids, cocaine and alcohol and obviously have a genetic co-determination for polysubstance dependence (22). This predisposition genetically maps to a segment of chromosome 10 harboring the murine μ opioid receptor gene (*Oprm*). Recent findings in this C57B1/6J strain indicate differential agonist regulation of striatal *Oprm* expression (49). Taken together, this pattern may be related to expressional alterations in central reward structures that may be involved in the genetic co-determination of several forms of substance dependence.

A series of clinical and experimental evidence has suggested an involvement of opioidergic regulatory mechanisms in the disposition to opioid and cocaine dependence and alcoholism (20,21). Interestingly, an upregulation of MOR binding has been detected by PET in cocaine-dependent men and was associated with cocaine craving in living human subjects (50). Moreover, an increase of *OPRM1* expression could be compatible with the therapeutic properties of the hMOR antagonist naltrexone, which prevents relapse and craving in alcoholics and exerts differential subjective alcohol responses in subjects with a family history of alcoholism (51,52). As with any newly proposed risk factor, independent tests of our results are necessary. Moreover, evaluation of possible functional implications of the described *OPRM1* variations *in vitro* and *in vivo* will be necessary to test the hypothesis that *OPRM1*-related mechanisms may be involved in the pathophysiology of addiction crossing pharmacological boundaries.

To conclude, this study shows the importance of comparing entire candidate gene sequences and future challenges of the genetics of complex disease. The development and application of approaches to the analysis of genotype–phenotype relationships against a background of high natural genome sequence variability can lead to testable genetic and functional hypotheses.

MATERIALS AND METHODS

Subjects

Individuals from residential and non-residential addiction treatment programs were included, if they met modified Research Diagnostic Criteria (RDC) (53) and DSM-IV criteria for dependence on opioids or cocaine and whether systematic family history information revealed at least two first degree relatives with a past or present diagnosis of substance dependence. Detailed diagnostic procedures, validation of diagnoses and family histories have been described (44,53). All drug-dependent individuals gave extensive histories of dependence on opioids or cocaine, in which jobs, family, houses and all financial resources were lost in pursuit of drugs. They had serious multi-year drug dependence, with age at onset below 20 years. Although a primary drug dependence was identified (either cocaine or opioids) for each patient, nearly all these patients satisfied diagnostic criteria for dependence or abuse of multiple substances, including cocaine, alcohol and opioids. Normal volunteers were recruited from the same residential area (26). These individuals had an SADS-L (Schedule for Affective Disorders and Schizophrenia—Lifetime Version) interview and a systematic family history was obtained. Individuals with a personal or family history of drug abuse, dependence, alcohol dependence or a major psychiatric disorder (schizophrenia, unipolar or bipolar illness) were excluded from the final control group (26). The volunteers were matched by gender and ethnic background to the substance-dependent group. Recruiting of controls from the same residential area as the substance-dependent group provides some control over different proportions of African and European ancestry among the African-Americans. No attempt was made to match the normal volunteers by age. The mean ages for African-American controls ($n = 51$) and cases ($n = 158$) were 35.1 ± 5.5 and 30.1 ± 8.9 years, respectively. Fifty milliliters of EDTA-treated venous blood were obtained for DNA extraction (54). Two hundred and fifty individuals, 172 African-Americans, 66 European-Americans and 12 individuals of other ethnic backgrounds, were analyzed by MSC. The relatively high proportion of African-Americans among the sample reflected the demographics of the Philadelphia neighborhoods served by the programs from which the patients were drawn. On completion of the analysis of the results obtained by MSC, additional cases and controls, who had been recruited according to the same ascertainment and diagnostic criteria, were genotyped by means of allele-specific PCR and included in the final statistical analysis.

MSC

The principle of MSC in brief (M.R. Hoehe *et al.*, in preparation) is as follows. Genes (both strands) are dissected into multiple target DNA segments that are PCR amplified and pooled to perform multiple (to date maximally up to 55) sequencing reactions simultaneously in one reaction tube. Pools of base-specific termination reactions are resolved on a sequencing gel and transferred by direct transfer electrophoresis (DTE) onto a nylon membrane. By sequentially hybridizing the membrane with different probes each highly specific for one of the original target DNA segments and by repeating the hybridization cycle as many times as there are PCR products in the original pool, sequences of all DNA segments can be read. The loading format allows identification of sequence differences between individual target DNA segments immediately by visual inspection of images (skilled pattern analysis) (Fig. 2).

Gene dissection and oligonucleotide design

Dissection of *OPRM1*, both strands, into target DNA segments is illustrated in Figure 1. PCR primers were designed to allow amplification at 60°C annealing temperature and probes designed to allow hybridization at 42°C. Specific information on the target DNA segments/PCR fragments and the oligonucleotides (PCR and sequencing primers and probes) defining the two plex pools that cover both strands of the gene will be made available on request. The PrimerSelect program (DNASTAR package) was used.

PCR and preparation of single-stranded template

Two sets of 15 PCR products per individual were generated to cover forward and reverse strands (Fig. 1), specifically 5′ biotinylated antisense primers were introduced to allow direct solid-phase sequencing of the coding strand. Products were amplified separately in PTC 225 Tetrads (MJ Research, Oldendorf, Germany). The reaction mixtures contained 100 ng of genomic DNA, 10 mM Tris–HCl pH 8.3, 1.5 mM $MgCl_2$, 50 mM KCl, 200 µM of each nucleotide, 30 pmol of each primer, 3 U *Taq* polymerase and H_2O in a final volume of 50 µl each. Polymerase was added during a time interval of 4 min at 88°C, following a denaturation phase of 94°C for 4 min, followed by 30–35 cycles at 94°C (15 s), 60°C (15 s) and 72°C (30 s), or at 94°C (15 s), 60°C (30 s) and 72°C (1 min). All 15 PCR products per strand per individual (10 µl of each equivalent to ~200 ng of PCR DNA) were pooled and purified via the 96-well QIAvac manifold system (Qiagen, Hilden, Germany) according to the manufacturer's protocols. For solid phase-based multiplex sequencing of the coding strand, single-stranded (ss) PCR DNA was isolated by magnetic bead–streptavidin/biotin interaction. Beads had been washed according to recommendations (Dynal, Oslo, Norway), which had been modified to allow simultaneous immobilization of 15 PCR products, utilizing only about a quarter of the amount of beads recommended. Then, the pooled purified PCR products in a final volume of 160 µl were added to equal volumes of pre-washed beads and were incubated for 15 min at room temperature. Melting DNA duplexes and separation of DNA strands were performed according to the manufacturer's protocol; however, 4- to 5-fold volumes of buffers and solutions were used to accommodate for the 15-fold amount of processed PCR DNA. At the end of the processing, beads bound with ss PCR DNA were resuspended in 16 µl of H_2O.

Fifteen plex solid-phase T7/Mn sequencing of PCR DNA

T7 sequencing reactions were performed according to the manufacturer's protocol (Amersham, Braunschweig, Germany), utilizing, however, the 2-fold amount of reagents compared with the 15-fold amount required if the reactions had have been processed individually, with the exception of the amounts of the 15 different sequencing primers. The specific modifications for 15 plex sequencing will be described in detail elsewhere (M.R. Hoehe *et al.*, in preparation). One microliter of sample was loaded onto the sequencing gel.

Fifteen plex cycle sequencing of PCR DNA

Complementary strand sequencing was performed with 15 plex cycle sequencing utilizing 10 µl, equivalent to ~50 ng, of each PCR product, Thermo Sequenase and reaction conditions as described (Amersham), with the exception that the sequencing primer was replaced by a mixture of 15 primers, 0.5 pmol per primer, respectively. Thus, 15 sequencing reactions were carried out with 1× reagents. The major part of the gene was analyzed by 15 plex sequencing; in a second pass of sequencing directed towards the 5′ end of regulatory sequence and additional 3′ intron 1 sequence (Fig. 1), direct sequencing of these three PCR segments, or an upgraded 18 plex pool, were used.

DTE

Ninety-six reactions (equivalent to the pooled base-specific termination reactions of 24 individuals), 1 µl each, were loaded onto an ultrathin, 0.125 mm, 5% polyacrylamide gel, resolved by size at 3150 V, transferred onto a nylon membrane (32 × 45 cm; neutral 1.2 µm Biodyne A; Pall, Dreieich, Germany) by DTE (55), following a modified speed gradient to guarantee band distances of ~0.8 mm. Membranes were cross-linked by UV light for 30 s.

Sequential probing

Oligonucleotides were end-labeled according to the manufacturer's recommendations (USB, Braunschweig, Germany). After prehybridization at 42°C for at least 15 min, one terminal transferase reaction was diluted into 8 ml of hybridization buffer [7% SDS, 10% polyethylene glycol (PEG), 0.25 M NaCl, 0.051% H_3PO_4, 82 mM $Na_2HPO_4{\cdot}2H_2O$, 10 mM EDTA (free acid), 32 mM NaOH] and incubated at 42°C overnight. The membranes were washed at room temperature in 1% SDS, 0.022% H_3PO_4, 69 mM $NaH_2PO_4{\cdot}H_2O$, 5 mM EDTA (free acid), 32 mM NaOH by use of an automated device (Umweltund Ingenieurtechnik, Dresden, Germany). Membranes were exposed to Phospho-Fluor-Screens up to 24 h at room temperature and scanned (Phospho-Fluor-Imager; Molecular Dynamics, Krefeld, Germany). Radioactive probes were removed with 0.2% SDS and 2 mM EDTA pH 8.3, at least twice for 5–10 min at 65°C.

Allele-specific PCR (AS-PCR)

In order to test the hypothesis that the combination of five variants, which constitute the pattern, specifically variants at nucleotide positions –1793, –1699, –1320, –111 and +17 (A6V), may reflect a genetic haplotype, a series of AS-PCR experiments was performed. The PCR product generated from the first PCR was used as template with two (reference sequence and mutant) reverse primers for a second round of PCR, then these new products were used as templates for a third round with two new (reference sequence and mutant) reverse primers, etc. First, a large PCR fragment of ~2470 bp including all five polymorphic sites was generated according to the PCR conditions described earlier as part of the MSC protocol; forward and reverse primers were MOR1X-108F (5′-GGACTTTCATTG-TACTGGTAGA-3′) and Intron1R (5′-TTACCTGACAATC-ACATACATGAC-3′). Ten nanograms of this PCR product was then assayed with 20 pmol primers MOR1X-108F and the reverse primers MOR1X+17 (5′-GCTGGCGTTCGTGGG-GG/A-3′) with either G or A at its 3′ end. AS-PCR reactions were performed at 68°C annealing temperature and conditions as described earlier. The PCR products obtained with both primers MOR1X+17 (5′-GCTGGCGTTCGTGGGGG/A-3′) were then used separately for AS-PCR reactions utilizing sequentially as reverse primers with variable positions at their 3′ ends: –111(G/A), –1320(T/C), –1699+T(G/T), –1793(T/A). Again, at each step, the PCR products were used separately, using as template 1 µl of the PCR product, which had been diluted 5–12 times. For further details regarding sequence-specific PCR reactions see Sander *et al.* (56). The DNA from individuals carrying the pattern was shown to amplify consistently, when PCR products generated with the mutant reverse primers were used as templates for amplification with the other mutant reverse primers.

Image analysis and genotyping

Computer images were genotyped on PCs by skilled pattern analysis independently by two expert readers, who were blind to diagnoses. Variation data from both strands, each read by the two readers, were compared and scored, if totally consistent. If not, data were checked for transfer or reading errors. In case the genotype could not be consistently resolved, the DNA segment was resequenced. Each genotype entering final analysis was based on a minimum of three consistent scores.

Haplotype analysis

To predict the most likely haplotype pair for each genotype in a given sample, we have developed and applied in this study the program MULTIHAP (http://mahe.bioinf.mdc-berlin.de) which derives the maximum likelihood estimation of the underlying haplotype frequencies by an EM algorithm, which is similar to that of Excoffier and Slatkin (57) or Hawley and Kidd (58). Besides giving the maximum likelihood estimation of the haplotype frequencies in the population, MULTIHAP has been specifically designed to provide access to the haplotype decomposition of the sample. This forms the basis for an optimal choice of the initial haplotype frequencies as starting points for the EM iterations, which reduces computing time considerably compared with other programs (57,58) and predicts the most likely haplotype pair for each genotype after maximum likelihood estimation. The haplotype frequencies estimated for our sample by application of MULTIHAP were cross-checked against the results obtained by application of the program of Excoffier and Slatkin (57) to this sample; the results were in excellent agreement. Our haplotype program uses similar algorithms for the estimation of haplotype frequencies in a population; it differs from the other haplotype programs in that it predicts the most likely haplotype pair for each genotype in a sample that means that we obtain two haplotypes for each genotype. The prediction of these haplotype pairs is necessary for our subsequent analysis.

Classification of haplotypes (genotypes)

Haplotypes (genotypes) were progressively classified by means of the procedure CLUSTER (method = BAVERAGE, measure = SEUCLID) from the package SPSS for Windows NT, release 7.5. This is a hierarchical cluster analysis procedure, which uses the squared Euclidean distance as measure and the between-groups linkage as the cluster method. The variants were coded with 1 (identical with the reference sequence) or 2 (different from the reference sequence) in a haplotype, with 1 (homozygous for nucleotides/sequences identical with the reference sequence), 2 (heterozygous) or 3 (homozygous for nucleotides/sequences different from the reference sequence) in a genotype, respectively. This procedure is based on an algorithm that starts with each haplotype (independent of phenotype) in a separate cluster and merges step by step the two most similar clusters until one final cluster is left. The results of the cluster analysis are illustrated by a dendrogram.

Construction of an evolutionary tree

Construction was performed by means of the distance-based neighbor-joining algorithm described by Saitou and Nei (37), using the computer program NJ.PAS (http://smiler.lab.nig.ac.jp). The distance matrix was derived from the unweighted pair-wise Hamming distances between the haplotypes, measuring the number of different variants. The principle of this method is to identify pairs of neighbors that minimize the total branch length at each stage of clustering, starting with a star-like tree. The branch lengths as well as the topology of a parsimonious tree can rapidly be obtained by application of this method. Specifically, this method is not based on a substitution model. Therefore, the NJ method has been often found to be superior to the other established tree reconstruction methods (59). Moreover, this method benefits from its high computational efficiency.

Association analysis and identification of risk patterns

Taking the results of the CLUSTER procedure as a starting point, a series of $(2 \times n_i)$-contingency tables were analyzed successively, each table corresponding to exactly one classification step of the i classification steps, each of these defined by an equal distance cluster coefficient; n_i symbolizes the number of classes in the i-th classification step. In each table, the haplotype frequencies of the two different phenotypic groups (cases and controls) in the n_i classes were compared and a likelihood ratio χ^2 statistic was calculated. If necessary, exact tests (60) were performed to calculate significance levels by means of the program package SPSS for Windows NT version 7.5 with the EXACT test procedure. The exact test calculates an accurate significance level for sparse and unbalanced tables. It uses network algorithms based on the methods of Mehta and Patel (61–63). The results of the statistical evaluation at each iterative step are given as P-values. At the stage of maximum separation of phenotypes within the classes, the highest significance level should be reached. If at least two classes differed significantly, the clustered haplotypes were inspected for consensus patterns. Subsequently, the extracted consensus patterns were analyzed for association, using likelihood ratio χ^2 statistics. In these analytical procedures, corrections for multiple testing seem not to be required, because in the process of step-wise testing related to the iterative clustering steps each step depends on the preceding one as part of the same clustering process; this is no series of independent tests of the same null hypothesis in different data sets. In addition, an extensive computer simulation, which duplicated our experiment, was performed in order to examine whether our result could have been obtained by chance. We generated 10 000 repetitions of our experiment, randomizing the affection status (case or control). Randomization testing is a way of determining whether the null hypothesis is reasonable in this situation. In each of these 10 000 experiments, we have randomly reassigned 70 of the available 344 haplotypes to controls. Then we analyzed the series of $(2 \times n_i)$-contingency tables successively, each table corresponding to exactly one of the 20 classification steps in our experiment. In each table, the haplotype frequencies of the two different phenotypic groups in the n_i clusters were compared and a likelihood ratio χ^2 statistic was calculated; if necessary, the exact test was applied. If the exact test in tables with 23 or more clusters was not computable, the asymptotic statistic was calculated.

Then we counted all those experiments of the 10 000 for which the following criteria were met. (i) One or more contin-

gency tables were significant at the 5% level for the exact statistic, or at the 10% level for the asymptotic χ^2 statistics (if 23 or more clusters were given). The table for the next clustering step (if at least three clusters existed at a step) would not reach a significance level of 30% (because mixing a pattern-defining cluster with any other clusters should lead to a non-significant clustering step). (ii) At least one of the clusters in such a significant table contained more cases than it should in the case of independence, and the likelihood ratio χ^2 statistic for the consensus pattern from such a cluster reached the 5% level.

The simulation experiments generated at any step any possible cluster with any possible pattern (which could consist of any possible number of positions), without any restriction to the pattern. Two hundred and ninety-four of the 10 000 experiments (2.94%) fulfill the above-described conditions; of these, 105 showed a significant result at step 20, as was observed in our μ opioid receptor gene study. This is equivalent to a probability of 1.05% that our result could have been generated by chance. This corresponds well with the *P*-value of 0.017 derived by our approach.

Data analysis

Allele and genotype frequencies were calculated by application of the program package SPSS for Windows NT version 7.5. Hardy–Weinberg equilibrium was checked using the χ^2 test. All clustering methods applied to this data set were performed with the SPSS package.

ACKNOWLEDGEMENTS

The authors particulary thank W. Schmidt, GenProfile AG, for performing the cladistic analysis and W. Terhalle, GenProfile AG, for performing the simulation study as well as for helpful discussions. The authors are particularly grateful for the valuable contributions of H. Lehrach and R. Reinhard, Max-Planck-Institute for Molecular Genetics, Berlin. The authors are also grateful to M. Vingron, German Cancer Research Institute, Heidelberg, for advice in simulation studies. They also thank J. Ott, Laboratory of Statistical Genetics, Rockefeller University, New York, for helpful discussions. The authors acknowledge V. Höllt for sharing information on transcription regulatory motif sequences. The authors acknowledge the excellent technical assistance of I. Grunewald and P. Heere, as well as valuable help through B. Timmermann, L. Ohl and N. Gscheidel. The authors wish to acknowledge E. Gottheil, N. Wintering, P. DeMaria, S. Weinstein, C. Kaltenbach, C. Berrettini and A. Patkar for their help in patient recruitment. This project was supported by DOE grant DE-FG02-87ER60565 to G.M.C., by German Human Genome grant 01KW9607/0 and an Ernst Schering Research Foundation Award Scholarship to M.R.H., by funding from the European Social Funds (ESF) to K.K. and by NIH grant DA11835 to W.H.B.

REFERENCES

1. Collins, F.S. (1995) Positional cloning moves from perditional to traditional. *Nature Genet.*, **9**, 347–350.
2. Collins, F.S., Guyer, M.S. and Chakravarti, A. (1997) Variations on a theme: cataloging human DNA sequence variation. *Science*, **278**, 1580–1581.
3. Risch, N. and Merikangas, K. (1996) The future of genetic studies of complex human diseases. *Science*, **273**, 1516–1517.
4. Lander, E.S. (1996) The new genomics: global views of biology. *Science*, **274**, 536–539.
5. Chakravarti, A. (1998) It's raining SNPs, hallelujah? *Nature Genet.*, **19**, 216–217.
6. Nickerson, D.A., Taylor, S.L., Weiss, K.M., Clark, A.G., Hutchinson, T.G., Stengard, J., Salomaa, V., Vartiainen, E., Boerwinkle, E. and Sing, C.F. (1998) DNA sequence diversity in a 9.7 kb region of the human lipoprotein lipase gene. *Nature Genet.*, **19**, 233–240.
7. Timmermann, B., Mo, R., Luft, F.C., Gerdts, E., Busjahn, A., Omvik, P., Li, G.-H., Schuster, H., Wienker, T.F., Hoehe, M.R. *et al.* (1998) Beta-2 adrenoceptor genetic variation is associated with genetic predisposition to essential hypertension: the Bergen Blood Pressure Study. *Kidney Int.*, **53**, 1455–1460.
8. Pang, H., Koda, Y., Soejim, M., Schlaphoff, T., Du-Toit, E.D. and Kimura, H. (1999) Allelic diversity of the human plasma α-(1,3) fucosyltransferase gene (*FUT6*). *Ann. Hum. Genet.*, **63**, 277–284.
9. Fan, F., Liu, Ch., Tavaré, S. and Arnheim, N. (1999) Polymorphisms in the human DNA repair gene *XPF*. *Mutat. Res. Genomics*, **406**, 115–120.
10. Hoehe, M.R., Wendel, B., Köpke, K., Flachmeier, C., Kieffer, B.L., Berrettini, W.H. and Church, G.M. (1998) Genetic variability of the human mu opioid receptor gene and its implication for substance abuse. *Am. J. Med. Genet.*, **81**, 507.
11. Rieder, M.J., Taylor, S.L., Clark, A.G. and Nickerson, D.A. (1999) Sequence variation in the human angiotensin converting enzyme. *Nature Genet.*, **22**, 59–62.
12. Weiss, K.M. (1996) Is there a paradigm shift in genetics? Lessons from the study of human diseases. *Mol. Phylogenet. Evol.*, **5**, 259–265.
13. Terwilliger, J.D. and Weiss, K.M. (1998) Linkage disequilibrium mapping of complex disease: fantasy or reality? *Curr. Opin. Biotechnol.*, **9**, 578–594.
14. Weiss, K.M. (1995) *Genetic Variation and Human Disease: Principles and Evolutionary Approaches.* Cambridge University Press, Cambridge, UK.
15. Clark, A.G., Weiss, K.M., Nickerson, D.A., Taylor, S.L., Buchanan, A., Stengård, J., Salomaa, V., Vartiainen, E., Perola, M., Boerwinkle, W.H. *et al.* (1998) Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. *Am. J. Hum. Genet.*, **63**, 595–612.
16. Willard, H.F. and Davies, K.E. (1998) Genetics of disease. Complex genetics, complex diseases. *Curr. Opin. Genet. Dev.*, **8**, 271–273.
17. Matthes, H.W., Maldonado, R., Simonin, F., Valverde, O., Slowe, S., Kitchen, I., Befort, K., Dierich, A., Le-Meur, M., Dolle, P. *et al.* (1996) Loss of morphine-induced analgesia, reward effect and withdrawal symptoms in mice lacking the mu-opioid-receptor gene. *Nature*, **383**, 819–823.
18. Kieffer, B.L. (1999) Opioids: first lessons from knockout mice. *Trends Phys. Sci.*, **20**, 19–26.
19. Hoehe, M., Duka, T., Doenicke, A. and Matussek, N. (1984) Dose-dependent influence of fentanyl on prolactin, growth hormone, and mood. *Neuropeptides*, **5**, 269–272.
20. Kreek, M.J. (1996) Opiates, opioids and addiction. *Mol. Psychiatry*, **1**, 232–254.
21. Herz, A. (1997) Endogenous opioid systems and alcohol addiction. *Psychopharmacology*, **129**, 99–111.
22. Berrettini, W.H., Ferraro, T.N., Alexander, R.C., Buchberg, A.M. and Vogel, W.H. (1994) Quantitative trait loci mapping of three loci controlling morphine preference using inbred mouse strains. *Nature Genet.*, **7**, 54–58.
23. Lander, E.S. and Schork, N.J. (1994) Genetic dissection of complex traits. *Science*, **265**, 2037–2048.
24. Wendel, B. and Hoehe, M.R. (1998) The human mu opioid receptor gene: 5′ regulatory and intronic sequences. *J. Mol. Med.*, **76**, 525–532.
25. Church, G.M. and Kieffer-Higgins, S. (1988) Multiplex DNA sequencing. *Science*, **240**, 185–188.
26. Berrettini, W.H., Hoehe, M.R., Ferraro, T.N., DeMaria, P.A. and Gottheil, E. (1997) Human mu opioid receptor gene polymorphisms and vulnerability to substance abuse. *Addiction Biol.*, **2**, 303–308.
27. Bergen, A.W., Kokoszka, J., Peterson, R., Long, J.C., Virkkunen, M., Linnoila, M. and Goldman, D. (1997) Mu opioid receptor gene variants: lack of association with alcohol dependence. *Mol. Psychiatry*, **2**, 490–494.
28. Bond, C., LaForge, K.S., Tian, M., Melia, D., Zhang, S., Borg, L., Gong, J., Schluger, J., Strong, J.A., Leal, S.M. *et al.* (1998) Single-nucleotide polymorphism in the human mu opioid receptor gene alters beta-endorphin binding and activity: possible implications for opiate addiction. *Proc. Natl Acad. Sci. USA*, **95**, 9608–9613.

29. Mestek, A., Hurley, J.H., Bye, L.S., Campbell, A.D., Chen, Y., Tian, M., Liu, J., Schulman, H. and Yu, L. (1995) The human mu opioid receptor: modulation of functional desensitization by calcium/calmodulin-dependent protein kinase and protein kinase C. *J. Neurosci.*, **15**, 2396–2406.
30. Sirand-Pugnet, P., Durosay, P., Brody, E. and Marie, J. (1995) An intronic (A/U)GGG repeat enhances the splicing of an alternative intron of the chicken beta-tropomyosin pre-mRNA. *Nucleic Acids Res.*, **23**, 3501–3507.
31. Yan, H., Papadopoulos, N., Marra, G., Perrera, C., Jiricny, J., Boland, C.R., Lynch, H.T., Chadwick, R.B., de la Chapelle, A., Berg, K. *et al.* (2000) Conversion of diploidy to haploidy. *Nature*, **403**, 723–724.
32. Nei, M. (1987) *Molecular Evolutionary Genetics.* Columbia University Press, New York, NY.
33. Templeton, A.R., Boerwinkle, E. and Sing, C.F. (1987) A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. I. Basic theory and an analysis of alcohol dehydrogenase activity in *Drosophila*. *Genetics*, **117**, 343–351.
34. Sing, C.F., Havilland, M.B., Zerba, K.E. and Templeton, A.R. (1992) Application of cladistics to the analysis of genotype–phenotype relationships. *Eur. J. Epidemiol.*, **8** (Suppl. 1), 3–9.
35. Valdes, A.M. and Thomson, G. (1997) Detecting disease-predisposing variants: the haplotype method. *Am. J. Hum. Genet.*, **60**, 703–716.
36. Knapp, M. (1991) *Statistische Methoden zur Assoziations- und Linkage-Analyze bei Kernfamilien.* [*Statistical methods for association and linkage analysis in nuclear families.*] Thesis, University of Bonn, Germany.
37. Saitou, N. and Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**, 406–425.
38. Arabie, P. and Hubert, L.J. (1996) An overview of combinatorial data analysis. In Arabie, P., Hubert, L.J. and De Soete, G. (eds), *Clustering and Classification.* World Scientific, Singapore, New Jersey, London, Hong Kong, pp. 5–63.
39. Templeton, A.R. (1995) A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping or DNA sequencing. V. Analysis of case/control sampling designs: Alzheimer's disease and the apoprotein E locus. *Genetics*, **140**, 403–409.
40. Templeton, A.R. (1996) Cladistic approaches to identifying determinants of variability in multifactorial phenotypes and the evolutionary significance of variation in the human genome. In: Variation in the human genome. *Ciba Found. Symp.*, **197**, 259–283.
41. Schork, N.J. and Schork, C.M. (1998) Issues and strategies in the genetic analysis of alcoholism and related addictive behaviors. *Alcohol*, **16**, 71–83.
42. Kidd, K.K. (1993) Associations of disease with genetic markers: déjà vu all over again. *Am. J. Med. Genet.*, **48**, 71–73.
43. Carey, G. (1994) Genetic association study in psychiatry: analytical evaluation and a recommendation. *Am. J. Med. Genet.*, **54**, 311–317.
44. Berrettini, W.H. and Persico, A.M. (1996) Dopamine D2 receptor gene polymorphisms and vulnerability to substance abuse in African Americans. *Biol. Psychiatry*, **40**, 144–147.
45. Barr, C.L. and Kidd, K.K. (1993) Population frequencies of the A1 allele at the dopamine D2 receptor locus. *Biol. Psychiatry*, **34**, 204–209.
46. Delbrück, S.J.W., Wendel, B., Grunewald, I., Sander, T., Morris-Rosendahl, D., Crocq, M.A., Berrettini, W.H. and Hoehe, M.R. (1997) A novel allelic variant of the human serotonin transporter gene regulatory polymorphism. *Cytogenet. Cell Genet.*, **79**, 214–220.
47. Tishkoff, S.A., Dietzsch, E., Speed, W., Pakstis, A., Kidd, J.R., Cheung, K., Bonne-Tamir, B., Santachiara-Benerecetti, A.S., Moral, P., Krings, M. *et al.* (1996) Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. *Science*, **271**, 1380–1387.
48. Uhl, G.R., Sora, I. and Wang, Z. (1999) The μ opiate receptor as candidate gene for pain: polymorphisms, variations in expression, nociception, and opiate responses. *Proc. Natl Acad. Sci. USA*, **96**, 7752–7755.
49. Petruzzi, R., Ferraro, T.N., Kürschner, V.C., Golden, G.T. and Berrettini, W.H. (1997) The effects of repeated morphine exposure on μ opioid receptor number and affinity in C57BL/6J and DBA/2J mice. *Life Sci.*, **61**, 2057–2064.
50. Zubieta, J.K., Gorelick, D.A., Stauffer, R., Ravert, H.T., Dannals, R.F. and Frost, J.J. (1996) Increased μ opioid receptor binding detected by PET in cocaine-dependent men is associated with cocaine craving. *Nature Med.*, **2**, 1225–1229.
51. Volpicelli, J.R., Alterman, A.I., Hayashida, M. and O'Brien, C.P. (1992) Naltrexone in the treatment of alcohol dependence. *Arch. Gen. Psychiatry*, **49**, 876–880.
52. King, A.C., Volpicelli, J.R., Frazer, A. and O'Brien, C.P. (1997) Effect of naltrexone on subjective alcohol response in subjects at high and low risk for future alcohol dependence. *Psychopharmacology*, **129**, 15–22.
53. Mazure, C. and Gershon, E.S. (1979) Blindness and reliability in lifetime psychiatric diagnosis. *Arch. Gen. Psychiatry*, **36**, 521–525.
54. Lahiri, D.K. and Nurnberger Jr, J.I. (1991) A rapid non-enzymatic method for the preparation of HMW DNA from blood for RFLP studies. *Nucleic Acids Res.*, **19**, 5444.
55. Richterich, P. and Church, G.M. (1993) DNA sequencing with direct transfer electrophoresis and nonradioactive detection. *Methods Enzymol.*, **218**, 187–222.
56. Sander, T., Gscheidl, N., Wendel, B., Samochowiec, J., Smolka, M., Rommelspacher, H., Schmidt, L.G. and Hoehe, M.R. (1998) Human μ-opioid receptor variation and alcohol dependence. *Alcohol Clin. Exp. Res.*, **22**, 2108–2110.
57. Excoffier, L. and Slatkin, M. (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol. Biol. Evol.*, **12**, 921–927.
58 Hawley, M.E. and Kidd, K.K. (1995) HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes. *J. Hered.*, **86**, 409–411.
59. Hasegawa, M. and Fujiwara, M. (1993) Relative efficiencies of the maximum likelihood, maximum parsimony, and neighbor-joining methods for estimating protein phylogeny. *Mol. Phylogenet. Evol.*, **2**, 1–5.
60. Agresti, A., Mehta, C.R. and Patel, N.R. (1990) Exact inference for contingency-tables with ordered categories. *J. Am. Stat. Assoc.*, **85**, 453–458.
61. Mehta, C.R. and Patel, N.R. (1983) A network algorithm for performing Fisher's exact test in r × c contingency tables. *J. Am. Stat. Assoc.*, **78**, 427–434.
62. Mehta, C.R. and Patel, N.R. (1986) A hybrid algorithm for Fisher's exact test on unordered r × c contingency tables. *Commun. Stat.*, **15**, 387–403.
63. Mehta, C.R. and Patel, N.R. (1986) FEXACT: a FORTRAN subroutine for Fisher's exact test on unordered r × c contingency tables. *ACM Transact. Math. Software*, **12**, 154–161.
64. Wang, J.B., Johnson, P.S., Persico, A.M., Hawkins, A.L., Griffin, C.A. and Uhl, G.R. (1994) Human μ opiate receptor: cDNA and genomic clones, pharmacologic characterization and chromosomal assignment. *FEBS Lett.*, **338**, 217–222.
65. Auffray, C., Lillie, J.W., Korman, A.J., Boss, J.M., Frechin, N., Guillemot, F., Cooper, J., Mulligan, R.C. and Strominger J.L. (1987) Structure and expression of HLA-DQ alpha and -DX alpha genes: interallelic alternate splicing of the HLA-DQ alpha gene and functional splicing of the HLA-DQ alpha gene using a retroviral vector. *Immunogenetics*, **26**, 63–73.
66. Antonarakis, S.E. and the Nomenclature Working Group (1998) Recommendations for a nomenclature system for human gene mutations. *Hum. Mutat.*, **11**, 1–3.
67. Beutler, E., McKusick, V.A., Motulsky, A.G., Scriver, C.R. and Hutchinson, F. (1996) Mutation nomenclature: nicknames, systematic names and unique indentifiers. *Hum. Mutat.*, **8**, 203–206.
68. Ad Hoc Committee on Mutation Nomenclature (1996) Update on nomenclature for human gene mutations. *Hum. Mutat.*, **8**, 197–202.