European Journal of Human Genetics (2005) 13, 698–706

www.nature.com/ejhg

REVIEW

Complex trait mapping in isolated populations: Are specific statistical methods required?

Catherine Bourgain*,[1] and Emmanuelle Génin[1]

[1]*INSERM U535, Hôpital Paul Brousse, Villejuif, France*

In this paper, we review the statistical methods that can be used in isolated populations to map genes involved in complex diseases. Our intention is to highlight the fact that if the features of population isolates may help in the identification of susceptibility factors for complex traits, the choice and design of methods for statistical analysis in these populations deserve particular care. We show that methods designed for outbred samples are generally not appropriate for isolated populations and could lead to false conclusions.
European Journal of Human Genetics (2005) **13,** 698–706. doi:10.1038/sj.ejhg.5201400
Published online 23 March 2005

Keywords: isolates; inbreeding; kinship; statistical methods; gene mapping; complex trait

Introduction

Genetic studies in founder populations have led to many successes in the identification of the mutations responsible for various rare monogenic diseases. Efficient linkage disequilibrium mapping strategies have been designed to take full advantage of the existence of unique founder mutations introduced only once in the populations and thus shared by virtually all the affected individuals. Indeed, the existence of founder mutations for rare mendelian diseases has been described in a wide range of populations, spanning from small strongly inbred isolates to much larger populations founded by a few thousand individuals, a few centuries ago. Heutink and Oostra[1] have proposed to classify founder populations according to the number of generations since their foundation – very old isolates (>100 generations), young isolates (<100 generations) and very young isolates (<20 generations) – arguing that the older the populations, the better they are for localizing the mutation. The number of founders and the growth pattern of the population also have crucial consequences on genetic characteristics.

The genetic factors involved in complex diseases are alleles that are neither necessary nor sufficient for disease expression but that increase the risk, in often rather modest proportions and through complex interactions with many other genetic and environmental risk factors. The question as to whether these alleles are likely to be frequent or not is still open. Recent examples of both very frequent[2] and rare susceptibility alleles[3] have been described for various diseases. However, the frequencies are always higher than those for rare monogenic mutations. Further, in the case of relatively rare alleles, they are not present in all the affected individuals but only in a small subset. This change in characteristics of the genetic factors under study has called for the development of new mapping methods based either on linkage or/and on association information.

As in the case of outbred populations, the genetic study of complex traits in isolated populations requires a change in methodology compared with monogenic disease studies. Indeed, the probability of observing a founder allele, introduced only once in the population, when considering a common susceptibility allele, is likely to be negligible. Methodologies relying on that assumption are thus inappropriate. However, isolated populations may still

*Correspondence: Dr C Bourgain, Genetique Epidemiologique et Structure des Populations Humaines, INSERM U535, Hôpital Paul Brousse, Batiment Leriche, BP 1000, 94817 Villejuif Cedex, France.
Tel: +33 1 45 59 53 85; Fax: +33 1 45 59 53 31;
E-mail: bourgain@vjf.inserm.fr
Received 19 October 2004; revised 25 January 2005; accepted 3 February 2005

present interesting features with regard to complex trait mapping. Environmental risk factors are generally more uniform in isolated populations than in large outbred populations and thus the genetic effects may be easier to identify in the former. Greater genetic homogeneity is also expected in these populations because of the limited number of founders (and thus a limited gene pool) and because of the absence of migration, which virtually rules out the risk of unidentified population stratification. Furthermore, the availability of extensive genealogical records can provide large genealogies, potentially very informative for linkage analysis. Finally, linkage disequilibrium may extend over larger regions than in outbred populations, increasing the power of association study for gene detection. We note that whereas these populations are alternatively described as 'founder' or 'isolated', the choice of 'isolated populations' better reflects the properties likely to be useful for complex trait mapping.

We do not intend to discuss whether isolated populations are the 'El Dorado' of genetic studies in this paper, many authors have discussed this issue before.[1,4–7] As the number of genetic studies in isolated population is rapidly increasing and because the first successes are beginning to appear, we instead propose to review and discuss the statistical methods available for complex trait mapping in isolated populations, their advantages and their limitations.

Apart from their isolation, a key characteristic of these populations in terms of methodology is that, contrary to outbred populations, the probability for two random individuals to be related is not negligible. Further, inbreeding might also be present. In this case, not only are two individuals potentially correlated but the two alleles in a random individual may also be correlated. The existence of inter- and intra-individual correlations has important consequences on most mapping methods. In particular, the distinction between linkage information and association information might become tricky. We first review the methods for linkage studies of qualitative and quantitative traits and then focus on association studies. We show how classical methods might not be valid in isolated populations and present both the different strategies available to correct existing methods as well as the new methods specifically developed to make use of some particular properties of isolated populations. We separate the different methods depending on the extent of genealogical information available. We would like to note that the problem of correlations existing among random individuals is not specific to isolated populations but also concerns studies of extended genealogies in outbred populations. The problems and methods discussed in the present paper may thus be of a more general interest to human geneticists.

Linkage analysis

Qualitative traits

Linkage analysis allows a comprehensive scan of the entire genome for disease genes in a hypothesis-independent manner.[8] It looks for shared segments of DNA among related patients that exceed the amount expected on the basis of their relationship pattern. In large outbred populations, this is usually performed on samples of independent affected sib-pairs. In isolated populations, affected sib-pairs are often not independent but related to an extent that depends on the demography of the population. Moreover, affected siblings and their parents may be inbred, resulting in an excess of shared segments of DNA on the entire genome as compared to what is expected in the absence of inbreeding. Not accounting for this increase in the expected sharing probability when performing a linkage test on inbred sib-pairs enhances the risk of falsely concluding linkage with a given region of the genome as shown by Génin *et al*[9] in the situation where parents are not typed, and by Leutenegger *et al*[10] in more general situations. Specific allele-sharing statistics have been proposed by McPeek[11] for inbred pairs.

When genealogical data are available, one can use them to trace back the relationships between the different affected individuals and perform linkage analysis on extended pedigrees. The approach has in particular been very fruitful in Iceland where the strategy of using general families that extended beyond the nuclear family has led to the detection of significant linkage for a number of complex diseases.[12–24] If additional replications of these findings in other populations are still required for validation, the success of this genealogical approach may be a consequence of the population history of Iceland, as explained by Gulcher *et al*.[8] However, both the important sample sizes available in the different studies and the high density of markers used in the genome scans are certainly other key parameters of the success. We note that the genealogies used in these Icelandic studies are not extremely large. Affected individuals are related at five meioses at most (eg they are first or second cousins). One may then wonder whether going back further in the past is of additional usefulness in linkage analysis. Indeed the chance that affected individuals within the same pedigree share genetic risk factors identical by descent decreases rapidly when the relationship between these individuals becomes more remote and this is especially true for common risk factors. Even in the case of a simple dominant disease segregating in a pedigree, the evidence of linkage provided by a pair of distant affected relatives first increases with decreasing values of the kinship coefficient, reaches a maximum that depends on marker allele frequencies and then decreases.[25] Further, pedigree complexity and inbreeding loops greatly increase the computational burden of linkage analysis and the different computer programs available are limited in terms of

number of individuals to include in a single pedigree. Programs using exact multipoint estimation of identity by descent (IBD), where information provided by neighbouring markers is used to better estimate the IBD sharing at a given marker, such as Allegro,[26] require $2N-F$ to be less than 26, where F and N represent the number of founders and nonfounders in the pedigree, respectively. Programs using Monte Carlo Markov Chain (MCMC) approaches to IBD estimation, such as Simwalk2,[27] can analyse larger pedigrees but are also limited. Pedigrees and inbreeding loops have thus often to be broken into smaller units. Such breaking should be performed very carefully to minimize false specification of expected sharing probabilities and possibilities of spurious linkage detection.

When genealogical data are neither available nor reliable, one may opt for a strategy that consists in contrasting the observed sharing between pairs of affected relatives to the sharing observed over the whole genome. Indeed, genomic data can be used to estimate the separation distance between two affected individuals[28–32] or the inbreeding coefficient of an individual.[33]

Quantitative traits

In the context of quantitative trait linkage analysis, variance component (VC) approaches are used on relatively large pedigrees. Briefly, the principle of VC is to consider that a phenotypic trait, Y, can be decomposed into the addition of a fixed effect, one (or several) quantitative trait locus (QTL) random effect(s) and an environmental random effect. Both the QTL and environmental effects are expected to have normal distributions with mean 0 and standard deviation, respectively, σ_g and σ_e. Assuming that the trait Y is approximately normally distributed in the population, it is possible to write the joint likelihood of a sample and test the existence of a QTL around the studied marker(s) by rejecting the null hypothesis $\sigma_g^2 = 0$ (no linkage), using for example, a likelihood ratio test (LRT). To write the likelihood, the covariance between any pair of individuals in the sample must be available, which, in turn, requires the characterization of the IBD sharing for all the pairs in the sample. If exact multipoint IBD sharing estimation is limited to moderate-sized pedigrees as already noted for qualitative traits, approximate methods, such as correlation-based[34] or MCMC[35] algorithms, allow to use VC for much larger pedigrees. An illustration of the use of VC in isolated populations is the work of Williams-Blangero *et al*,[36] who have conducted a linkage analysis on susceptibility to Ascaris infection (a roundworm) based on a 444-member pedigree from the Jirels, an isolated Nepalese population. Individuals were selected based only on pedigree informativeness and not with respect to Ascaris phenotype. Owing to a non-normal distribution of the trait, even after transformation, a robust test[37] was chosen and *P*-values were validated using simulations. A total of 6209 pairs of relatives were informative for the analysis and allowed the detection of two QTLs, one on chromosome 1 and the other on chromosome 13. In inbred pedigrees, additional components of the variance are required to fully describe even simple models. Estimation of these components relies on the calculation of additional identity coefficients between the pairs of relatives, a task that can seriously increase the computational burden. However, as shown by Abney *et al*,[38] given the low power to estimate these additional VCs, even in inbred samples, neglecting them in the analyses should not impact the power to detect linkage. Finally, VC approaches are not robust to departures from normality.[39] When such a departure is due to a selective sampling (the trait is normal in the population but not in the selected sample), the regression-based method proposed by Sham *et al*[40] and implemented in the software MERLIN-REGRESS is an interesting alternative available for pedigree data.

In extremely large and complex pedigrees, these methods may be computationally infeasible, especially in the context of genome screens. While analysing the Hutterite pedigree (see Table 2 for a brief description of the Hutterites) with a MCMC method for IBD-sharing estimation, Chapman *et al*[41] had to break it into subpedigrees. As discussed for the qualitative traits, reducing pedigree complexity may entail a loss of power. Dyer *et al*[42] showed how breaking a 1544 individual Hutterite pedigree into three subpedigrees, divided by 2 the relative efficiency to detect a QTL responsible for 20% of the phenotypic variance of a quantitative trait with a total heritability of 50%. However, these authors did not study the sensitivity of this result to the genetic model considered for the QTL. Recently, Falchi *et al*[43] proposed a new systematic method for pedigree breaking that maximizes useful information for linkage while minimizing the burden in IBD calculation. They show that the loss of power when using subpedigrees depends on the genetic model.

Abney *et al*[44] have proposed a regression-based linkage method for quantitative traits in complex and inbred pedigrees. The method relies on the existence of regions that are homozygous by descent (HBD, the two homologous regions are both copies of the same ancestral region) in inbred individuals, to detect QTLs that act recessively. HBD sharing is estimated using multipoint marker information and the complete pedigree information. Excess HBD sharing is expected for markers linked to QTLs acting recessively. The trait is regressed on different covariates including the HBD status. The test is equivalent to a *t*-test for the coefficient of the HBD covariate, corrected exactly for the known correlations among the individuals computed using the pedigree information. Abney *et al*[44] also proposed a permutation test to correct for multiple testing over the genome that preserves the correlation structure of the data. Indeed, for permutation tests to be valid, the elements to be permuted must be exchangeable, which is

not the case when individuals are related. In the VC context, Iturria *et al*[45] have attempted to solve a similar problem by approximately maintaining the familial correlation structure. The procedure of Abney *et al*[44] maintains this correlation structure in an exact manner provided that the genealogy is correctly specified and the trait under study has a multivariate normal distribution. This permutation procedure is extendable to a wide set of methods relying on similar linear models for the data, including VC approaches.

Association studies

Whereas linkage studies yield relatively broad locations for susceptibility loci, association studies may be used to test the role of particular candidate genes. As they are sensitive to ignored population substructures, case–control studies in outbred populations are used very cautiously and family-based association tests are often preferred to avoid the detection of spurious association. Yet, in the absence of population stratification, case–control tests are more powerful. They may thus regain interest in isolated populations.

Possible bias

Two tests are classically used to test for association, the case–control allele-based (CC-χ^2_{allelle}) and the case–control genotype-based (CC-χ^2_{genotype}) χ^2 tests, which contrast allele or genotype frequencies between a sample of cases and a sample of controls. Under the null hypothesis of no association, CC-χ^2_{allelle} and CC-χ^2_{genotype} follow a χ^2 distribution. However, for this null distribution to be valid, the cases and the controls must be independent, which might not be true in isolated populations. Bourgain *et al*[46] have illustrated the increase of type I error (probability of detecting an association when there is no association) of the CC-χ^2_{allelle} test in samples drawn from the Hutterite population. Table 1 shows similar results for the CC-χ^2_{allelle} and CC-χ^2_{genotype} in two different samples: the highly inbred Hutterite sample and a non inbred sample of related cases and controls from the GAW12-simulated genealogies (see Box 1 for a brief description of the two data sets).

Table 1 Empirical type I error of the CC-χ^2_{genotype}, CC-χ^2_{allele}, CC-QLS and CC-corrχ^2 tests using either the χ^2 distribution or a resampling procedure to get significance. Nominal type I error is 5%

		Hutterite sample SNP frequency		*GAW12 isolate sample SNP frequency*	
P-value computed with …	*Test statistic*	*0.5*	*0.2*	*0.5*	*0.2*
χ^2 distribution	CC-χ^2_{genotype}	0.12	0.12	0.13	0.13
	CC-χ^2_{allele}	0.15	0.14	0.14	0.14
Resampling procedure	CC-χ^2_{genotype}	0.14	0.13	0.13	0.15
	CC-χ^2_{allele}	0.14	0.13	0.11	0.15
χ^2 distribution	CC-QLS	0.051	0.050	0.046	0.053
	CC-corrχ^2	0.050	0.049	0.052	0.051

5000 simulations of the Hutterite and GAW12 samples, performed for two allele frequency sets of SNP.

Box 1

GAW12 data

The data simulated for the 12th Genetic Analysis Workshop[70] consist in samples of 1000 individuals with phenotype and genotype data. These individuals are actually the living members of 23 noninbred and independent extended genealogies totaling 1497 individuals. The mean kinship coefficient between the 1000 living individuals is relatively low (0.0018) because the 23 genealogies are independent, but the standard error of the kinship is high (0.0166). In all, 50 simulated replicates based on the same genealogies were available. To study the properties of the case–control tests, we chose one of these replicates in which 281 individuals out of the 1000 were cases and the remaining 719 were controls (replicate number 5). We performed our own genotype data simulations for the cases and controls. Alleles in the founders of the genealogies were randomly and independently drawn from a given allele frequency distribution. Mendelian transmission of these alleles was then simulated throughout each genealogy.

Hutterite data

The Hutterites are a North American religious isolate originating from Tyrol whose entire population can be traced back in the 1700s/1800s. The S-leut Hutterites of South Dakota are descendants of only 64 Hutterite ancestors. More than 12 000 individuals are included in the complete genealogy. We considered a sample of 310 atopic cases and 391 controls, described in Bourgain *et al*.[46] The entire genealogy of this sample could be constructed from the large Hutterite pedigree, yielding a 1623-person pedigree that included all known ancestors of the sample. The mean inbreeding in the 701-individual sample considered was 0.033 with SD = 0.015, and the mean kinship coefficient was 0.043 with SD = 0.033. We performed our own simulations using the real genealogy and status of the 701 individuals, as for the GAW12 data.

When the probability to detect a false association is fixed at 5%, it is in fact greater than 10%, in both inbred (Hutterites) and noninbred samples (GAW12) with related individuals. Newman *et al*[47] have illustrated the same problem in the case of quantitative trait analysis in Hutterite samples. Considering phenotypes such as IgE level, LDL or BMI and regressing the trait on age, sex and genotype, these authors showed how ignoring the pedigree structure increases the type I error by a factor of 10–20 (they observed an empirical type I error of 10–22% for a nominal type I error of 1%). Génin and Clerget-Darpoux[48] focused on the problem of inbreeding in samples of independent individuals. When cases and controls are not correctly matched on the basis of inbreeding, the type I error of the CC-χ^2_{genotype} is modestly inflated. However, when cases and controls are correctly matched for inbreeding, inbreeding can increase the power of the test, especially for recessive or quasi-recessive disease susceptibility factors.

Caution with permutation procedures to assess significance

To perform valid tests, one might consider using classical test statistics and get accurate *P*-values by permutation strategies. However, as already outlined in the context of linkage analysis, existing correlations among individuals must be maintained when conducting a permutation procedure. In particular, the classical permutation test for qualitative traits, where statuses are randomly reassigned to genotypes to create dummy 'null case control samples' on which the statistic is computed (the permutation being repeated a large number of times to get the distribution of the statistic when there is no allele frequency difference between cases and controls), might not be valid. As an illustration, we present in Table 1 the type I error of this permutation strategy for the CC-χ^2_{allele} and CC-χ^2_{genotype} tests, in the Hutterite and GAW12 samples. Empirical type I errors are much larger than the expected 5%, demonstrating that such permutation strategies do not protect against spurious conclusions. The different solutions for performing valid tests of association depend on the amount of genealogical information available. We start by presenting the methods usable in the absence of genealogical information and end with methods that require the knowledge of the entire pedigree.

Genomic controls

Devlin and Roeder[49] proposed the use of genomic controls (GC) to prevent from spurious signal detection in association studies. The primary concern of these authors was to control for population stratification, but they also recognized the impact of what they called 'cryptic relatedness' (unrecognized relationship among some individuals in the sample) on association studies. The general principle of GC approaches relies on the demonstration by Devlin and Roeder,[49] that the effects of cryptic relatedness and population substructure on test statistics of interest are essentially constant across the genome, under certain conditions, and do not vary with individual locus properties (number of alleles, allele frequencies). Consequently, the test statistic inflation due to cryptic relatedness is the same for all markers throughout the genome. These authors suggested the use of null markers (eg, polymorphisms unlikely to affect susceptibility) across the genome to estimate the effects of confounding and to remove these effects from the association test statistics. In practice, when there is no association between the marker and the disease but cryptic relatedness is present, the CC-χ^2_{allele} statistic follows a χ^2 distribution multiplied by a scaling factor λ. λ is constant over the genome and only depends on the relationship between all the individuals of the sample. λ is estimated using a robust estimator based on the median[49] value or on the mean value[50] of the CC-χ^2_{allele} statistics over all the control markers. Bacanu *et al*[51] suggested that 70 markers should be used for a good estimation of λ. In their original paper, Devlin and Roeder proposed the GC approach for the Armitage's trend test,[52,53] a genotype-based association test that is robust to departure from HW but makes the assumption of additive effects for the two alleles of an individual. However, the GC principle is applicable to a wide class of statistics for marker association testing. Tzeng *et al*[54] have recently proposed a method using GC in haplotype-based case–control analysis.

In 2002, Bacanu *et al*[55] adapted GC for association studies of quantitative traits, using linear regression. A phenotypic trait is regressed on different environmental covariates and on one or several genotype covariates with possible interactions terms. Testing for association reduces to test whether the regression coefficients for the genotype covariates are significantly different from 0 using a *t*-test. As for the qualitative traits, Bacanu *et al*[55] have shown that the inflation factor of the *t*-test, λ, due to population substructure or cryptic relatedness, only depends on the sample composition and not on the properties of the individual loci. λ can be estimated, as in the qualitative trait, using a robust estimator based on the median value of the *t*-test over all the control markers. The same principle can be used while simultaneously testing the effect of multiple loci with *F*-statistics.

TDT and related approaches

When parents are available, family-based association tests such as the TDT[56] may also be considered though Spielman and Ewens[57] underlined that the TDT is not a valid test of association if the families have affected members in multiple generations. Génin *et al*[58] extended this result to samples of independent case–parent trios where cases or parents are inbred. However, the TDT remains a valid test for linkage and its power increases with the strength of

association. To overcome the problem of meiose dependence that can arise when related affected are analysed simultaneously, different extensions of the TDT have been proposed, where an empirical robust variance is computed (see Clayton[59] for the case of multiple affected sibs or Lake *et al*[60] in more general situations such as pedigrees including multiple nuclear families). The FBAT software[60–62] implements this latter approach and allows the inclusion of any type of family. The PDT[63] is based on a similar strategy, but only considers two types of informative families (trios with an affected child and both parents genotyped and discordant sibships of at least one affected and one unaffected sib with different genotypes) and discards others even though they could bring additional information. This robust variance approach does not require the knowledge of the precise genealogical links between the individuals. However, when all the affected individuals in a sample are correlated within a single pedigree (as it is the case for instance with the Hutterite pedigree) and declared as such, the robust variance approach has virtually no power. Large pedigrees must thus be broken into smaller subpedigrees to use this approach. Neglecting the correlations between the different subpedigrees should not introduce a detectable bias in the test provided that most parent genotypes are available. This still needs to be demonstrated and the optimal breaking strategy remains to be defined.

Different extensions of the TDT have been proposed for haplotypes analyses where multiple linked loci are considered simultaneously. Apart from the problem of ambiguities in haplotypes assignments (see, for instance, Clayton[64] or Zhao *et al*[65]), haplotypes analyses pose a multiple-testing problem that can be solved as in Clayton and Jones,[59] who propose to test the global null hypothesis that all haplotype effects are 0. However, if the number H of haplotypes is large, this test may lack power as the number of degrees of freedom is $H-1$ and Clayton and Jones[59] proposed to group haplotypes based on a similarity index defined as the length, around a focal point, of the continuous region over which haplotypes are identical by state. Bourgain *et al*[66] proposed another approach, the Maximum Identity Length Contrast (MILC), which compares the mean lengths of haplotype identity among all transmitted haplotypes and among all nontransmitted haplotypes. More recently, Seltman *et al*[67] proposed a grouping of haplotype based on their phylogeny and Zhang *et al*[68] generalized the MILC approach. The impact of including related cases has only been evaluated in the context of MILC,[69] where the authors have shown that closely related cases may be analysed simultaneously provided that one is not the parent of another. Indeed the null hypothesis tested by MILC is the absence of any genetic risk factor involved in the disease in the studied region and not a composite null hypothesis of no association or no linkage as for the TDT or related family-based association tests.

To our knowledge, there is no systematic power comparison of the TDT (or related approaches) and the GC tests in the context of cryptic relatedness. Bacanu *et al*[51] extensively compare the two approaches for qualitative traits, in the presence of population stratification. They show that GC performs better than the TDT as long as the scenario is different from 'a few highly differentiated subpopulations'. For the case of cryptic relatedness, they only note that the GC adjustment when notable levels of kinship are present in the sample has a substantial cost in power, because λ can become quite large. They suggest that family-based methods are likely to be more powerful in such situations. However, the need for genotype information on family members, such as parents or sibs, for the TDT to be most powerful, can drastically reduce the number of cases eligible for a study, a concern that may be particularly relevant for late-onset diseases. By allowing the recruitment of larger samples, case–control strategies may thus prove to be more efficient.

Association tests when the entire genealogy is available

When the genealogy is entirely known, it is preferable to use this information. Bourgain *et al*[46] have recently proposed two methods for case–control association studies, suitable for any set of related individuals, provided that their genealogy is known: the case–control quasi-likelihood score test (CC-QLS) and the case–control corrected χ^2 test (CC-corrχ^2). The methods are suitable for large inbred pedigrees. The principle of the CC-corrχ^2 is similar to GC as it consists in the computation of a correction factor, λ, for the CC-χ^2_{allele} test. The difference is that the value of λ is derived analytically using the extensive pedigree information. Bourgain *et al*[46] showed that λ depends only on the values of all the kinship and inbreeding coefficients of the individuals included in the sample. CC-QLS and CC-corrχ^2 have similar forms, but the CC-QLS actually compares allele frequencies in cases and controls estimated while taking into account the known correlations among the individuals, whereas CC-corrχ^2 compares allele frequencies estimated by simply counting the number of alleles in each group. Bourgain *et al*[46] showed that CC-QLS is asymptotically locally more powerful than CC-corrχ^2. We present in Table 1 the empirical type I errors of these tests in the Hutterite and GAW12 samples. They are not significantly different from the nominal type I errors, showing that these tests correctly control for the presence of correlations among the individuals in the samples, provided that genealogical data are exhaustive. No power comparison of these latter approaches with GC has been published yet. However, when extensive genealogical information is available, exact computation of the correction factor as in CC-QLS or CC-

corrχ^2 should be more powerful. A test similar to CC-corrχ^2 was used by Gretarsdottir *et al*[14] and Styrkarsdottir *et al*[17] in the Icelandic population.

Abney *et al*[44] have proposed two different methods to test for association of quantitative traits in samples with known genealogy using the linear regression framework. Here again, the method roughly corresponds to an exact derivation of the inflation factor λ of the *t*-test described in Bacanu *et al*[55] instead of an estimation through control markers as in the case of GC. More specifically, Abney *et al* proposed the 'allele-specific HBD' method (ASHBD) and the 'general two-allele model' method (GTAM). The ASHBD uses the same framework as the HBD linkage method described above, but models the effect of a particular allele rather than the effect of a main locus. The method also uses multipoint marker information to estimate the HBD status. Both inbreeding and a recessive effect of the allele at the locus tested are required for the ASHBD method to be of interest. GTAM is a single point method suitable for general genetic models and for inbred or outbred pedigrees. The model does not include an HBD status covariate but a covariate, indicating the number of alleles of a particular type at the locus tested and a second covariate modeling the genotypic effect. To test whether these two covariates have regression coefficients significantly different from zero, the method corresponds to an F-test corrected exactly for the known correlations among the individuals. The permutation procedure proposed to correct the HBD linkage test for multiple testing over the genome is also applicable to both the ASHBD and GTAM tests.

Conclusion

In reviewing the literature on isolated populations, the number of papers praising their merits for mapping genes involved in complex disease susceptibility is impressive. But it is also striking to note the lack of discussion about the need for specific statistical methods to correctly and best search for genetic risk factors in these populations. The review of the statistical methods, presented in this study, is not intended to be exhaustive. We only tried to present some of the available methods, their advantages and limits, for both linkage and association studies, and considering qualitative and quantitative complex traits. Our intention was to highlight the fact that if the features of population isolates may facilitate the identification of susceptibility factors for complex traits, these studies deserve particular care in the choice and design of statistical methods. To illustrate how the use of methods designed for outbred samples could lead to false conclusions when applied to isolated populations, we considered two extreme examples: a highly inbred isolate (the Hutterites) and data from noninbred extended genealogies where all the affected individuals were considered, including first- and second-degree relatives. As shown by Grettarsdottir *et al*,[14] removing first- and second-degree relatives from the sample minimizes the bias in noninbred samples. However, since the number of cases may be limited in relatively small isolates, using methods allowing for an inclusion of all cases can be crucial.

Depending on whether extended genealogies are available or not, the methods that can be used differ. When the entire genealogy of the population is available, as it is the case in the Hutterite population, methods that take advantage of this information to characterize the correlations existing between the individuals and to account for them in the tests should obviously be preferred. When genealogies are not known or not accurate, then one may opt for methods that contrast what is observed at a given marker to what is observed on average over the whole genome. These GC approaches require additional genotyping of markers, but are probably a good alternative to extensive genealogical studies.

Acknowledgements

We thank Carole Ober for kindly accepting to let us use the Hutterite pedigree in our simulations and the Genetic Analysis Workshop grant number, GM31575 for GAW12 data. We thank Carole Ober, Françoise Clerget-Darpoux, Marie-Claude Babron and two anonymous reviewers for helpful comments on the manuscript.

References

1 Heutink P, Oostra BA: Gene finding in genetically isolated populations. *Hum Mol Genet* 2002; **11**: 2507–2515.
2 Horikawa Y, Oda N, Cox NJ *et al*: Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus. *Nat Genet* 2000; **26**: 163–175.
3 Hugot JP, Chamaillard M, Zouali H *et al*: Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* 2001; **411**: 599–603.
4 Peltonen L: Positional cloning of disease genes: advantages of genetic isolates. *Hum Hered* 2000; **50**: 66–75.
5 Shifman S, Darvasi A: The value of isolated populations. *Nat Genet* 2001; **28**: 309–310.
6 Escamilla MA: Population isolates: their special value for locating genes for bipolar disorder. *Bipolar Disord* 2001; **3**: 299–317.
7 Wright AF, Carothers AD, Pirastu M: Population choice in mapping genes for complex diseases. *Nat Genet* 1999; **23**: 397–404.
8 Gulcher JR, Kong A, Stefansson K: The role of linkage studies for common diseases. *Curr Opin Genet Dev* 2001; **11**: 264–267.
9 Genin E, Clerget-Darpoux F: Consanguinity and the sib-pair method: an approach using identity by descent between and within individuals. *Am J Hum Genet* 1996; **59**: 1149–1162.
10 Leutenegger AL, Genin E, Thompson EA, Clerget-Darpoux F: Impact of parental relationships in maximum lod score affected sib-pair method. *Genet Epidemiol* 2002; **23**: 413–425.
11 McPeek MS: Optimal allele-sharing statistics for genetic mapping using affected relatives. *Genet Epidemiol* 1999; **16**: 225–249.
12 Thorgeirsson TE, Oskarsson H, Desnica N *et al*: Anxiety with panic disorder linked to chromosome 9q in Iceland. *Am J Hum Genet* 2003; **72**: 1221–1230.
13 Sveinbjornsdottir S, Hicks AA, Jonsson T *et al*: Familial aggregation of Parkinson's disease in Iceland. *N Engl J Med* 2000; **343**: 1765–1770.

14 Gretarsdottir S, Thorleifsson G, Reynisdottir ST *et al*: The gene encoding phosphodiesterase 4D confers risk of ischemic stroke. *Nat Genet* 2003; **35**: 131–138.
15 Stefansson SE, Jonsson H, Ingvarsson T *et al*: Genomewide scan for hand osteoarthritis: a novel mutation in matrilin-3. *Am J Hum Genet* 2003; **72**: 1448–1459.
16 Kristjansson K, Manolescu A, Kristinsson A *et al*: Linkage of essential hypertension to chromosome 18q. *Hypertension* 2002; **39**: 1044–1049.
17 Styrkarsdottir U, Cazier JB, Kong A *et al*: Linkage of osteoporosis to chromosome 20p12 and association to BMP2. *PLoS Biol* 2003; **1**: E69.
18 Bjornsson A, Gudmundsson G, Gudfinnsson E *et al*: Localization of a gene for migraine without aura to chromosome 4q21. *Am J Hum Genet* 2003; **73**: 986–993.
19 Gretarsdottir S, Sveinbjornsdottir S, Jonsson HH *et al*: Localization of a susceptibility gene for common forms of stroke to 5q12. *Am J Hum Genet* 2002; **70**: 593–603.
20 Reynisdottir I, Thorleifsson G, Benediktsson R *et al*: Localization of a susceptibility gene for type 2 diabetes to chromosome 5q34–q35.2. *Am J Hum Genet* 2003; **73**: 323–335.
21 Hakonarson H, Bjornsdottir US, Halapi E *et al*: A major susceptibility gene for asthma maps to chromosome 14q24. *Am J Hum Genet* 2002; **71**: 483–491.
22 Stefansson H, Sigurdsson E, Steinthorsdottir V *et al*: Neuregulin 1 and susceptibility to schizophrenia. *Am J Hum Genet* 2002; **71**: 877–892.
23 Hicks AA, Petursson H, Jonsson T *et al*: A susceptibility gene for late-onset idiopathic Parkinson's disease. *Ann Neurol* 2002; **52**: 549–555.
24 Karason A, Gudjonsson JE, Upmanyu R *et al*: A susceptibility gene for psoriatic arthritis maps to chromosome 16q: evidence for imprinting. *Am J Hum Genet* 2003; **72**: 125–131.
25 Genin E, Bellis G, Clerget-Darpoux F: Information provided by pairs of distantly affected relatives to search for genes involved in rare autosomal dominant diseases. *Ann Hum Genet* 1997; **61** (Part 1): 25–36.
26 Gudbjartsson DF, Jonasson K, Frigge ML, Kong A: Allegro, a new computer program for multipoint linkage analysis. *Nat Genet* 2000; **25**: 12–13.
27 Sobel E, Lange K: Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. *Am J Hum Genet* 1996; **58**: 1323–1337.
28 Cheung VG, Nelson SF: Genomic mismatch scanning identifies human genomic DNA shared identical by descent. *Genomics* 1998; **47**: 1–6.
29 Durham LK, Feingold E: Genome scanning for segments shared identical by descent among distant relatives in isolated populations. *Am J Hum Genet* 1997; **61**: 830–842.
30 Mirzayans F, Mears AJ, Guo SW, Pearce WG, Walter MA: Identification of the human chromosomal region containing the iridogoniodysgenesis anomaly locus by genomic-mismatch scanning. *Am J Hum Genet* 1997; **61**: 111–119.
31 Nelson SF, McCusker JH, Sander MA, Kee Y, Modrich P, Brown PO: Genomic mismatch scanning: a new approach to genetic linkage mapping. *Nat Genet* 1993; **4**: 11–18.
32 Smalley SL, Woodward JA, Palmer CG: A general statistical model for detecting complex-trait loci by using affected relative pairs in a genome search. *Am J Hum Genet* 1996; **58**: 844–860.
33 Leutenegger AL, Prum B, Genin E *et al*: Estimation of the inbreeding coefficient through use of genomic data. *Am J Hum Genet* 2003; **73**: 516–523.
34 Almasy L, Blangero J: Multipoint quantitative-trait linkage analysis in general pedigrees. *Am J Hum Genet* 1998; **62**: 1198–1211.
35 Heath SC, Snow GL, Thompson EA, Tseng C, Wijsman EM: MCMC segregation and linkage analysis. *Genet Epidemiol* 1997; **14**: 1011–1016.
36 Williams-Blangero S, VandeBerg JL, Subedi J *et al*: Genes on chromosomes 1 and 13 have significant effects on Ascaris infection. *Proc Natl Acad Sci USA* 2002; **99**: 5533–5538.
37 Blangero J, Williams JT, Almasy L: Robust LOD scores for variance component-based linkage analysis. *Genet Epidemiol* 2000; **19** (Suppl 1): S8–S14.
38 Abney M, McPeek MS, Ober C: Estimation of variance components of quantitative traits in inbred populations. *Am J Hum Genet* 2000; **66**: 629–650.
39 Amos CI: Robust variance-components approach for assessing genetic linkage in pedigrees. *Am J Hum Genet* 1994; **54**: 535–543.
40 Sham PC, Purcell S, Cherny SS, Abecasis GR: Powerful regression-based quantitative-trait linkage analysis of general pedigrees. *Am J Hum Genet* 2002; **71**: 238–253.
41 Chapman NH, Leutenegger AL, Badzioch MD *et al*: The importance of connections: joining components of the Hutterite pedigree. *Genet Epidemiol* 2001; **21** (Suppl 1): S230–S235.
42 Dyer TD, Blangero J, Williams JT, Goring HH, Mahaney MC: The effect of pedigree complexity on quantitative trait linkage analysis. *Genet Epidemiol* 2001; **21** (Suppl 1): S236–S243.
43 Falchi M, Forabosco P, Mocci E *et al*: A genomewide search using an original pairwise sampling approach for large genealogies identifies a new locus for total and low-density lipoprotein cholesterol in two genetically differentiated isolates of Sardinia. *Am J Hum Genet* 2004; **75**: 1015–1031.
44 Abney M, Ober C, McPeek MS: Quantitative-trait homozygosity and association mapping and empirical genomewide significance in large, complex pedigrees: fasting serum-insulin level in the Hutterites. *Am J Hum Genet* 2002; **70**: 920–934.
45 Iturria SJ, Williams JT, Almasy L, Dyer TD, Blangero J: An empirical test of the significance of an observed quantitative trait locus effect that preserves additive genetic variation. *Genet Epidemiol* 1999; **17** (Suppl 1): S169–S173.
46 Bourgain C, Hoffjan S, Nicolae R *et al*: Novel case–control test in a founder population identifies P-selectin as an atopy-susceptibility locus. *Am J Hum Genet* 2003; **73**: 612–626.
47 Newman DL, Abney M, McPeek MS, Ober C, Cox NJ: The importance of genealogy in determining genetic associations with complex traits. *Am J Hum Genet* 2001; **69**: 1146–1148.
48 Genin E, Clerget-Darpoux F: Association studies in consanguineous populations. *Am J Hum Genet* 1996; **58**: 861–866.
49 Devlin B, Roeder K: Genomic control for association studies. *Biometrics* 1999; **55**: 997–1004.
50 Reich DE, Goldstein DB: Detecting association in a case–control study while correcting for population stratification. *Genet Epidemiol* 2001; **20**: 4–16.
51 Bacanu SA, Devlin B, Roeder K: The power of genomic control. *Am J Hum Genet* 2000; **66**: 1933–1944.
52 Armitage P: Tests for linear trends in proportions and frequencies. *Biometrics* 1955; **11**: 375–386.
53 Sasieni PD: From genotypes to genes: doubling the sample size. *Biometrics* 1997; **53**: 1253–1261.
54 Tzeng JY, Devlin B, Wasserman L, Roeder K: On the identification of disease mutations by the analysis of haplotype similarity and goodness of fit. *Am J Hum Genet* 2003; **72**: 891–902.
55 Bacanu SA, Devlin B, Roeder K: Association studies for quantitative traits in structured populations. *Genet Epidemiol* 2002; **22**: 78–93.
56 Spielman RS, McGinnis RE, Ewens WJ: Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* 1993; **52**: 506–516.
57 Spielman RS, Ewens WJ: The TDT and other family-based tests for linkage disequilibrium and association. *Am J Hum Genet* 1996; **59**: 983–989.
58 Genin E, Todorov AA, Clerget-Darpoux F: Properties of the transmission-disequilibrium test in the presence of inbreeding. *Genet Epidemiol* 2002; **22**: 116–127.
59 Clayton D, Jones H: Transmission/disequilibrium tests for extended marker haplotypes. *Am J Hum Genet* 1999; **65**: 1161–1169.

60 Lake SL, Blacker D, Laird NM: Family-based tests of association in the presence of linkage. *Am J Hum Genet* 2000; **67**: 1515–1525.
61 Horvath S, Xu X, Laird NM: The family based association test method: strategies for studying general genotype–phenotype associations. *Eur J Hum Genet* 2001; **9**: 301–306.
62 Rabinowitz D, Laird N: A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. *Hum Hered* 2000; **50**: 211–223.
63 Martin ER, Monks SA, Warren LL, Kaplan NL: A test for linkage and association in general pedigrees: the pedigree disequilibrium test. *Am J Hum Genet* 2000; **67**: 146–154.
64 Clayton D: A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission. *Am J Hum Genet* 1999; **65**: 1170–1177.
65 Zhao H, Zhang S, Merikangas KR *et al*: Transmission/disequilibrium tests using multiple tightly linked markers. *Am J Hum Genet* 2000; **67**: 936–946.
66 Bourgain C, Genin E, Quesneville H, Clerget-Darpoux F: Search for multifactorial disease susceptibility genes in founder populations. *Ann Hum Genet* 2000; **64**: 255–265.
67 Seltman H, Roeder K, Devlin B: Transmission/disequilibrium test meets measured haplotype analysis: family-based association analysis guided by evolution of haplotypes. *Am J Hum Genet* 2001; **68**: 1250–1263.
68 Zhang S, Sha Q, Chen HS, Dong J, Jiang R: Transmission/disequilibrium test based on haplotype sharing for tightly linked markers. *Am J Hum Genet* 2003; **73**: 566–579.
69 Bourgain C, Genin E, Holopainen P *et al*: Use of closely related affected individuals for the genetic study of complex diseases in founder populations. *Am J Hum Genet* 2001; **68**: 154–159.
70 Almasy L, Terwilliger JD, Nielsen D, Dyer TD, Zaykin D, Blangero J: GAW12: simulated genome scan, sequence, and family data for a common disease. *Genet Epidemiol* 2001; **21** (Suppl 1): S332–S338.