ELSEVIER

Biological Psychology 61 (2002) 53–71

BIOLOGICAL PSYCHOLOGY

www.elsevier.com/locate/biopsycho

Gene finding strategies

Jacqueline M. Vink *, D.I. Boomsma

Department of Biological Psychology, Free University of Amsterdam, Van der Boechorstraat 1, 1081 BT Amsterdam, The Netherlands

Abstract

Both linkage and association methods have been used to localise and identify genes related to behaviour and other complex traits. The linkage approach (parametric or non-parametric) can be used for whole genome screens to localise genes of unknown function. The parametric linkage approach is very effective for locating single-gene disorders and is usually based on large family pedigrees. The non-parametric method is useful to detect quantitative trait loci (QTLs) for complex traits and was originally developed for sib pair analyses. Genetic association studies are most often used to test the association of alleles at a candidate gene with a disease or with levels of a quantitative trait. Allelic association between a trait and a marker can be studied in a case–control design, but because of possible problems due to population stratification, within-family designs have been proposed as the optimal test for association.

Keywords: (Non) parametric linkage; Genetic association; Quantitative trait loci; Population stratification; Candidate genes

1. Introduction

Behavioural genetic studies (including twin studies) have shown that genetic influences often contribute to individual differences in behaviour. Behavioural traits are complex, reflecting the aggregate effect of multiple, possibly interacting genetic and environmental determinants. Molecular genetic methods have been applied to complex and quantitative traits trying to identify genes responsible for the moderate to high heritabilities seen for behavioural traits (e.g. Gayan and Olson, 1999; Gershon, 2000; Faraone and Doyle, 2001), but with the availability of relatively

* Corresponding author. Tel.: +31-20-444-8787; fax: +31-20-444-8832
E-mail address: jm.vink@psy.vu.nl (J.M. Vink).

0301-0511/02/$ - see front matter
PII: S0301-0511(02)00052-2

cheap and easy DNA marker typing, many more molecular genetics studies of behavioural traits can be expected in the next few years. In this respect, the completion of the human genome sequence will be valuable in locating and identifying genes involved in human behaviour (International Human Genome Sequencing Consortium, 2001; International SNP Map Working Group, 2001; Peltonen and McKusick, 2001; Venter et al., 2001). This article reviews strategies for gene finding in humans, especially linkage and association methods. The gene finding strategies that we will discuss have been mostly applied to disease phenotypes. They are, however, increasingly applied to behaviour in a broader sense. We will illustrate gene finding for behavioural traits with examples of phenotypes taken from various research fields such as addiction and personality and psychophysiological traits considered to be risk markers or risk factors for disease. We will discuss EEG power as one of the examples, and other papers in this special issue will provide a number of further examples (e.g. Busjahn et al., this issue; Porjesz et al., this issue; Snieder et al., this issue). Variations and extensions of linkage and association methods are summarised and combined linkage and association tests are introduced as a tool for testing for genuine associations, as well as for fine mapping of broad linkage regions.

2. Linkage

Genes contribute to variation in both normal behaviour and behavioural disorders (Sullivan and Kendler, 1999; Plomin and Crabbe, 2000; Plomin et al., 2000; Bouchard and Loehlin, 2001). Some disorders have a simple Mendelian mode of transmission in which a specific mutation confers the certainty of developing the disorder, in other words a single gene is responsible for the disorder. Many single gene diseases and disorders are listed in full in the 'Mendelian Inheritance in Man' (McKusick, 1998) and its freely available online version (ftp://www.ncbi.nlm.nih.-gov/omim; updated every day). A general strategy to find genes for Mendelian traits is called classical linkage and is based on Fisher's theory of likelihood inference (Fisher, 1918). It is referred to as being parametric or model-based because an explicit genetic model for the disease or trait locus has to be provided. Classical linkage analysis models the distance between a DNA marker locus and a putative disease locus in small numbers of large multigenerational families (pedigrees) consisting of both affected and unaffected family members. It is the method of choice for the genetic mapping of single-gene diseases, especially when these diseases are rare. Classical linkage requires that a model for the disease or trait locus is specified a priori, in terms of allelic frequencies, penetrance and mode of action (recessive or dominant). Complete penetrance implies that all individuals with a high-risk genotype (genotype dd in the case of a recessive disorder and genotypes Dd and DD in the case of a dominant disorder) will develop the disorder. If there are individuals with a high-risk genotype who do not develop the disease, then the penetrance of the genotype is said to be incomplete. Individuals without a high-risk

genotype who develop a disorder that is phenotypically indistinguishable from the genetic form, are called phenocopies (Sham, 1998).

In linkage analysis a number of DNA markers of known location, evenly dispersed throughout the entire genome, are measured in individuals from multiple generations. DNA markers can be mutations in a single base pair (Single Nucleotide Polymorphisms (SNPs)) or a variable number of repeats of two or more base pairs (microsatellites), as described in more detail by Slagboom and Meulenbelt (in this issue). They need not to be part of a functional gene—they are just landmarks of known location in the genome. For each DNA marker, evidence for linkage is derived using statistical procedures that trace the co-segregation of the trait (and thus in many instances the gene) and a specific variant of the DNA marker along familial lineages in extended pedigrees. The genetic distance between a marker locus of known position and a disease/trait locus (of unknown position) is estimated by observing the segregation of the marker locus in a pedigree together with the disease status/trait (Fig. 1). The second law of Mendel states that the inheritance of one gene is not affected by the inheritance of another gene (law of independent assortment). This law applies if two loci are on different chromosomes or are far apart on the same chromosome, because recombination between the loci will prevent alleles from being transmitted together. The closer two loci are on the same chromosome, the less likely crossovers during meiosis will be and the fewer recombinants will be observed in the offspring. Starting with the known position of a marker locus, it can thus be tested whether another locus is genetically close (linked) by counting the number of recombinations that occurred between both loci in a given number of meioses. The probability that two alleles at different loci on the same chromosome are derived from different parental chromosomes (i.e. recombinant) is called the recombination fraction. The recombination fraction ranges from $\theta = 0$ (tight linkage) to $\theta = 0.5$ (no linkage). If the loci are tightly linked, alleles from both loci are always inherited together in a pedigree. The recombination fraction can be taken as a measure of the genetic distance, or map distance, between gene loci. The unit of measurement is 1 map unit or 1 centimorgan (cM), corresponding approximately to a recombination fraction of 1%.

In parametric linkage analysis, it is standard practice to summarise the results of a linkage analysis in the form of a LOD score function (Morton, 1955). LOD score stands for the logarithm of the odds that the locus is linked to the trait and indicates the strength of the linkage (Fig. 2). LOD scores are expressed according to the following equation:

$$\text{LOD score} =^{10} \log \frac{\text{Likelihood of the observed genotypes given } \theta \text{ is less than 0.5 (linkage)}}{\text{Likelihood of the observed genotypes given } \theta = 0.5 \text{ (no linkage)}}$$

Evidence for linkage is said to be present when the maximal LOD-score exceeds a pre-defined threshold, which depends on the size of the genome and the number of markers (Lander and Kruglyak, 1995). The LOD-score is a function of the unknown recombination fraction θ. It is customary to plot LOD score against different recombination fractions in order to obtain an impression of the relative support for different values of the recombination fraction and thus the distance between the

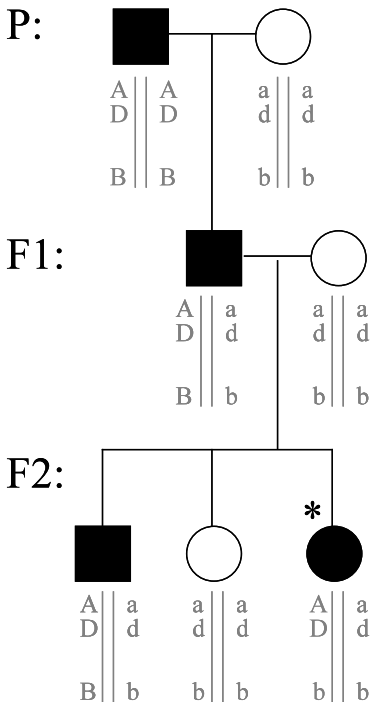

Fig. 1. Linkage and recombination. Artificial family with assumed disease locus. P, parents; F1, first generation offspring; F2, second generation offspring; square, male; circle, female. The disease locus carries a variation which is reflected in D or d; when having allele D, the individual is affected and having allele d, the individual is unaffected. The affected individuals are shown in black and the unaffected individuals in white. There are two markers; marker 1 with polymorphism A or a and marker 2 with polymorphism B or b. Both parents are doubly homozygous: the father is homozygous for the A and B marker (haplotype AB), and the mother is homozygous for the a and b variant of the marker (haplotype ab). The son in the first generation has received marker A and B from the father and a and b from the mother (who newly enters the pedigree at F1). In the F2 generation recombination has occurred (individual is marked with an *), this individual carries haplotype Ab. The other two individuals are non-recombinant. The recombination fraction is the number of recombinations divided by the total number of meioses is, so when recombination occurs in one of the 100 meioses, the recombination fraction is 1%. In this example, full linkage will be found between marker A and the disease/trait (they always co-segregated together), but less so between marker B and the disease/trait (they co-segregate in only two of three meisoses).

marker and the disease locus. The chromosomal region surrounding a marker with a significantly high LOD-score under the optimal recombination fraction will be selected for fine-mapping, which is essentially a repetition of the same procedure but now with many additional markers concentrated in the area of interest on a single chromosome. If the region containing the putative gene is sufficiently small, the DNA in the entire region is sequenced in full to find genetic variants (polymorphisms). The next step could be an association study (described below). The entire process from significant LOD scores to the actual allelic variants is usually

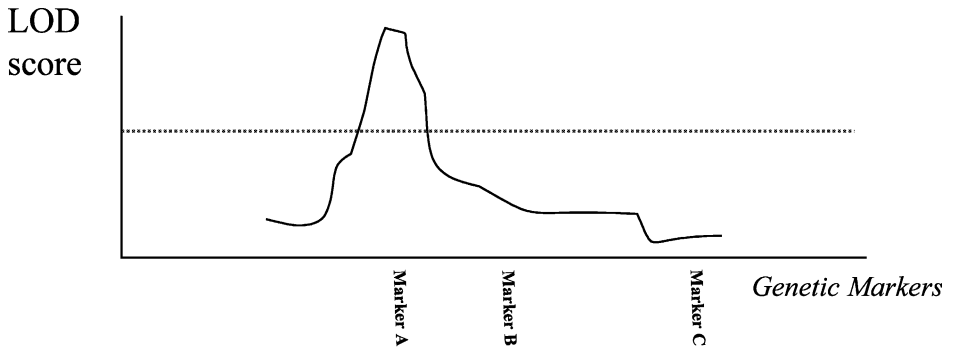

Fig. 2. LOD-score. A significant LOD-score will be found if the locus of a marker (in this example marker A) is linked to the trait. The higher the LOD-score the tighter the linkage. The chromosomal region surrounding a marker with a significantly high LOD-score will be selected for fine-mapping. As the disease locus significantly co-segregates with marker A, the position of the disease locus is probably (relatively) close to the locus of marker A.

summarised as '*positional cloning*'. The circa 1500 disease genes now listed in the Online Mendelian Inheritance in Man catalogue have largely been detected by this process.

The identification of a locus for the human low-voltage EEG on chromosome 20q is an early example of a classical linkage approach. The inter individual variability of the human EEG is largely determined genetically (Beijsterveld et al., 1994; Vogel and Motulsky, 1997; Beijsterveld and van Baal, this issue). Some EEG variants were shown to follow a simple mode of inheritance. In the case of the low-voltage EEG the familial transmission pattern was found to follow an autosomal dominant mode of inheritance. Steinlein et al. (1992) studied a total of 22 blood and serological markers (as a proxy for the underlying polymorphisms) and 73 DNA markers (restriction fragment length polymorphisms (RFLPs)) in 17 families with 191 individuals. The markers were distributed over all autosomal chromosomes. The frequency of the low-voltage EEG allele was taken to be 0.02. An autosomal dominant mode of inheritance with full penetrance was assumed. Linkage analysis carried out for all families pooled together yielded no significant evidence for linkage. This 'null' finding provided a nice example of a nasty complexity that may arise in linkage analyses when variations in recombination fraction occur because of the existence of multiple disease loci. A marker that is close to a particular disease locus will demonstrate linkage in families where the disease is caused by alleles at that locus. In other families, in which the disease is caused by alleles at other loci, the marker will show no linkage with the disease. This heterogeneity is known as *locus heterogeneity*. The results of the Steinlein study provided evidence for locus heterogeneity with respect to the low-voltage EEG variant. One of the markers, the CMM6 (D20S19), localised on the distal part of chromosome 20q showed linkage in some of the families (maximum LOD score 3.13 and recombination fraction 0) and exclusion of linkage in the other families. In short, two types of families were found: with and without linkage to chromosome 20q. Within the first

type of family the autosomal dominant inherited low-voltage EEG is determined by a gene located close to the highly polymorphic marker CMM6 on chromosome 20q. In the second type of family this phenotype is caused by another gene, or genes, located elsewhere (Anokhin et al., 1992; Steinlein et al., 1992).

3. Non-parametric linkage

Most complex traits are multifactorial, i.e. they are influenced by a number of different genes, environmental factors, their possible interactions, and possibly a third source of variation that consist of nonlinear epigenetic processes (Molenaar et al., 1993). Traits that are influenced by the developmental interplay of many genes and environmental factors are usually quantitative traits, and each of the genes that influence such quantitative traits is called a polygene. The chromosomal region (or locus) where such a polygene can be found is called a *quantitative trait locus* (QTL). Typically, the word 'quantitative' is used when 'continuous' is meant, and variation in the phenotype shows a normal distribution. However, for some quantitative traits the scale of measurement can also be discrete. In the case of a binary disease phenotype (affected/unaffected) the penetrance, or probability of being affected, is often transformed to a probit (or logit), giving rise to what is called the 'liability' to disease. This liability can be thought of as the underlying vulnerability to the disease and is treated as a continuous phenotype (Falconer and Mackay, 1996; Elston, 2000).

To detect QTLs, non-parametric or model free linkage analysis uses a similar linkage concept as described above, but unlike parametric linkage, no explicit model of the disease is required for this type of genomic search. Non-parametric methods were originally developed for sibling pairs but have been extended to general pedigrees. In this kind of analysis, several hundreds of DNA markers are obtained from siblings and (optimally) their parents and allele sharing between siblings (or other relatives) is investigated. There are two definitions of allele sharing, *identity-by-state* (IBS) and *identity-by-descent* (IBD). Two alleles of the same form (i.e. having the same DNA sequence) are said to be IBS. If, in addition to being IBS, two alleles are descended from the same ancestral allele, then they are said to be IBD. Full siblings both receive an allele from the father and an allele from the mother. Let the variable D_f be 1 when both siblings have received the same paternal allele, and 0 otherwise. Similarly, let the variable D_m be 1 if the two siblings have received the same maternal allele, and 0 otherwise. The total IBD value of the sibling pair, D, is defined as the sum of D_f and D_m and, therefore, can be 0,1,2 with the probabilities 1/4, 1/2, 1/4, respectively (Fig. 3). Linkage of a marker to a QTL implies that the differences in the trait between the relative pairs will be smaller if they share the same variant of the marker, obtained from the same ancestor (IBD), (Haseman and Elston, 1972).

The original sib pair method was based on the idea that linkage is supported if sibling pairs with two affected or two unaffected siblings are significantly more alike in terms of allele sharing at a marker locus compared with sibling pairs with just one

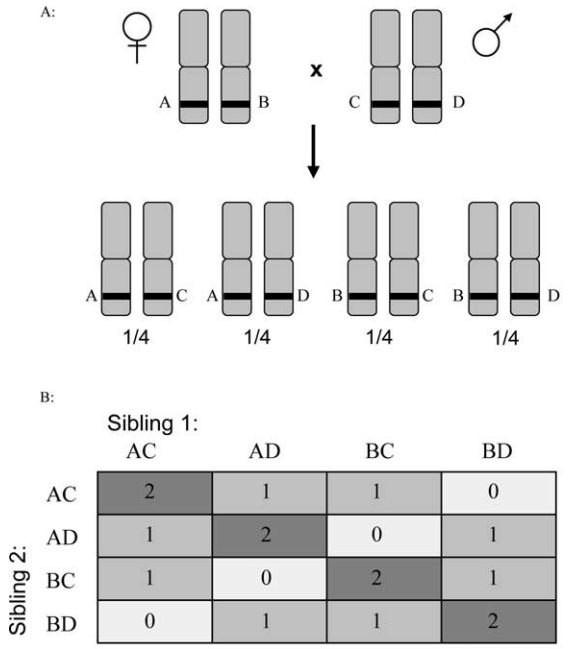

Fig. 3. IBD. (A) Graph showing the possible allele combination for children from a mother with allele A and B and a father with allele C and D. The chance for each combination (AC, AD, BC and BD) in the offspring is 1/4. (B) Identical by descent. The probability that two siblings share two parental alleles (IBD = 2) is 4/16 = 1/4. The probability that they do not share parental alleles is also 4/16 = 1/4, but the probability that they share one parental allele is 8/16 = 1/2.

affected member. This sib pair test was refined to give rise to the currently popular *affected sib pair* (ASP) method. Attention was focused exclusively on sibling pairs in which both members are affected, since such pairs are often more informative than unaffected sibling pairs, or sibling pairs with one affected and one unaffected member (Sham, 1998). The ability of the ASP method to detect a disease susceptibility locus depends on the contribution the locus makes to family resemblance, which is often measured in terms of the increased risk to relatives of an affected proband as compared with the population prevalence (Risch, 1990a,b). For ASP studies, this can be measured by the sibling risk ratio λ_s of the risk to a sib of an affected proband versus the population prevalence. This λ_s is an overall risk ratio that summarises the collective effect of all the disease loci plus any other non-genetic familial resemblance; higher λ_s indicates stronger familial effects (Lathrop and Weeks, 2000).

An example of non-parametric analyses of a dichotomous trait in ASPs is the study of Straub et al. (1999). This is the first published report of a complete genome scan designed to detect genes that influence the risk of nicotine dependence. A genome scan using 451 DNA markers was conducted to identify chromosomal regions linked to nicotine dependence in a sample from Christchurch, New Zealand

(201 ASPs from 130 families). Non-parametric linkage scores (z_{all}) were obtained under the assumption of locus heterogeneity. The z_{all} statistic is a 'similarity statistic' for affected relatives, and is defined as the average of the possibilities that relatives are IBS. The best result was with marker D2S1326 on chromosome 2. Straub et al. also found a number of large chromosomal regions where many consecutive markers yielded small but positive z_{all} scores. Selected regions of chromosomes were further investigated by additional genotyping of the Christchurch sample and an independent sample from Richmond, VA (190 ASPs from 91 families). For example, the analyses of the DNA markers on chromosome 2 in the Christchurch sample showed six positive z_{all} scores in a region over 19 cM. The best result marker D2S1326 ($z_{all} = 2.65$, $P = 0.0011$) was roughly in the middle of this region. In the Richmond sample there is a cluster of seven markers on chromosome 2 which all have positive z_{all} scores and the best result for the Richmond sample was marker D2S442 (z_{all} 1.05) which is located about 2 cM of the D2S1326 marker. Straub et al. (1999) found regions on chromosome 2, 4, 10, 16, 17 and 18 that merit further study. However, they also concluded that when simply judged against the usual standards of linkage significance, none of the individual regions yielded strong evidence. It is probable that the size of the available sample provided only limited power to detect linkage. This illustrates that it is difficult to detect genes of small effect, or genes that are influencing risk in only a small proportion of the families (Straub et al., 1999).

4. Statistical methods and power to detect linkage

Different methods can be used to calculate linkage. In general, the methods can be divided in regression analyses and maximum likelihood methods (variance component models). Haseman and Elston introduced in 1972 an elegant regression method to test for linkage for quantitative traits. Evidence that a marker is linked to the trait is obtained by regressing the squared trait difference between phenotypes of siblings on the proportion of marker alleles shared identical-by-descent (IBD) (π) (Haseman and Elston, 1972). A major drawback of this method is that it requires large numbers of sibs to detect significant evidence for linkage. The variance-component models were originally developed for the partitioning of phenotypic variation into genetic and environmental components from correlational data from pairs of relatives (usually twins), but have now been extended for QTL analysis (Fig. 4). With this approach not only the differences within sibships but also the differences between sibships can be modelled as a function of the QTL. Moreover, the full IBD distribution can be used in the estimation procedure and the model generalises quite easily to larger sibships and to mulivariate phenotypes (Martin et al., 1997; Boomsma and Dolan, 2000; Neale, 2000; Dolan et al., 1999; Fulker and Cherny, 1996). Under assumptions of (multivariate) normality, the parameters depicted in Fig. 4 can be estimated with maximum-likelihood methods, which are available in standard software packages such as LISREL or MX (Jöreskog and Sörbom, 1989; Neale et al., 1999).

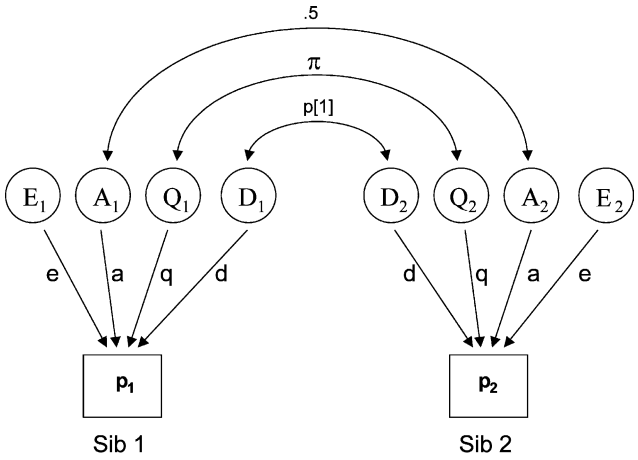

Fig. 4. Path-diagram with QTL effect. Path-diagram with observed phenotypes (p_1 and p_2) in sib 1 and sib 2 represented by squares, and latent variables E (individual-specific environment), A (additive genetic background), Q (additive QTL effect) and D (non-additive QTL effects) represented by circles. The path coefficients of each latent variable on the observed phenotypes are estimated (a, q, d, e). The correlation between additive QTL effects equals the proportion of alleles shared IBD (π) and the correlation between non-additive QTL effects is p[1]: the probability that siblings share all alleles identical-by-descent. The significance of the QTL effect is tested by constraining the path from QTL to phenotype at zero and test if this leads to decrease in the goodness of fit statistic.

A major problem of linkage analysis of complex traits with multiple contributing loci is the lack of statistical power. The main challenge thus is to develop linkage methods that have the highest statistical power to detect QTLs of small effect. The extensions of variance components methods listed above, such as multivariate approaches and testing larger sibships all lead to an increase in power. Visscher and Hopper (2001) compared the statistical power of linear regression and maximum likelihood methods to map QTLs for univariate traits from unselected sib pair data, and determined which methods are superior under which set of population parameters. Their derivations of statistical power for regression and maximum likelihood methods provide a simple way to compare alternative methods. If there are many covariates to be adjusted, a full maximum likelihood approach is recommended because regression methods have the drawback that they cannot perform multivariate analyses (Visscher and Hopper, 2001).

An important factor for power is the magnitude of the heritability or the familial risk ratio. Risch (1990a,b) describes the power to detect linkage as a function of the risk ratio λ_s (familial risk ratio) by using affected sibling pairs and assuming a fully informative marker and a recombination fraction of 0 between marker and QTL. For a sample of 200 affected sibling pairs the power to detect linkage is 0.4 when $\lambda_s = 2$, while the power to detect linkage is > 0.9 when $\lambda_s = 4$ (Risch, 1990a,b). In the same paper, Risch also shows that many of the power estimates are too optimistic if some of their common assumptions are violated. For most estimates it is assumed that the marker and the disease susceptibility locus are completely linked, and that

markers are completely informative. Risch demonstrated the potentially damaging effect on the power to detect linkage when the distance (recombination fraction) between marker and QTL is large. For a sample of 300 affected sibling pairs and a λ_s of 3, the power to detect linkage is 0.85 when the recombination fraction is 0, while the power is 0.4 and 0.15 when the recombination fraction is 0.05 and 0.10, respectively. Those problems, however, can be overcome to some extent by the use of multiple linked markers in multipoint analyses. The use of multipoint analyses requires prior knowledge of the relative positions of several marker loci in the chromosomal region of interest. The positions of the markers can then be fixed, while, in an iterative procedure, the putative position of the trait locus is varied from the one end of the region through the other end of the region. A LOD score is then calculated for each of the tested positions.

A genome-wide scan that provides a good illustration of non-parametric linkage analyses for a quantitative trait is a multipoint analysis for personality traits and a set of genetic markers. Cloninger et al. (1998) measured personality with the Tridimensional Personality Questionnaire and tested each of the four personality dimensions for linkage across all chromosomes. Genotyping was carried out on 987 individuals in 105 pedigrees, including 758 sibling pairs in 177 nuclear families. Multipoint variance components analysis was performed to estimate the genetic variance attributable to the QTL(s) linked to a genetic marker. Strong evidence was found that a genetic locus on 8p21–23 (marker D8S1106) accounted for most of the additive genetic variance in harm avoidance (anxiety-proneness vs. risk taking), (LOD score = 3.2, $P = 0.0006$). Cloninger et al. (1998) also tested for epistatic interactions (possibility of interactions between alleles at different loci) and found strong evidence for epistasis between the locus on 8p and others on chromosome 18p, 20p and 21q (LOD 5.1, $P = 0.000007$).

5. Allelic association

Linkage is usually genome-wide, while association studies are limited to candidate genes or candidate regions. Furthermore, linkage analyses must be carried out in pedigrees (families and sibling pairs), while association can be performed at the population level. With allelic association studies an association between a disease and a specific allele can be detected in groups of unrelated cases (e.g. patients) and controls (e.g. healthy subjects). Association can be found either with functional genetic variants that have biological consequences related to disease, or with other variants that are in linkage disequilibrium with these variants. Linkage disequilibrium occurs when a marker allele (i.e. a SNP) and the QTL are so close on the chromosome that they co-segregate in the population over many generations of meiotic recombination. Association studies are similar in design to classic case–control studies in epidemiology. DNA is collected from all participants and the trait is compared across the various allelic variants of the DNA marker. Vice versa, frequencies of the various allelic variants may be compared in subjects with particular phenotypes, to detect an association between a particular allele and the

occurrence of the phenotype. The advantage over linkage analysis is that association studies can detect the region of a QTL that has only very small effects on the trait (Risch and Merikangas, 1997). Provided that either the selection of cases does not introduce population stratification or that the analyses properly control for such stratification, association studies provide a good complement to the linkage strategy. Screening the entire genome with association, however, requires huge numbers of markers (linkage requires only a few hundred markers) and is not currently feasible. Allelic association, therefore, has been used primarily with candidate genes.

6. Candidate genes

The ideal candidate gene has been shown to be functional: it influences the concentration of the (iso)form of a protein, its functionality or efficiency, or perhaps most importantly, its responsiveness to environmental factors triggering the expression of the gene. The problem with a candidate gene approach for most complex traits is the potentially huge proportion of genes, which can serve as candidates. Several strategies are possible to select an optimal set of candidate genes. First, genes that are part of physiological systems known to influence the trait can be tested as candidates. Secondly, genes or chromosomal regions that are known to influence the trait in animals can be tested as candidate genes (or regions) in humans. Candidate genes for smoking and nicotine dependence, for example, could be genes that are involved in dopamine activity (because the dopamine reward pathway plays a critical role in substance use) but also genes that are involved in nicotine metabolism and genes involved in personality (sensation seeking, neuroticism, depression). Using such candidates, several associations between the dopamine receptor genes and substance use have been reported. A significant effect was found for the dopamine transporter gene; individuals with a particular variant of this gene (SLC6A3-9) were significantly less likely to be smokers, especially if they also had a certain variant of the D2 dopamine receptor (DRD2-A2), (Lerman et al., 1999). The long form of the D4 receptor gene is more frequent in individuals with high quantity/frequency of drug use compared with controls (Vandenbergh et al., 2000). The results of a population-based association study of substance abuse and a microsatellite at the dopamine D5 receptor locus (DRD5) in a sample of European–American males and females found that the DRD5 locus is involved in the variation of substance abuse liability (Vanyukov et al., 1998). Duaux et al. (2000) reviewed molecular genetic studies in drug abuse; results of several association studies reported positive association between drug disorder and polymorphisms of several dopaminergic receptor genes (DRD1, DRD2, DRD3, DRD4). A problem with the candidate gene approach is that by looking for candidates among the pathways that we already know, we may still overlook the essential genes, because of our ignorance of other biological systems involved.

7. Within-family association studies

Another problem of association studies is the danger that a spurious association is found between the trait of interest and any locus that differs in allele frequency between subpopulations. This situation is illustrated by the 'chopstick gene' story described by Hamer and Sirota (2000). They describe a hypothetical study in which DNA markers were assessed in students who often used chopsticks and students who did not. One of the DNA markers showed a huge correlation to chopstick use. Of course this gene had nothing to do with chopstick use, but just happened to have different allele frequencies in Asians and Caucasians, who differ in chopstick use for purely cultural rather than biological reasons. Witte et al. (1999) have evaluated the asymptotic bias in relative risk estimates resulting from using population controls when there is confounding due to population stratification. The direction of the bias is what one would expect from the usual principles of confounding in epidemiology: if the allele frequencies and baseline risks are both higher in a population, the bias is positive; if different, the bias is negative. Case–control studies of genetic associations thus can lead to false positive as well as to false negative results.

To prevent significant findings due to population stratification, within-family association designs have been developed, because family members are usually well matched on a number of traits that could give rise to stratification effects (Spielman et al., 1993). Most available family based tests for association were initially developed for binary traits, such as the Transmission Disequilibrium association Test (TDT) and the Haplotype Relative Risk Test (HRR). Those tests usually collect DNA samples in affected individuals and their biological parents. Affected individuals must have received one or two susceptibility alleles from their parents. These alleles transmitted from parents to the affected individual can be viewed as a group of 'case' alleles. The non-transmitted alleles from the parents can be considered as 'control' alleles. In other words, those tests only needs affected individuals and their parents, no control group is required (Terwilliger and Ott, 1994).

In a different approach the effects of genotypes on phenotypic means are partitioned into between-family and within-family components, by comparing the association of alleles and trait values across siblings from different families to the association of alleles and trait values across siblings within the same family. Sibling pairs are by definition ethnically and racially homogeneous and any difference in trait scores between siblings of different genotypes at a candidate marker, therefore, reflect true genetic association. By partitioning the mean effect of a locus into a between and a within-sibship component, spurious associations due to population stratification and admixture are controlled for (Abecasis et al., 2000; Fulker et al., 1999). An early example is the study of Lesch et al. (1996) that demonstrated that the observed associations between a polymorphism in the serotonin transporter gene (5-HTTLPR) and personality are the result of genetic transmission rather than population stratification (Lesch et al., 1996). The study population included 459 siblings from 210 independent families, of which 78 sibling pairs from 61 independent families had discordant 5-HTTLPR genotypes (one or two copies of

the short form vs. homozygous for the long form). The difference in personality scores between siblings with the long form and siblings with the short form of the 5-HTTLPR genotype was statistically significant. Most importantly, highly comparable results were obtained by population-based or across-pedigrees analyses.

The literature on family-based methods rapidly grows. Some methods extend the original tests to accommodate multi allelic markers, variable pedigree constellations, multiple loci, and quantitative traits. Family-based association studies are comprehensively described by Zhao (2000), Schulze and McMahon (2002).

8. Fine mapping

To detect which candidate gene in a linkage region is the causal gene, Fulker et al. (1999) introduced a systematic approach for the simultaneous analysis of both association and linkage for quantitative traits in sib pairs. If significant linkage is detected while also modelling association, the putative locus modelled in the association is not the functional gene. If linkage evidence vanishes when simultaneously modelling association, the marker may be the QTL itself (or in very strong linkage disequilibrium with it). A simultaneous test for linkage and association can be carried out using multipoint IBD information to model sib pair covariances (test of linkage) and a decomposition of the mean phenotype into allelic effects (test of association) between and within families. The novel joint analyses of both linkage and association is made possible by a statistical approach unified by the use of maximum likelihood and of a common biometrical model for the simultaneous analysis of means and covariance matrices. Cardon and Abecasis (2000) evaluated the behaviour of the association and linkage parameters in the model of Fulker et al. which may facilitate fine-mapping studies of complex traits that aim to localise QTLs by assessment of association with many markers in a candidate region of interest. An extension to the method of Fulker et al. is proposed by Posthuma et al. (2002), and allows the use of (variable) sibship sizes greater than two, the estimation of additive and dominance association effects, and the use of multiple alleles. These extensions can be implemented without parental genotypes but are most powerful when these genotypes are available.

9. Animal models in genetics

The oldest technique in behavioural genetics with animals is that of artificial selection. Mice (or other animals) are selected on their scores on tests for open field activity, behaviour in a maze or their behavioural response for e.g. sensitivity, tolerance, dependence and preference for alcohol or nicotine. Extreme scoring animals are mated and selection lines are created for high scoring animals and for low scoring animals. If such selection is possible, this proves that the trait is influenced by genetic factors (Crabbe et al., 1999). High and low scoring strains will differ at loci that influence the trait on which selection was based. In contrast inbred

strains are created by repeated matings between brothers and sisters. Within an inbred strain, all same-sex animals are essentially monozygotic twins and have two identical copies of a single allele at each locus. Crabbe (2002) describes a study in which different inbred strains of mice were offered a choice between a bottle filled with tap water and one containing alcohol. The differences among strains in preference for alcohol far exceeded the within-strain differences; suggesting that these preferences have a genetic basis. Many different inbred strains of mice are available for genetic mapping experiments. Initial identification of QTLs involves examining many individual animals and correlating the possession of specific alleles at genetic markers with the degree of quantitative trait expressed (Crabbe et al., 1999). QTLs in mice have been found for several drug-sensitivity genes. For example, QTL analyses revealed that several genetic markers in inbred mice were associated with ethanol consumption levels, including markers for the D2 dopamine receptor (Philips et al., 1994; Buck et al., 2000). Another opportunity to evaluate the roles of gene products in animals is the genetic engineering approach. In this approach, mice that are made to lack ('knock-out'), under-express or over-express specific genes are studied. Several studies in knock-out mice have demonstrated the effects of specific genes on behavioural responses to drugs. For example, knocking out a serotonin receptor gene in mice leads to increased alcohol consumption and to increased vulnerability to cocaine (Rocha et al., 1998). Rubinstein et al. (1997) found supersensitivity to alcohol, cocaine, and methamphetamine in mice whose dopamine D4 receptor was knocked out.

A new development is the use of expression arrays or so called 'gene chips'. Thousands of individual gene sequences can be bound to tiny chips (glass plates). When a sample of DNA or RNA is applied, those genes actively express in the sample, bind to their embedded ligand and the resulting interaction is visualised. At least 6000 mouse brain DNA probes are available on chips, and can be used to study gene expression under different conditions of e.g. environmental exposure, thus identifying genes (Crabbe, 2002). For example, Freeman et al. (2001) used expression arrays to identify cocaine-regulated genes by comparing the gene-expression in rats treated with cocaine versus control rats. The findings suggest altered expression of genes with a number of different functions in the rat hippocampus after cocaine administration (a.o. induction of potassium channel 1.1, protein tyrosine kinase 2), (Freeman et al., 2001). As the homologous region of murine genes in the human genome is often known, the genes/regions identified in mouse studies (or other animal studies) can be regarded as plausible candidate genes/regions in human genetic studies (Picciotto et al., 2000).

10. Discussion

The major strength of linkage is that it is systematic in the sense that a few hundred DNA markers can be used to scan the entire genome. In contrast, allelic association with a quantitative trait can only be detected if a DNA marker is the QTL itself or very close to it, so tens of thousands of DNA markers would need to be

genotyped to scan the entire genome On the other hand, association studies can detect QTLs with only small effects on the trait, whereas linkage may not. Linkage and association analysis are, therefore, fully complementary approaches; association studies can be used as an approach to isolate a susceptibility gene in a region that has first been identified by linkage. Evidence for a substantial genetic contribution, in terms of the sibling recurrence ration (λ_s) for all-or-none traits and heritability (h^2) or sibling correlation (r) for the quantitative traits, is a prerequisite for embarking on gene mapping studies. Nonetheless, even for confirmed heritable complex diseases, linkage and association approaches have met with limited success so far.

Altmuller et al. (2001) reviewed 101 whole genome scans of complex human disease, which were found by a systematic Medline search. These linkage studies were compared with regard to design, method and relative 'success'. Most studies (66.3%) did not show significant linkage (using the criteria of Lander and Kruglyak, 1995) and the results of studies of the same disease were often inconsistent. Altmuller et al. concluded that no single study design consistently produces more-significant results. The only factors independently associated with increased study success were (1) an increase in the number of individuals studied and (2) studies of subjects drawn from only one ethnic group (Altmuller et al., 2001; Guo, 2002). An efficient method to realize the power of gene detection in large samples is to phenotype a large samples and to select a subgroup of the most informative families for genotyping. Selection and genotyping of extremely discordant and concordant sibling pairs can increase the power to detect linkage without the need to genotype the entire sample from which the extreme pairs were drawn (Risch and Zhang, 1995; Dolan et al., 1999; Abecasis et al., 2001). Several studies of anxiety and depression have described such methods for selection of extreme discordant and concordant sib pairs (Boomsma et al., 2000; Kirk et al., 2000; Martin et al., 2000). The limited success of linkage studies has led to the proposal that genetic association studies may offer a better alternative to find genes for complex traits in humans. A review of meta-analysis studies on genetic associations in human disease by Ionnidis et al. (2001), however, shows that this promise has not yet been realised. Of the 36 traits considered (based on a total of 370 studies) only eight traits/diseases showed statistically significant associations in the meta-analysis (a.o. ischaemic stroke/ECE, bladder cancer/NAT2). In eight other traits, the first study revealed statistically significant associations and subsequent research reported either trends that suggested association or statistically significant associations. At the end of the meta-analysis, however, in only four of the eight traits did the evidence for genetic association achieve formal statistical significance. When an initial study suggests a stronger genetic effect than is found in subsequent studies, this can be caused by sampling bias (the most prominent findings represent an extreme sample and associations may be less extreme in new studies), by publication bias, by inflation of the size of a genetic effect (if based only on a single study with impressive results) and by a large statistical uncertainty in the first study (Ionnidis et al., 2001; Vieland, 2001). We concur with Iaonnidis and colleagues that association studies require cautious replication—and we believe that this applies to all gene findings using either linkage or association methods.

We have focused in this review mostly on the methodology of linkage studies employing sib-pair designs, which can also be used as a tool in association studies by decomposing the association effect into a between and within-families component. Other designs, such as linkage and association studies in large pedigrees or isolated populations may be less feasible for the complex traits studied in the fields of human behaviour and psychophysiology. More distant relatives or isolated populations are useful to detect rare susceptibility genes, whereas closer relatives, such as siblings are required for studies of common diseases and traits.

Acknowledgements

This work was supported by the Netherlands Organisation for Scientific Research (NWO 985-10-002) and ZonMW/NIDA (3100.0038).

References

Abecasis, G.R., Cardon, L., Cookson, W.O.C., 2000. A general test of association for quantitative traits in nuclear families. Am. J. Hum. Genet. 662, 279–292.

Abecasis, G.R., Cookson, O.C., Cardon, L., 2001. The power to detect linkage disequilibrium with quantitative traits in selected samples. Am. J. Hum. Genet. 68, 1463–1474.

Altmuller, J., Palmer, L.J., Fischer, G., Scherb, H., Wjst, M., 2001. Genomewide scans of complex human diseases: true linkage is hard to find. Am. J. Hum. Genet. 69, 936–950.

Anokhin, A., Steinlein, O., Fischer, C., Mao, Y., Vogt, P., Schalt, E., Vogel, F., 1992. A genetic study of the human low-voltage electroencephalogram. Hum. Genet. 90, 99–112.

Beijsterveld, V.C., Boomsma, D.I., 1994. Genetics of the human electroencephalongram (EEG) and event-related brain potentials (ERPs): a review. Hum. Genet. 94, 319–330.

Boomsma, D.I., Dolan, C.I., 2000. Multivariate QTL analysis using structural equation modelling: a look at power under simple conditions. In: Spector, H.S.T.D., MacGregor, A.J. (Eds.), Advances in Twin and Sib-Pair Analysis. Greenwich Medical Media, London, pp. 203–218.

Boomsma, D.I., Beem, A.L., Berg van den, M., Dolan, C.V., Koopmans, J.R., Vink, J.M., Geus de, E.J.C., Slagboom, P.J., 2000. Netherlands Twin Family Study of anxious depression (NETSAD). Twin Res. 3, 323–334.

Bouchard, T.J.J., Loehlin, J.C., 2001. Genes, evolution, and personality. Behav. Genet. 31 (3), 243–273.

Buck, K., Lischka, T., Dorow, J., Crabbe, J., 2000. Mapping quantitative trait that regulate sensitivity and tolerance to quinpirole, a dopamine mimetic selective for D2/D3 receptors. Am. J. Med. Genet. 96, 696–705.

Cardon, L.R., Abecasis, G.R., 2000. Some properties of a variance components model for fine-mapping quantitative trait loci. Behav. Genet. 30 (3), 235–243.

Cloninger, C.R., Eerdewegh, V.P., Goate, A., Edenberg, H.J., Blangero, J., Hesselbrock, V., Reich, T., Nurnberger, J., Schuckit, M., Porjesz, B., Crowe, R., Rice, J.P., Foroud, T., Przybeck, T.R., Almasy, L., Bucholz, K., Wu, W., Shears, S., Carr, K., Crose, C., Willig, C., Zhao, J., Tischfield, J.A., Li, T.K., Coneally, P.M., Begleiter, H., 1998. Anxiety proneness linked to epistatic loci in genome scan of human personality traits. Am. J. Med. Genet. 81, 313–317.

Crabbe, J.C., 2002. Genetic contributions to addiction. Annu. Rev. Psychol. 53, 435–462.

Crabbe, J.C., Philips, T.J., Buck, K.J., Cunningham, C.L., Belknap, J.K., 1999. Identifying genes for alcohol and drug sensitivity: recent progress and future directions. Tins 22 (4), 173–178.

Dolan, C.V., Boomsma, D.I., Neale, M.C., 1999. A note on the power provided by sibships of sizes 2, 3, and 4 in genetic covariance modeling of a codominant QTL. Behav. Genet. 29 (3), 163–170.

Duaux, E., Krebs, M.O., Loo, H., Poirier, M.F., 2000. Genetic vulnerability to drug abuse. Eur. Psychiatry 15, 109–114.
Elston, R.C., 2000. Introduction and overview. Stat. Methods Med. Res. 9, 527–541.
Falconer, D.S., Mackay, T.F.C., 1996. Quantitative Genetics. Longman Group, Essex.
Faraone, S.V., Doyle, A.E., 2001. The nature and heritability of attention-deficit/hyperactivity disorder. Child Adolesc. Psychiatr. Clin. North Am. 10 (2), 299–316.
Fisher, R.A., 1918. The correlation between relatives on the supposition of Mendalian inheritance. Trans. R. Soc. Edinburgh 52, 399–433.
Freeman, W.M., Brebner, K., Lynch, W.J., Robertson, D.J., Vrana, K.E., 2001. Cocaine-responsive gene expression changes in rat hippocampus. Neuroscience 108 (3), 371–380.
Fulker, D.W., Cherny, S.S., 1996. An improved multipoint sib-pair analysis of quantitative traits. Behav. Genet. 26 (5), 527–531.
Fulker, D.W., Cherny, S.S., Sham, P., Hewitt, J.K., 1999. Combined linkage and association sib-pair analysis for quantitative traits. Am. J. Hum. Genet. 64, 259–267.
Gayan, J., Olson, R.K., 1999. Reading disability: evidence for a genetic etiology. Eur. Child Adolesc. Psychiatry 8 (suppl 3), 52–55.
Gershon, E.S., 2000. Bipolar illness and shizophrenia as oligogenic diseases: implications for the future. Biol. Psychiatry 47, 240–244.
Guo, S.W., 2002. Sibling recurrence risk ratio as a measure of genetic effect: caveat emptor. Am. J. Hum. Genet. 70, 818.
Hamer, D., Sirota, L., 2000. Beware the chopstick gene. Mol. Psychiatry 5, 11–13.
Haseman, J.K., Elston, R.C., 1972. The investigation of linkage between a quantitative trait and a marker locus. Behav. Genet. 2 (1), 3–19.
International Snp Map Working Group, 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409, 928–933.
International Human Genome Sequencing Consortium, 2001. Initial sequencing and analysis of the human genome. Nature 409, 860–921.
Ionnidis, J.P.A., Ntzani, E.E., Trikalinos, T.A., Contopoulos-Ioannidis, D.G., 2001. Replication validity of genetic association studies. Nat. Genet. 29, 306–309.
Jöreskog, K.G., Sörbom, D., 1989. LISREL 7: A Guide to the Program and Applications, second ed.. Scientific Software International, Chicago.
Kirk, K., Birley, A., Statham, D.J., Haddon, B., Lake, R.I.E., Andrews, J.G., Martin, N.G., 2000. Anxiety and depression in twin and sib pairs extremely dicordant and concordant for neuroticism: prodromus to a linkage study. Twin Res. 3, 299–309.
Lander, E., Kruglyak, L., 1995. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat. Genet. 11 (November).
Lathrop, G.M., Weeks, D.E., 2000. Methods for mapping complex disease traits. In: Bishop, T., Sham, P. (Eds.), Analysis of Multifactorial Disease. BIOS Scientific Publishers, Oxford, UK, pp. 15–31.
Lerman, C., Caporaso, N.E., Audrian, J., Main, D., Bowman, E.D., Lockshin, B., Boyd, N.R., Shields, P.G., 1999. Evidence suggesting the role of specific genetic factors in cigarette smoking. Health Psychol. 18 (1), 14–20.
Lesch, K.P., Bengel, D., Heils, A., Sabol, S.Z., Greenberg, B.D., Petri, S., Benjamin, J., Muller, C.R., Hamer, D.H., Murphy, D.L., 1996. Association of anxiety-related traits with a polymorphism in the serotonin transporter gene regulatory region. Science 274 (5292), 1527–1531.
Martin, N., Boomsma, D.I., Machin, G., 1997. A twin prolonged attach on complex traits.
Martin, N., Goodwin, G., Fairburn, C., Wilson, R., Allison, D., Cardon, L., Flint, J., 2000. A population-based study of personality in 34000 sib-pairs. Twin Res. 3, 310–315.
McKusick, V.A., 1998. Medalian Inheritance in Man. In: A Catalog of Human Genes and Genetic Disorders. John Hopkins University Press, Baltimore, MD.
Molenaar, P.C.M., Boomsma, D.I., Dolan, C.V., 1993. A third source of developmental differences. Behav. Genet. 23 (6), 519–531.
Morton, N.E., 1955. Sequential test for the detection of linkage. Am. J. Hum. Genet. 7, 277–318.

Neale, M.C., 2000. QTL mapping with sib-pairs: the flexibility of MX. In: Spector, H.S.T.D., MacGregor, A.J. (Eds.), Advances in Twin and Sib-Pair Analysis. Greenwich Medical Media, London, pp. 203–218.

Neale, M.C., Boker, S.M., Xie, G., Maes, H.H., 1999. MX: Statistical Modeling, fifth ed.. Department of Psychiatry, Richmond, VA.

Peltonen, L., McKusick, V.A., 2001. Dissecting human disease in the postgenomic era. Science 291, 1224–1229.

Philips, T.J., Crabbe, J.C., Metten, P., Belknap, J.K., 1994. Localization of genes affecting alcohol drinking in mice. Alcohol Clin. Exp. Res. 18 (4), 931–941.

Picciotto, M.R., Caldarone, B.J., Kink, S.L., Zachariou, V., 2000. Nicotine receptors in the brain: links between molecular biology and behavior. Neuropsychopharmacology 22 (5), 451–465.

Plomin, R., Crabbe, J.C., 2000. DNA. Psychol. Bull. 126 (6), 806–828.

Plomin, R., DeFries, J.C., McClearn, G.E., McGuffin, P., 2000. Behavioral Genetics. Worth Publishers, New York.

Posthuma, D., Geus de, E.J.C., Boomsma, D.I., Neale, M.C., 2002. Combined linkage and association test in Mx. In: Posthuma, D. Genetic variation and cognitive ability. Amsterdam: PrintPartners Ipskamp, pp. 153–174.

Risch, N., 1990. Linkage strategies for genetically complex traits. II. The power of affected relative pairs. Am. J. Hum. Genet. 46, 229–241.

Risch, N., 1990. Linkage strategies for genetically complex traits. III. The effect of marker polymorphism on analysis of affected relative pairs. Am. J. Hum. Genet. 46, 242–253.

Risch, N., Zhang, H., 1995. Extreme discordant sib pairs for mapping quantitative trait loci in humans. Science 16 (268(5217)), 1584–1589.

Risch, N., Merikangas, K., 1997. The future of genetic studies of complex human diseases. Science 275 (5281), 1516–1517.

Rocha, B.A., Scearce-Levie, K., Lucas, J.J., Hiroi, N., Castanon, N., Crabbe, J.C., Nestler, E.J., Hen, R., 1998. Increased vulnerability to cocaine in mice lacking the serotonin-1B receptor. Nature 14 (393(6681)), 175–178.

Rubinstein, M., Philips, T.J., Bunzow, J.R., Falzone, T.L., Dziewczapolski, O., Zhang, G., Fang, Y., Larson, J.L., McDougall, J.A., Chester, J.A., Saez, C., Pugsley, T.A., Gershanik, O., Low, M.J., Grandy, D.K., 1997. Mice lacking dopamine D4 receptors are supersensitive to ethanol, cocaine and methamphetamine. Cell 19 (90(6)), 991–1001.

Schulze, T.G., McMahon, F.J., 2002. Genetic association mapping at the crossroads: which test and why? Overview and practical guidelines. Am. J. Hum. Genet. 114 (1), 1–14.

Sham, P., 1998. Statistics in Human Genetics. Oxford University Press, New York.

Spielman, R.S., McGinnis, R.E., Ewens, W.J., 1993. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am. J. Hum. Genet. 52 (3), 506–516.

Steinlein, O., Anokhin, A., Yping, M., Schalt, E., Vogel, F., 1992. Localization of a gene for the human low-voltage EEG on 20q and genetic heterogeneity. Genomics 12, 69–73.

Straub, R.E., Sullivan, P.F., Ma, Y., Myakishev, M.V., Harris-Kerr, C., Wormley, B., Kadami, B., Sadek, H., Silverman, M.A., Webb, B.T., Neale, M.C., Bulik, C.M., Joyce, P.R., Kendler, K.S., 1999. Susceptibility genes for nicotine dependence: a genome scan and followup in an independent sample suggest that regions on chromosones 2, 4, 10, 16, 17 and 18 merit further study. Mol. Psychiatry 4, 129–144.

Sullivan, P.F., Kendler, K.S., 1999. The genetic epidemiology of smoking. Nicotine Tobacco Res. 1, S51–S57.

Terwilliger, J.D., Ott, J., 1994. Handbook of human genetic linkage. Baltimore, MD, The John Hopkins University Press.

Vandenbergh, D.J., Rodriguez, L.A., Hivert, E., Schiller, J.H., Villareal, G., Pugh, E.W., Lachman, H., Uhl, G.R., 2000. Long forms of the dopamine receptor (DRD4) Gene VNTR are more prevalent in substance abusers: no interaction with functional alleles of the Catechol-*O*-Methyltransferase (COMT) gene. Am. J. Med. Genet. 96, 678–683.

Vanyukov, M.M., Moss, H.B., Gioio, A.E., Hughes, H.B., Kaplan, B.B., Tarter, R.E., 1998. An association between a microsatellite polymorphism at the DRD5 gene and the liability to substance abuse: pilot study. Behav. Genet. 28 (2), 78–82.

Venter, C., Adams, M.D., Myers, E.W., Al, E., 2001. The sequence of the human genome. Science 291, 1304–1351.

Vieland, V.J., 2001. The replication requirement. Nat. Genet. 29, 244–245.

Visscher, P.M., Hopper, J.L., 2001. Power of regression and maximum likelihood methods to map QTL from sib-pair and DZ twin data. Ann. Hum. Genet. 65, 583–601.

Vogel, F., Motulsky, A.G., 1997. Human Genetics, Problems and Approaches. Springer, New York.

Witte, J.S., Gauderman, W.J., Thomas, D.C., 1999. Asymptotic bias and efficiency in case-control studies of candidate genes and gene-environment interactions: basis family designs. Am. J. Epidemiol. 149 (8), 693–705.

Zhao, H., 2000. Family-based association studies. Stat. Methods Med. Res. 9, 563–587.