

Practical Considerations for Dividing Data into Subsets Prior to PPL Analysis

M. Govil^a V.J. Vieland^b

^aDepartment of Oral Biology and Center for Craniofacial and Dental Genetics, School of Dental Medicine, University of Pittsburgh, Pittsburgh, Pa., ^bBattelle Center of Mathematical Medicine, The Research Institute at Nationwide Children's Hospital and The Ohio State University, Columbus, Ohio, USA

Key Words

Complex traits · Linkage analysis · Genetic mapping · Posterior probability of linkage · PPL · Sequential updating · Heterogeneity · Clinically defined subsets · Bayesian linkage analysis

Abstract

Objective: The PPL, a class of statistics for complex trait genetic mapping in humans, utilizes Bayesian sequential updating to accumulate evidence for or against linkage across potentially heterogeneous data (sub)sets. Here, we systematically explore the relative efficacy of alternative subsetting approaches for purposes of PPL calculation. **Methods:** We simulated genotypes for three pedigree sets (sib pairs; 2–3 generations; ≥ 4 generations) based on families from an ongoing study. For each pedigree set, 100 replicates were generated under different levels of heterogeneity (1000 under 'no linkage'). Within each replicate, updating was performed across subsets defined randomly (RAND2, RAND4), by true (TRUE) linkage status, with a realistic (REAL) classification, by individual pedigree (PED), or without any subsetting (NONE). **Results:** Under 'linkage', REAL yields larger PPLs compared to NONE, RAND2, RAND4, or PED. Under 'no linkage', RAND2, RAND4 and PED yield PPLs close to NONE. **Conclusions:** We have examined the impact of different subsetting strategies on the sampling behavior of the PPL. Our results underscore the utility of finding variables that can help delineate more

homogeneous data subsets and demonstrate that, once such variables are found, sequential updating can be highly beneficial in the presence of appreciable heterogeneity at a linked locus, without inflation at an unlinked locus.

Copyright © 2008 S. Karger AG, Basel

Introduction

The PPL is a class of statistics designed to address complications of genetic mapping for complex traits in humans [1, 2]. The PPL differs from other approaches in several respects, but of primary interest here is its use of Bayesian sequential updating to accumulate evidence for or against linkage across multiple, potentially heterogeneous sets of data. (For ease of exposition, we will assume throughout this discussion that the analysis of interest is two-point sex-averaged linkage analysis of a dichotomous trait. The PPL can at present also be used for multipoint analysis [3], quantitative trait analysis [4], quantitative trait threshold analysis [5], gene \times gene interaction analysis [6], and linkage disequilibrium mapping [7]. The statistic has also been extended to allow sex-specific two-point [8] and multipoint [9] analyses. Because the mechanism of sequential updating is the same in all cases, however, we do not address these forms of the PPL separately in this paper.)

Sequential updating works by accumulating the posterior density for the recombination fraction (in the case of two-point analysis) across data subsets, integrating out nuisance parameters of the trait model within each subset prior to updating the posterior density across the subsets. This procedure allows for the fact that parameters such as allele frequencies, penetrances, and the proportion of 'linked' families in a data set may vary across data sets, while the location of an underlying gene remains constant; and at the same time, because the trait parameters are integrated out rather than maximized over, there is no inherent 'penalty' or inflation in the distribution of the statistic as an artifact of the number of parameters included in the model. We have shown previously that sequential updating has distinct advantages over either 'pooling' results across data subsets in the presence of appreciable heterogeneity, or analyzing subsets independently [10, 11].

The original motivation for using sequential updating in the context of linkage analysis involved study designs in which a new set of families is used to follow up on a linkage signal obtained in a previous data set [12]. However, a variation on this theme is the use of sequential updating as a technique for measuring the aggregate linkage evidence *within a single data set*, when the families can be sorted into subsets based on clinical, geographic, or other features. For instance, if a particular clinical characteristic is present in some but not all families in the data set, and if investigators suspect that this characteristic might distinguish two genetic forms of the disease, then the overall evidence at a given point in the genome can be measured in the first clinically defined group, and then sequentially updated over the remaining families.

Whether the evidence should be accumulated across the two groups, or alternatively, whether the two clinical forms should be considered as two different diseases and analyzed separately, is a scientific question; the optimal procedure is the one that best matches the underlying biology, which is generally not known at time of data analysis. Thus the 'best' approach to subdividing a data set is to follow what the investigators believe to be the most compelling scenario. For example, if the clinical feature in question has proved to demarcate locus heterogeneity in the mouse, the possibility of two distinct genetic forms in the human warrants serious attention; on the other hand, if the characteristic is known to be the result of variation in an environment factor, then subsetting on its basis may not be warranted. In the absence of clear-cut evidence in either direction, the decision as to whether

and how much to subdivide the sample falls to the judgment of the investigators.

The immediate motivation for this paper was analysis of a large set of pedigrees ascertained for the presence of multiple individuals with cleft lip with or without cleft palate (CL/P) [13], for which we needed to make decisions regarding subsetting. CL/P is known to be a complex disease, in the sense that no single major gene has been discovered although the evidence of genetic etiology is strong [14], and environmental influences appear to play a role as well [see e.g., 15, 16]. We therefore felt it was safe to assume the presence of multiple CL/P-related genes in the data set, with appreciable variation across subsets of the data with respect to the relative importance of the different genes (including differences in allele frequencies, relevant environmental exposures, etc.). The question before us was how best to define subsets of the data, and how many subsets to consider? And while there were some obvious ways to subdivide the sample (e.g., by country of origin), there were several other possibilities as well (e.g., by various clinical features of potential but unknown genetic relevance).

This application raised several very practical questions regarding potential implications of how one breaks up a large dataset into smaller subsets for purposes of sequential updating. For instance, is it important to maintain a certain subset size, e.g., are larger subsets inherently better (say, more stable, with smaller variance) than smaller ones? Do the effects of sequential updating depend on the sizes of the individual pedigrees within the subsets, e.g., can subsetting introduce artifacts into the analysis when the pedigrees are very small? And what would happen if a variable chosen as the basis for the subsetting was in fact 'random' with respect to the underlying genetics, that is, if it did not demarcate more homogeneous subsets of the data? Finally, how careful do we have to be to avoid subsetting on a variable that we view as only somewhat likely to relate to underlying heterogeneity?

To date, we have not systematically explored the relative efficacy of alternative approaches to subsetting per se for purposes of PPL calculation. This is in part because – as a matter of philosophical principle – we view the optimal form of the PPL as the one that best reflects the underlying biology, rather than, say, the form with the best sampling characteristics. As a practical matter, however, we felt that considering the sampling behavior of the PPL under alternative approaches to subsetting was important, especially in the event that some practices might result in undesirable artifacts, in order to guide subsetting decisions in real applications.

Table 1. Characteristics of pedigree sets

Pedigree set	Total pedigrees in set	Pedigree size distribution avg [min, max]	Proportion of linked pedigrees across 100 replicates avg % [min, max]		ELOD across 100 replicates	
			$\alpha = 30\%$	$\alpha = 50\%$	$\alpha = 30\%$	$\alpha = 50\%$
SmallPed	140	4.00 [4, 4]	30 [19,41]	49 [40,61]	1.0510	4.3070
ModeratePed	114	7.66 [4, 16]	29 [20,47]	49 [36,61]	1.5984	5.5203
LargePed	104	18.12 [8, 53]	30 [18,40]	50 [40,62]	5.5611	15.4458

In this paper we examine the impact of different subsetting strategies on the sampling behavior of the PPL. In the process, we also document more systematically than had been done previously that sequential updating can be highly beneficial in the presence of inter-sample heterogeneity at a linked locus, without any inflation of results at an unlinked locus.

Methods

In this section we describe (1) the methods used to simulate the data; (2) the methods used to analyze the data, and (3) the different subsetting schemes to be compared.

Data Simulation Methods

Pedigree data were generated based on the observed pedigrees (structure and distribution of phenotypes) from one of the populations collected in connection with the CL/P project mentioned above, which are described in detail elsewhere [13]. We chose to work with real pedigrees rather than simulated structures for two reasons. First, we wanted to ensure that the answers we obtained applied directly to the CL/P study; second, this data set seemed ideal for the current investigation because of the variety of pedigree structures it contains, varying from 4-person pedigrees up to a few with more than 35 individuals, some including inbreeding loops.

Fixing the observed pedigree structures and phenotypes, we then simulated marker data using SLINK [17, 18]. We picked one marker from the original data as a model for missing information, and set any individual who was missing a genotype in the real data as missing in the simulation as well. (While the selected marker was one with a positive linkage signal in a preliminary analysis; in the present context, the choice of the marker – which serves only to give us a reasonable pattern of missing data – is arbitrary. Note that no follow-up genotyping had been done to fill in missing information subsequent to linkage analysis.) Marker genotypes were simulated under a single-locus model, allowing for locus heterogeneity (see below). The recombination fraction was fixed at 0.01; the disease allele frequency was 0.001; and the three penetrances were 0.40, 0, and 0 for DD, Dd, and dd respectively. Again, the choice of generating model is arbitrary and is not expected to have an impact on the comparisons among alternative subsetting procedures. What drives the behavior of the PPL in general is the magnitude of the expected posterior probability (or

Bayes Ratios, BR; see Appendix A) rather than the particular generating model [8]. Since the three pedigree sets under these two levels of heterogeneity give rise to a wide range of BR (see Results below), we can generalize from these results without the need to consider alternative versions of the generating trait model. The choice was, however, based on the maximizing model (MOD) for the selected marker. The simulated marker had 11 alleles, with the heterozygosity fixed at 80% to match the observed heterozygosity in the 13-allele real marker.

The value of the admixture parameter α (proportion of ‘linked’ families [19]) was set at 30, 50, or 100% (no heterogeneity). In the last case, however, the PPL was 1.0 in all replicates regardless of subsetting procedures, with MOD scores [20–22] ranging from 14–75. Thus we consider only the lower two levels of heterogeneity in what follows. In all cases, pedigrees were randomly assigned to the ‘linked’ or ‘unlinked’ groups based on the given probability (α), giving rise to considerable variability across replicates with respect to the true proportion linked in any given subset.

In order to consider the effect of pedigree size on subsetting, the 218 pedigrees were divided into three groups based on pedigree size: LargePed ($n = 104$), with four or more generations each; ModeratePed ($n = 114$), comprising the remaining pedigrees with <4 generations; and SmallPed ($n = 140$), which was created from the 12 affected sib pair (ASP) pedigrees in the original data set by randomly repeating these pedigrees while preserving the pattern of missing information from the 12 observed ASPs. The number of the original ASPs with missing genotypes for none, one, or both parents was 4, 6, 2, respectively.

For each type of data set (LargePed, ModeratePed, and SmallPed), 100 replicates were generated at each of the specified levels of α . To study what happens under ‘no linkage’, an additional 1,000 replicates were generated for each type at a marker simulated at recombination distance of 0.50.

Table 1 summarizes characteristics of the three pedigree sets. The average observed proportion of linked pedigrees for each set corresponds well with the generating levels of α , with considerable variability across replicates. The table also shows expected LOD scores (ELOD) [23] across the 100 replicates, computed at the generating trait model, which illustrate that the three pedigree sets do in fact differ appreciably in the amount of linkage information they contain. Overall, we are considering data sets with ELOD values ranging from 1 to over 15.

The time required to compute the PPL highlights another difference in the complexity of the three pedigree sets. Even on very fast nodes in our Linux cluster (dual AMD Athlons) it required approximately 1–2 h to complete a replicate for the ModeratePed and SmallPed, and approximately 11 h to complete a single repli-

cate for LargePed. Work on computational efficiency for all forms of the PPL is ongoing [24–26].

Data Analytic Methods

In what follows, we report means and standard deviations across the 100 replicates for each generating condition. In comparing results across subsetting schemes, statistical significance is assessed at the 5% significance level using paired sample t tests (for differences in the means) or variance ratio F tests (for differences in the variances).

The PPL has been described in detail elsewhere [1–12, 27, 28]. Relevant mathematical details are given in Appendix A. All PPL values, which represent the probability of linkage, are reported as percentages. By convention, we round a PPL $>2\%$, which is the stipulated prior probability of linkage [1, 29] based on a genome-scan design, to the nearest whole number; while any PPL $<2\%$ is reported to two decimal places. Note that the prior probability of linkage is in fact irrelevant in these simulations, for which the probability is either 1 or 0 by stipulation, depending on the generating scheme. However, we retain the usual form of the PPL as used in genome-wide linkage scans in order to maintain the usual scale.

Alternative Subsetting Schemes

We consider six different approaches to subsetting within each data set: (i) *no subsetting* (NONE), in which each replicate is analyzed as a single group; (ii) *two random subsets* (RAND2), in which each replicate is randomly divided into two (approximately) equally sized groups; (iii) *four random subsets* (RAND4), in which each replicate is randomly divided into four (approximately) equally sized groups; (iv) *'true' subsets* (TRUE), in which each replicate is divided into two groups, one comprising all and only the 'linked' pedigrees within that replicate (or the *linked subset*, LINK_{TRUE}), and the other comprising the remaining (unlinked) pedigrees; (v) *'realistic' subsets* (REAL), intended to mimic a reasonable yet imperfect classification variable, in which each linked pedigree is given an 80% probability of being included in the first of two subsets, while each unlinked pedigree is given an 80% chance of being included in other subset; and finally (vi) an extreme form of subsetting, in which each pedigree is considered as a separate subset (PED).

In generating RAND2, RAND4, and REAL, randomness was ensured by seeding the random number generator twice, first with the process ID, and second with the first random number generated using the process ID as seed. We have verified for all subsetting schemes that the proportion of linked pedigrees per data set follows the intended generating distribution.

The rationale for these subsetting schemes is as follows: NONE serves as a baseline, indicating the PPL that would be obtained if no subsetting was performed; RAND2 and RAND4 permit us to consider what happens when we subset on a variable that turns out to be (genetically) irrelevant, and also, whether the impact of so doing is dependent on subset sizes; TRUE serves as a model for the optimal procedure, in which subsetting is based on a clinical (or other) variable that perfectly demarcates 'linked' from 'unlinked' pedigrees, and thus represents the optimal subsetting scheme under the given generating conditions; and REAL models a more realistic situation in which imperfect knowledge exists regarding a factor for creating homogeneous subgroups of the data. PED is somewhat different from the other schemes in that we do

not view it as an option on scientific grounds in general applications; however, it is of particular interest from a mathematical point of view (see below for additional comments).

In addition to the subsetting schemes described above, we also report results for LINK_{TRUE}, the subset containing all and only the linked pedigrees in a replicate, considered by itself without any updating. This represents the maximum PPL that can be obtained for a given replicate.

Results

In this section we (1) show the effects of subsetting on a variable relevant to linkage, and consider the impact of different levels of heterogeneity and the different pedigree sets in this context; (2) consider the effects of subsetting on the basis of random (genetically irrelevant) variables, again, separately considering these effects for the different levels of heterogeneity and the different pedigree sets; (3) revisit each of these effects for data generated under 'no linkage', and finally (4) consider the performance of sequentially updating by individual pedigree under both 'linkage' and 'no linkage'.

(1) Effect of Subsetting Based on Variables Relevant to Linkage

Table 2 presents results at the linked marker, for all three data types and both levels of α . Considering first $\alpha = 30\%$ and SmallPed, we see that, as expected, the average PPL increases reading from left to right across the table. When just 30% of pedigrees on average are 'linked', then sequentially updating across TRUE subsets performs statistically significantly better than NONE, or simply pooling all families for a single analysis. While TRUE also outperforms REAL, REAL is significantly better than NONE as well. Finally, it is of interest to note that LINK_{TRUE}, which is obtained by ignoring all unlinked families, is essentially identical in performance to TRUE, illustrating the robustness of the PPL to inclusion of data sets with substantially different features (e.g., from different geographic locations with very different linkage patterns, or based on phenotypic variants that turn out to be genetically distinct).

Reading down the left-hand side of the table, we also see that while the PPL gets larger in magnitude as the pedigrees become more informative, the pattern of results within rows remains the same, with TRUE and LINK_{TRUE} performing virtually identically, TRUE performing better than REAL, and REAL performing statistically significantly better than NONE.

Table 2. Mean (standard deviation) of the PPL (%) when a relevant variable is used to subset the data at a linked locus

Pedigree set	$\alpha = 30\%$				$\alpha = 50\%$			
	NONE	REAL	TRUE	LINK _{TRUE}	NONE	REAL	TRUE	LINK _{TRUE}
SmallPed	15 (27)	25 (38)	33 (46)	33 (46)	53 (48)	54 (49)	55 (49)	55 (49)
ModeratePed	20 (33)	30 (42)	35 (45)	35 (45)	55 (48)	56 (48)	57 (47)	58 (47)
LargePed	40 (45)	45 (45)	57 (43)	58 (43)	68 (41)	71 (40)	87 (27)	88 (26)

Table 3. Mean (standard deviation) of the PPL (%) when an irrelevant variable is used to subset the data at a linked locus

Pedigree set	$\alpha = 30\%$			$\alpha = 50\%$		
	NONE	RAND2	RAND4	NONE	RAND2	RAND4
SmallPed	15 (27)	12 (23)	11 (19)	53 (48)	50 (47)	46 (45)
ModeratePed	20 (33)	17 (28)	15 (26)	55 (48)	54 (47)	50 (45)
LargePed	40 (45)	38 (44)	37 (44)	68 (41)	64 (43)	63 (44)

Finally, comparing the left-hand and right-hand sides of the table, we see that while the same patterns overall hold for both levels of α , there is a marked decrease in the difference between NONE and the various forms of updating, especially for the less informative pedigree sets. With greater homogeneity, the difference in performance between REAL and NONE is only marginally significant at best (for SmallPed and ModeratePed) or not statistically significant (LargePed). However, even here, the performance of TRUE continues to be statistically significantly better than both NONE and REAL, for all pedigree sets. It is also interesting to note that in all cases, TRUE and LINK_{TRUE} are again virtually identical.

In almost all cases, observed differences in the variances are not statistically significant, and the variances appear to be driven primarily by how close the average PPL is to a boundary (0 or 100%) rather than by the subsetting procedure itself, which is unsurprising given the bounded nature of the statistic. This pattern holds in all analyses, strongly suggesting that the approach to subsetting per se does not have a direct impact on variability in the sampling distribution of the PPL; rather, any impact on the variance is secondary to the impact on the magnitude of the PPL (on average) itself.

For a fuller view of the data, figure 1 shows the cumulative frequency distributions for the three pedigree sets, for both levels of heterogeneity. These graphs show that

over a broad range of PPL values, REAL does indeed outperform NONE for the less homogeneous data ($\alpha = 0.30$). They therefore underscore the utility of identifying clinical or other variables that can be used to demarcate more homogeneous subsets of the data. Further, with more homogeneous data, the difference between the three subsetting schemes essentially disappears for the SmallPed and ModeratePed, so that while subsetting is no longer particularly helpful, neither is it deleterious.

(2) Effect of Subsetting on Genetically Irrelevant Variables at a Linked Locus

We next consider the effect of subsetting on the basis of an irrelevant, or random, criterion. Table 3 provides results for the different pedigree sets for RAND2 and RAND4. To facilitate comparison, results for NONE are repeated in this table. Here as well, while the magnitude of the PPL increases from SmallPed to LargePed, the trend across subsets is consistent for the three pedigree sets.

For both levels of heterogeneity and all the pedigree sets, NONE outperforms both RAND2 and RAND4, and RAND2 outperforms RAND4. The differences are statistically significant in most cases (with the exception of NONE vs. RAND2 for ModeratePed and LargePed at $\alpha = 0.30$). Figure 2 shows the cumulative frequency distribution for the three pedigree sets, for both levels of hetero-

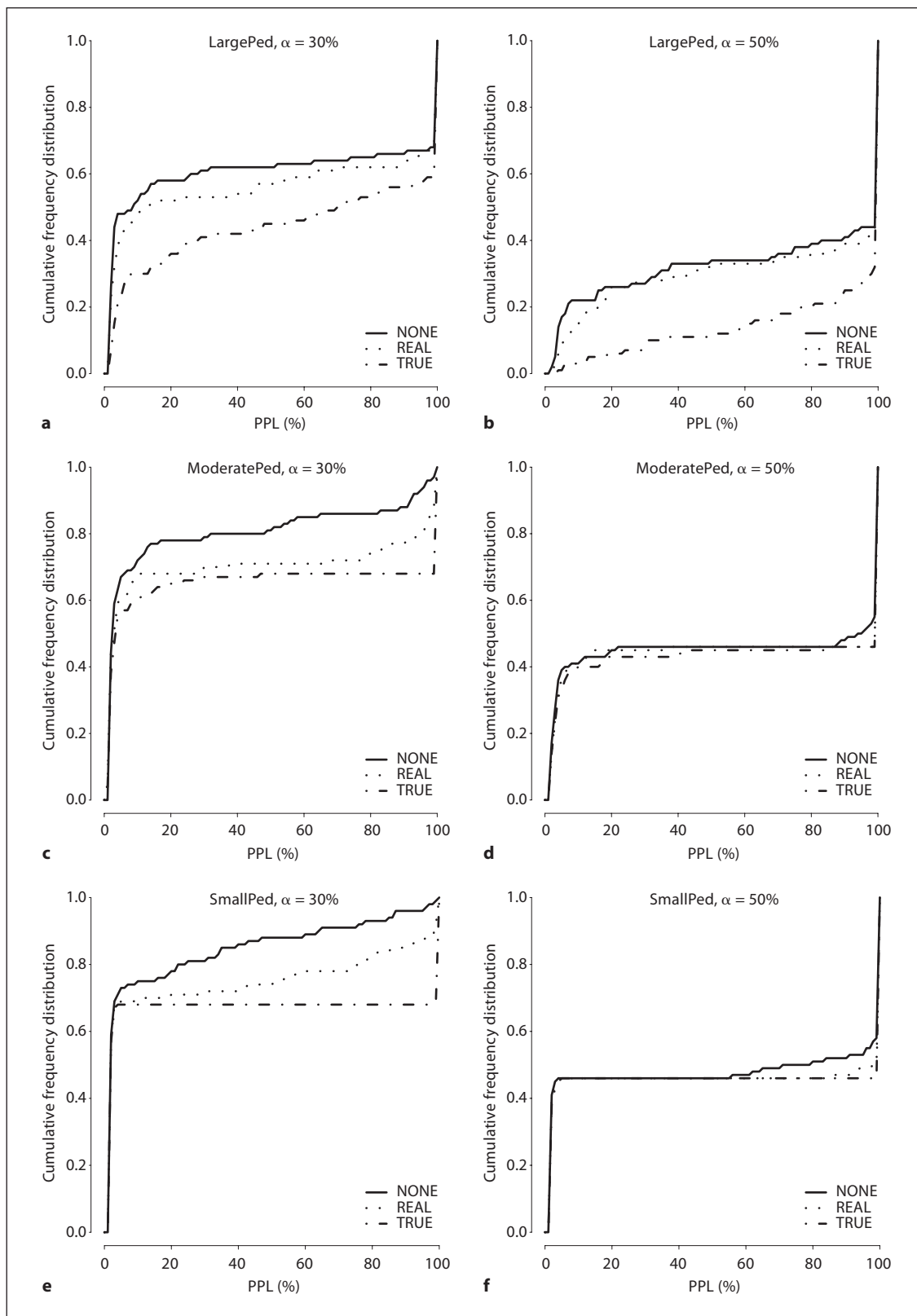


Fig. 1. a–f Cumulative frequency distribution of the PPL (%) when a relevant variable is used to subset the data at a linked locus.

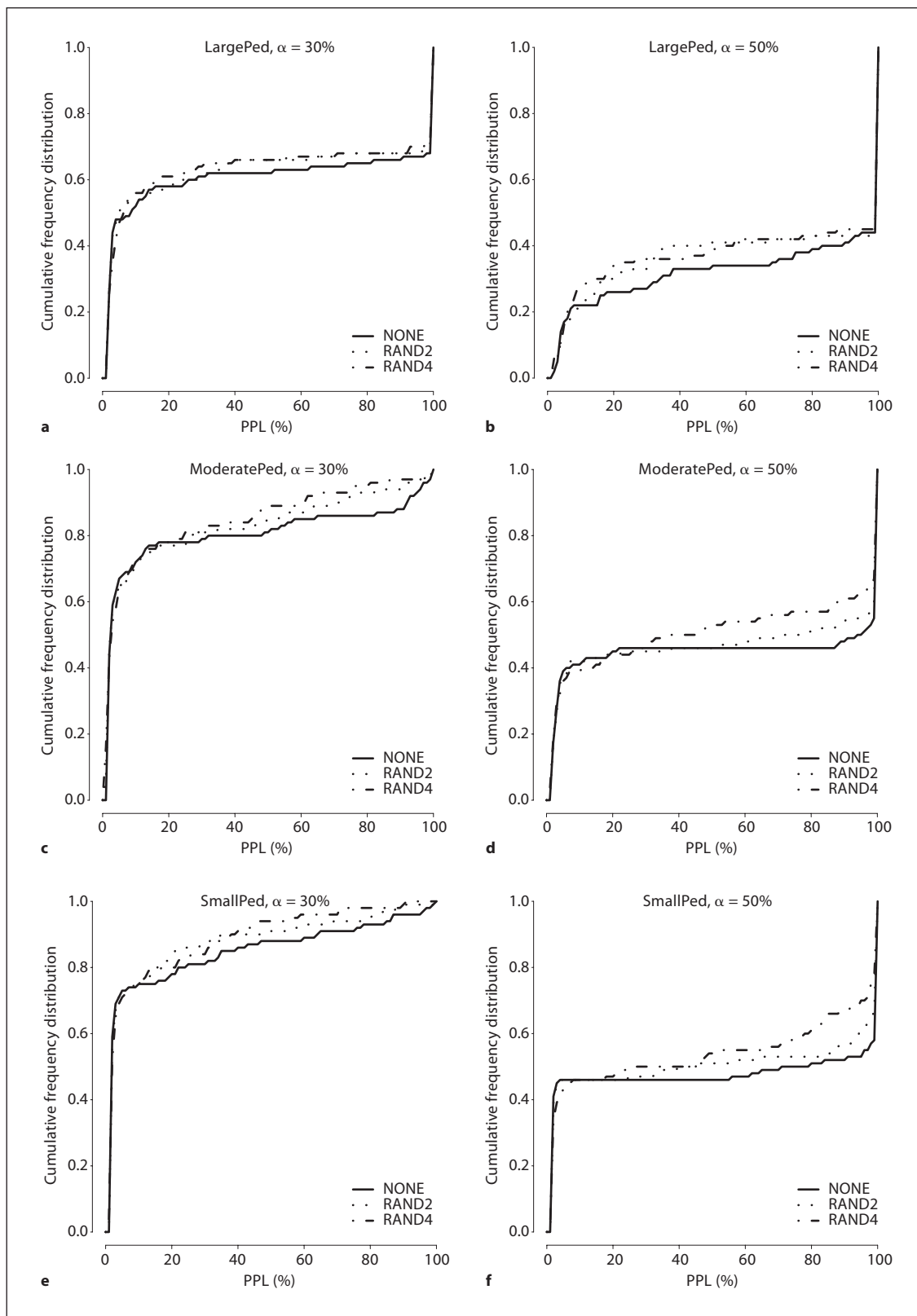


Fig. 2. a–f Cumulative frequency distribution of the PPL (%) when an irrelevant variable is used to subset the data at a linked locus.

Table 4. Mean (standard deviation) of the PPL (%) at an unlinked locus

Pedigree set	NONE	RAND2	RAND4
SmallPed	1.93 (1.50)	1.94 (1.87)	1.93 (1.35)
ModeratePed	1.76 (1.24)	1.71 (2.26)	1.70 (2.14)
LargePed	1.77 (0.97)	1.72 (1.56)	1.70 (2.43)

Table 5. Mean (standard deviation) of the PPL (%) for PED at a linked locus

Pedigree set	$\alpha = 30\%$		$\alpha = 50\%$	
	NONE	PED	NONE	PED
SmallPed	15 (27)	15 (23)	53 (48)	47 (44)
ModeratePed	20 (33)	22 (32)	55 (48)	55 (44)
LargePed	40 (45)	44 (43)	68 (41)	72 (37)

geneity. It can be seen that NONE is preferable to both RAND2 and RAND4 in all cases, although the difference between NONE and the two random forms of subsetting is not large.

These results indicate that subsetting on a variable that turns out to be genetically irrelevant can actually depress linkage signals in the presence of linkage, with the effect tending to be more deleterious as more subsets (and thus, smaller subsets) are created. Since in practice it is usually not possible to know with certainty whether the subsetting criteria being employed are relevant or irrelevant, this shows that some caution is warranted in using subsetting indiscriminately, insofar as true linkage signals can be diminished if the subsetting variable turns out to be genetically irrelevant. Thus there does not appear to be a rationale for arbitrarily dividing a large data set into smaller subsets for analysis, unless a specific genetic or clinical basis for so doing exists.

(3) Comparison Across Subsetting Schemes at an Unlinked Marker

Table 4 gives the mean PPL for all three pedigree sets at the unlinked marker. Note that in this case, there is just one ‘true’ subset, viz., the entire data set. Hence NONE and TRUE become identical procedures.

Every subsetting scheme yields a mean PPL <2%, indicating evidence against linkage. None of the pair-wise comparisons of the means are statistically significant. Figure 3 shows scatter plots for each of the random sub-

setting schemes compared to NONE, for each of the pedigree sets. As can be seen, the behavior of the PPL is similar across the subsetting schemes for all pedigree sets, with 98.8–99.8% of PPL values <10% for all cases.

We note that in the plots for ModeratePed and LargePed, there are a few ‘outlier’ replicates for which the random subsetting procedure yields a substantially larger PPL than NONE, and the variance is higher under random subsetting. Additionally, for all subsetting schemes and pedigree sets (except for RAND2 vs. RAND4 for ModeratePed), the difference in variances is statistically significant. (However, based on just 1,000 replicates the sampling distribution this far out in the tail is not represented with any precision.) While this may be another reason to be cautious in subsetting on variables that lack strong prima facie justification, nonetheless, overall these results confirm earlier work showing that subsetting, even on the basis of irrelevant variables, does not systematically lead to a higher PPL under ‘no linkage’.

(4) Effects of Treating Each Pedigree as Its Own Subset (PED)

Here we briefly consider an extreme form of subsetting (PED) in which each pedigree is considered as a separate subset and sequential updating is performed one pedigree at a time. PED is primarily of theoretical interest only. As a practical matter, PED would only seem appropriate for situations in which we would posit a ‘private’ mutation in each family (so that if it were to be used, it would be probably be reserved for very large, informative pedigrees, each sufficient to map a gene on its own), or under other conditions in which extreme locus and/or allelic heterogeneity might be anticipated. On the other hand, in this case sequential updating would be moot, since each pedigree would in effect represent a different and truly independent data set. For this reason, we find it hard to imagine situations in which PED would be the subsetting scheme of choice from a scientific point of view.

However, PED does represent a particularly interesting scheme from a theoretical view, because it permits us, in principle, to reap the benefits of sequential updating without any prior knowledge or measures of good subsetting variables. We therefore consider the behavior of PED for all three pedigree sets. (Note that in this section, PED is based on an underlying homogeneity likelihood, that is, with α omitted; see Appendix B for a brief note on use of the heterogeneity likelihood in this context.)

Table 5 provides the mean PPL for PED under ‘linkage’ for the three pedigree sets and the two heterogeneity

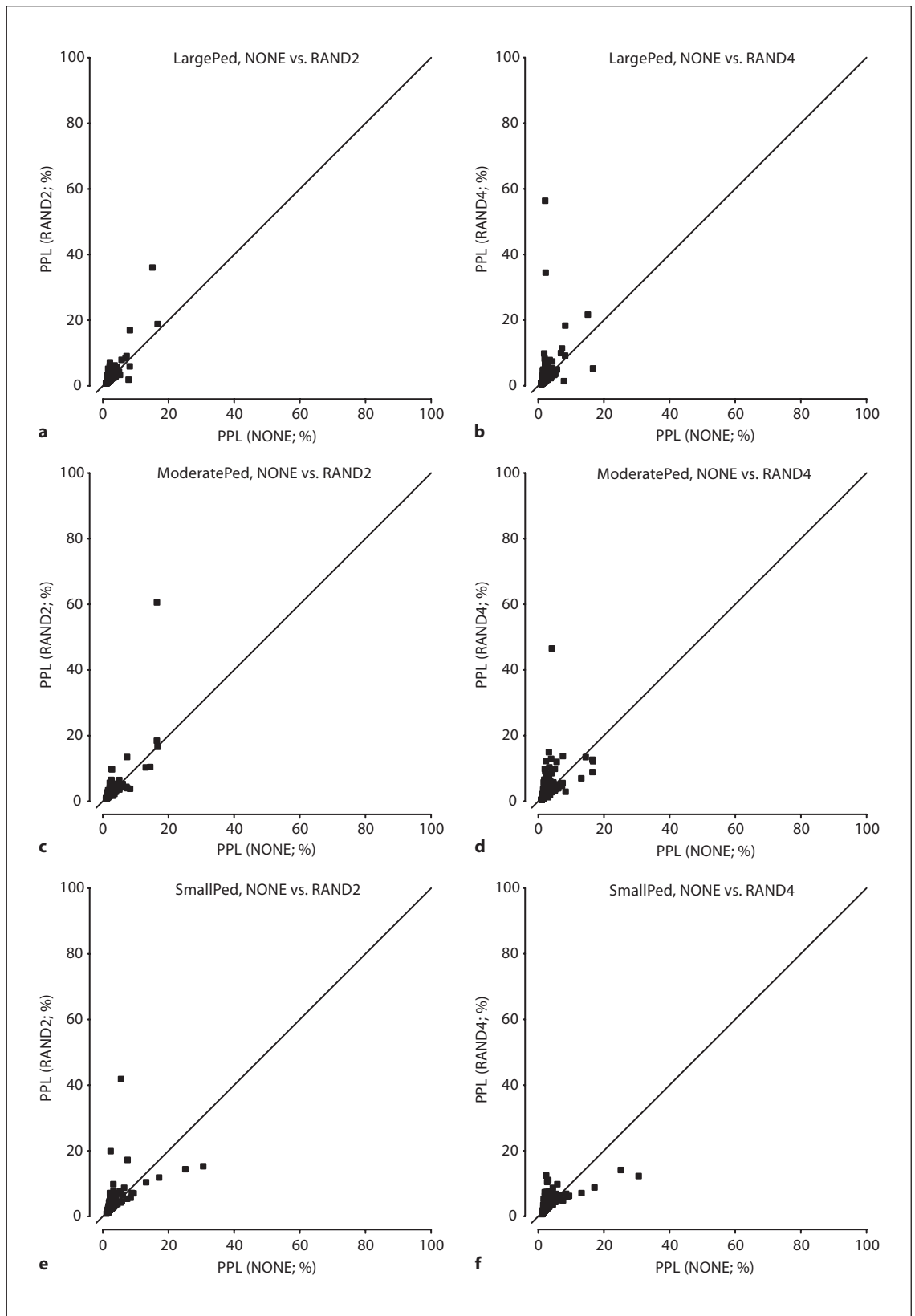


Fig. 3. a-f Scatter plot of the PPL (%) for NONE against RAND2 and RAND4 at an unlinked locus.

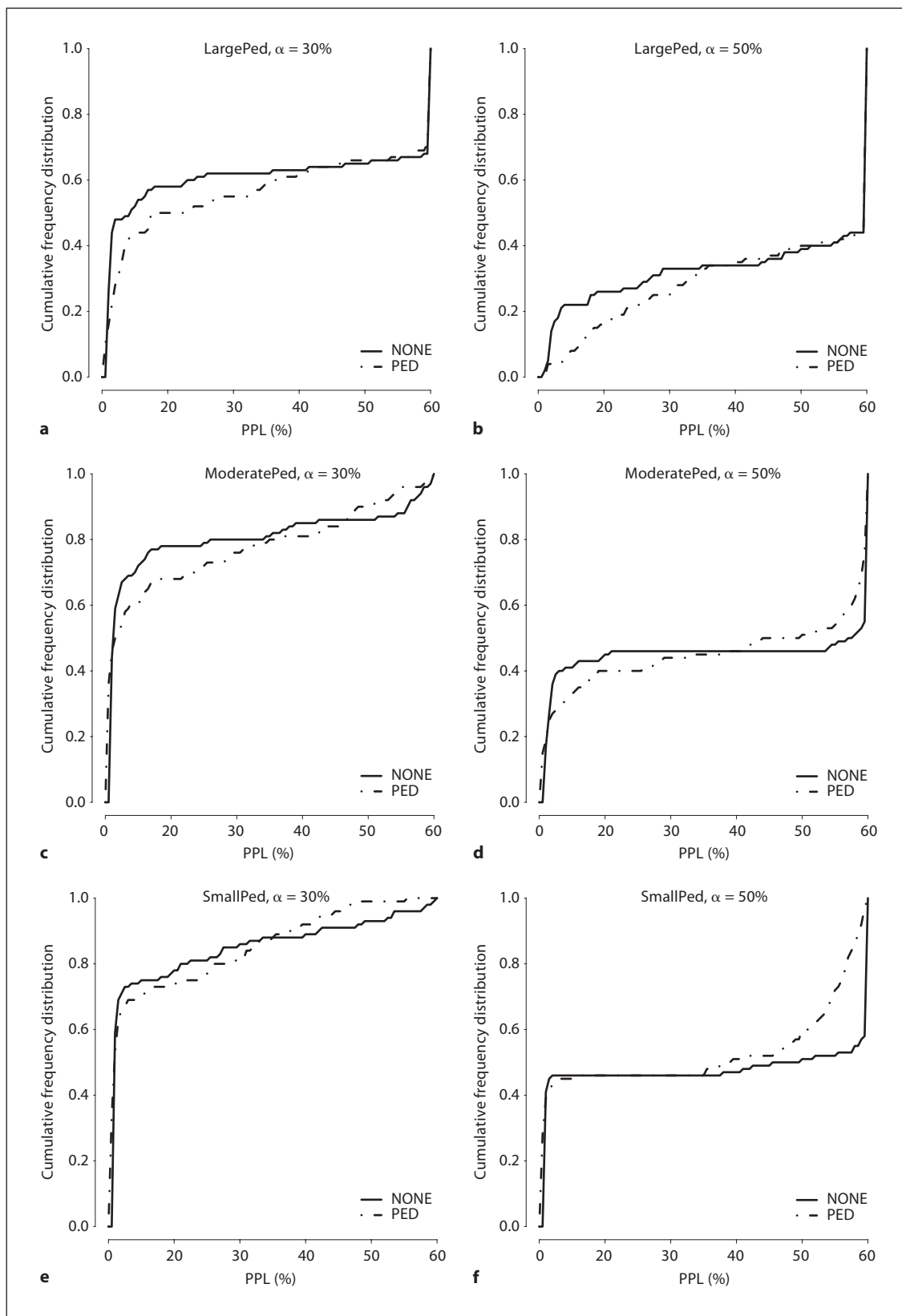


Fig. 4. a-f Cumulative frequency distribution of the PPL (%) for PED at a linked locus.

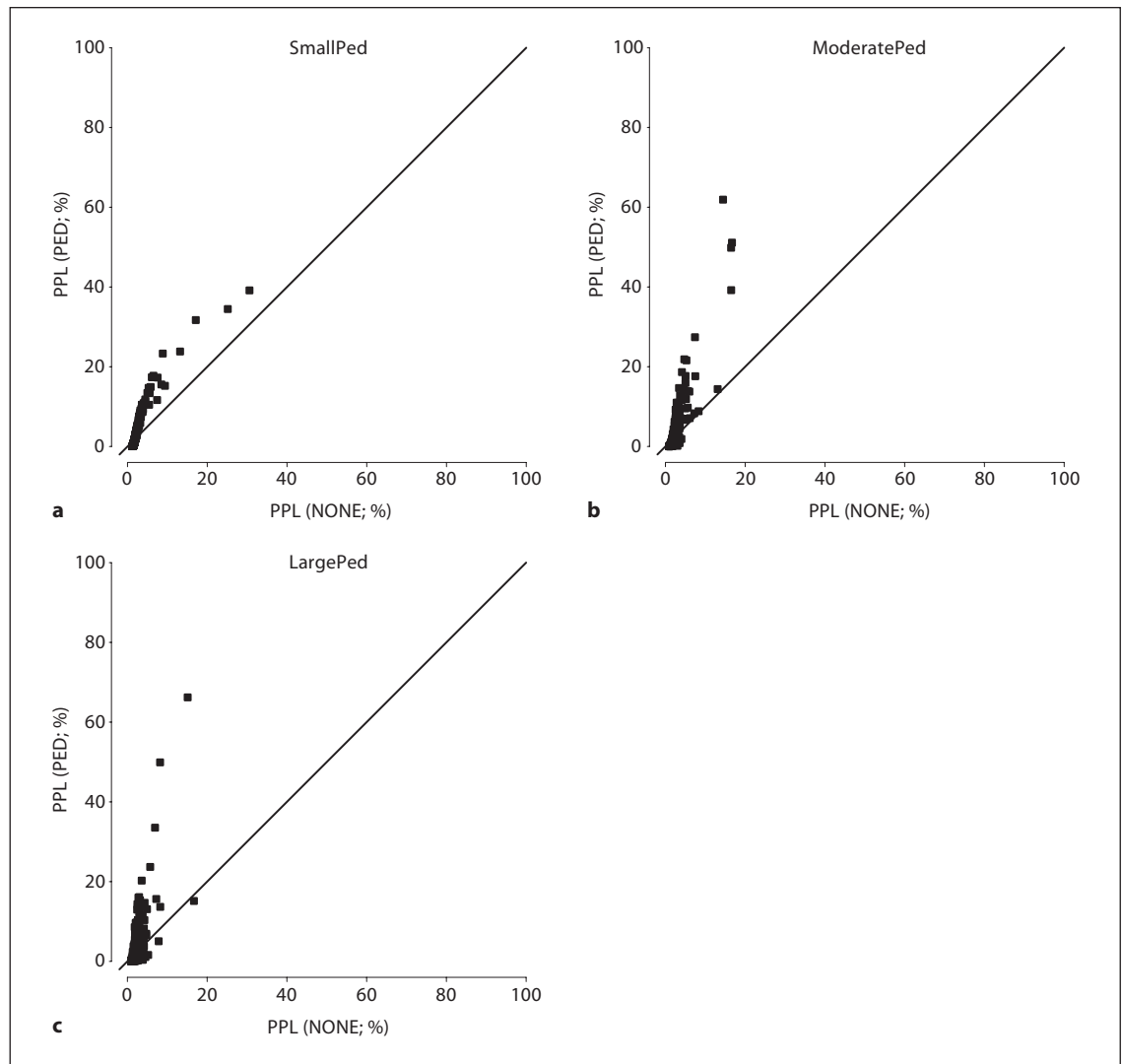


Fig. 5. a–c Scatter plot of the PPL (%) for NONE against PED at an unlinked locus.

levels. While the means for PED and NONE are similar in most cases, it is not surprising to see that PED appears to do better than NONE for more informative pedigrees and/or with greater heterogeneity. The pair-wise comparisons of the means highlight this point. With greater heterogeneity ($\alpha = 30\%$), the difference in means for SmallPed is not statistically significant, while for ModeratePed and LargePed the performance of PED is statistically significantly better than NONE. However, with more homogeneity ($\alpha = 50\%$), this trend is reversed. The performance of NONE compared to PED for SmallPed is statistically significant (p value: 2.17×10^{-11}) for the more homogeneous data; for ModeratePed, however, the

means are practically the same, while for LargePed although PED is still better in performance than NONE, the difference is not statistically significant. In all cases, differences in variances are statistically insignificant.

Figure 4 shows the corresponding cumulative frequency plots for PED compared with NONE. The behavior of PED is quite similar to NONE for LargePed and ModeratePed, while its performance in SmallPed is worse. Thus when there is linkage, considering the data as a single subset (NONE) appears to be comparable to or superior to considering the data to comprise one subset per pedigree (PED), especially with greater homogeneity, or less informative pedigrees. Of course, as stated earlier, in

Table 6. Mean (standard deviation) of the PPL (%) for PED at an unlinked locus

Pedigree set	NONE	PED
SmallPed	1.93 (1.50)	1.98 (3.11)
ModeratePed	1.76 (1.24)	1.45 (4.09)
LargePed	1.77 (0.97)	1.42 (3.66)

the case of very heterogeneous data with large, informative pedigrees, the individual pedigrees might presumably be independent data sets, in which case the question of sequential updating would be irrelevant. Hence, in the absence of any basis for subdividing the sample, it seems that from practical point of view, NONE is a better choice than PED.

Table 6 shows the mean PPL for PED under ‘no linkage’. As seen from the table, PED yields a mean PPL <2% for all pedigree sets, indicating evidence against linkage. However, there is statistically significantly more variability in the PPL for PED as compared to NONE, and 97.5–97.6% of the PPL values for PED are <10% (compare to 99.5–99.8% for NONE). Figure 5 shows the corresponding scatter plots, and makes clear the potential for PED to overstate the evidence for linkage at an unlinked locus.

It is of particular interest to note here that even sequentially updating by individual pedigree does not lead to appreciable ‘inflation’ of the PPL under ‘no linkage’. Recall that the underlying likelihood has 5 parameters, and these are being permitted to vary freely across individual pedigrees under the PED updating scheme. Yet the result is (essentially) no inflation of the mean PPL under ‘no linkage’. By contrast, it is well known that summing the maximum MOD (the LOD maximized over 5 parameters within each family) across families substantially affects the mean of the sampling distribution of the statistic.

Discussion

The PPL utilizes Bayesian sequential updating to accumulate evidence for or against linkage across multiple, potentially heterogeneous sets of data. In capitalizing upon the PPL’s use of sequential updating, it is of practical importance to know whether dividing data sets into subsets is helpful in the presence of linkage, whether it is harmful in the absence of linkage, and whether some subsetting schemes have better or worse performance than others. In this paper we have simulated various approaches to se-

quential updating across data subsets and characterized the results in terms of the corresponding sampling characteristics of the PPL; while varying pedigree size and levels of underlying heterogeneity in addition to the subsetting criteria. (Recall that there is another context – not considered in this paper – in which sequential updating is always advisable, and this is when following up in a new data set on a finding obtained in another set of families [2, 4, 11].)

Our simulations show that, in general, under conditions of ‘no linkage’, or of ‘linkage’ and relative homogeneity (say, 50% or so of families ‘linked’, on average, to the marker or location being analyzed), in the absence of any basis for subdividing the sample, the optimal procedure is to analyze all data as a single group. However, even in such cases, the impact of subdividing the data into what are in effect purely random subsets is not particularly deleterious, at least when we consider the impact on mean PPL. On the other hand, while the frequency of larger ‘outlier’ scores under ‘no linkage’ is essentially the same across subsetting schemes, the values of these outliers may be larger when updating over random subsets of the data. (Additional simulations, well beyond the scope of this paper in terms of computing resources, would be required to verify this.)

Under conditions of ‘linkage’ and appreciable heterogeneity (say, 30% or so of families ‘linked’ on average) there is a marked loss of information when data are pooled and analyzed as a single group, rather than divided into subsets on the basis of a genetically relevant variable and evaluated using sequential updating. This is true even when the classification variable is imperfect, misclassifying 20% of ‘linked’ and 20% of ‘unlinked’ pedigrees. By contrast, division of the data into random (genetically irrelevant) subsets yields lower mean PPL than simply analyzing the data as a single group. Also of note is the fact that dividing the data perfectly into one subset containing all the ‘linked’ pedigrees and another containing all the ‘unlinked’ pedigrees, and then sequentially updating across the two resulting data subsets, yields results virtually identical to what are obtained when analyzing only the ‘linked’ subset.

Overall, therefore, it appears that subsetting is useful when (i) there is appreciable heterogeneity (due to simple locus heterogeneity or other causes of differences across subsets), and/or (ii) there is a basis for subsetting that is likely to result, albeit perhaps imperfectly, in a higher percentage of ‘linked’ pedigrees in one of the resulting subsets than in the original set of data considered as a whole. The penalty for subsetting indiscriminately is a reduction in the PPL at linked loci, but overall the penalty under ‘no linkage’ appears to be negligible.

While we have attempted to systematically cover an important range of configurations with these simulations, clearly many additional cases could be considered as well (additional levels of α , different data set sizes, different numbers of subsets, different degrees of classification error, etc.). But specific implications of one approach or another will depend on the particulars of the underlying genetics for the particular phenotype(s) under investigation, along with a certain degree of ‘luck of the draw’ in assembling the particular data set at hand, and these are things that are never clear in practice at the time of analysis. It therefore strikes us as futile to attempt an exhaustive survey of all possible conditions of application. Nevertheless, the results shown above are consistent enough in pattern across what we view as a large enough range of generating conditions, that we are comfortable in extrapolating the primary results to general applications.

It appears that balancing the pros and cons of alternative subsetting schemes will remain something of an art form, and reasonable people could disagree on the optimal approach for any given data set. The good news is that the PPL appears to be quite robust to subsetting decisions in the presence of linkage, and relatively immune to the effects of subsetting in the absence of linkage. In particular, our results conclusively demonstrate that given a variable which can help delineate more homogeneous data subsets, sequential updating can be highly beneficial in the presence of appreciable heterogeneity at a linked locus, without inflation at an unlinked locus.

Acknowledgements

We thank Drs. Jeffery C. Murray, Mary L. Marazita, and Andrew C. Lidral for providing us with the pedigree structures and cleft lip/palate trait phenotypes used in this study. These pedigrees are part of a sample of cleft lip/palate families collected by an ongoing collaboration between The University of Iowa and the University of Pittsburgh, involving families from several different countries [family data collection supported by NIH grants R01-DE09886 [MLM], R01-DE012472 [MLM], R37-DE08559 [JCM], R01-DE016148 [MLM], P50-DE016215 [JCM, MLM], R21-DE016930 [MLM], R01-DE014667[ACL]]. We also thank Dr. Mark W. Logue for helpful discussions. This work was funded in part by NIH grant R03-DE-017167 to VJV.

Appendix A: Mathematical Details

The PPL is defined for two-point analysis as the probability that a trait gene is linked to the marker [1; for the multipoint analogue see 3]. For marker data M , and trait data T , the PPL is shown in Equation 1 [1, 28].

$$PPL = \frac{P(L) \int_{\theta \in [0, \frac{1}{2}]} \int_{\alpha} \int_{\mathbf{g}} P(M|T; \theta, \alpha, \mathbf{g}) f(\theta|L) f(\alpha) f(\mathbf{g}) d\theta d\alpha d\mathbf{g}}{[P(L) \int_{\theta \in [0, \frac{1}{2}]} \int_{\alpha} \int_{\mathbf{g}} P(M|T; \theta, \alpha, \mathbf{g}) f(\theta|L) f(\alpha) f(\mathbf{g}) d\theta d\alpha d\mathbf{g}] + (1 - P(L))P(M)} \quad (1)$$

In this equation, L indicates linkage and g , θ and α represent the trait parameters, recombination fraction and admixture parameter, respectively. As seen from the equation, the PPL includes a prior probability on linkage, $P(L)$. This [probability is fixed at 2% [1], indicating the small prior probability of linkage between two random loci [29]. The method also treats all the trait parameters as nuisance parameters by assigning independent, uniform, prior distributions to them and then integrating over their full space [28, 30].

Since

$$HLOD(\theta, \alpha) \triangleq \log_{10} \left[\frac{P(M, T; \theta, \alpha, \mathbf{g})}{P\left(M, T; \theta = \frac{1}{2}, \alpha, \mathbf{g}\right)} \right] \triangleq \log_{10} [P(M|T; \theta, \alpha, \mathbf{g})]$$

[20, 22], Equation 1 can also be expressed more simply in terms of the HLOD, as shown in Equation 2 [1].

$$PPL = \frac{P(L) \int_{\theta \in [0, \frac{1}{2}]} BR(\theta) f(\theta|L) d\theta}{P(L) \int_{\theta \in [0, \frac{1}{2}]} BR(\theta) f(\theta|L) d\theta + (1 - P(L))} \quad (2)$$

where

$$BR(\theta) = \int_{\alpha} \int_{\mathbf{g}} 10^{HLOD(\theta, \alpha, \mathbf{g})} f(\mathbf{g}) f(\alpha) d\mathbf{g} d\alpha$$

The PPL makes use of the Bayesian technique of sequential updating to accumulate linkage information across data sets/subsets [1, 11–12]. Given two or more sets of data, (D_1, D_2, \dots, D_n) , the procedure utilizes the marginal posterior density $f(\theta|D_i)$ of θ based on data set D_i as the prior density for θ in analyzing the next data set, D_{i+1} . Note that the PPL converges to 1 (if there is linkage) and to 0 (if there is no linkage) with increasing sample size, regardless of how the updating is carried out [27]; and the order in which the data sets are analyzed does not affect the final result [11].

Appendix B: Alternative Form of PED

In Results section 4 we considered the subsetting scheme PED based on an underlying homogeneity likelihood. It is also possible of course to retain the heterogeneity likelihood for PED, and this would have the advantage of maintaining the same likelihood form across all subsetting schemes. Here we briefly comment on

the effects of including or omitting the heterogeneity parameter α from the underlying likelihood. In what follows, we will use the notation $BR(\theta, \alpha)$ to explicitly denote the Bayes Ratio based on the heterogeneity likelihood (computed by integrating over all genetic parameters, including α); and $BR(\theta)$ to denote the Bayes Ratio based on the homogeneity likelihood. Note that this is a departure in notation from Appendix A, where the Bayes Ratio based on the heterogeneity likelihood is denoted by the simpler (and more usual) $BR(\theta)$.

When the PPL is computed for a single pedigree i , it can be shown that α is integrated out from the BR as a constant, as follows:

$$\begin{aligned} BR_i(\theta, \alpha) &= \int_{\alpha} \int_{\mathbf{g}} 10^{\log_{10}[\alpha 10^{LOD_i(\theta, \mathbf{g})} + (1-\alpha)]} f(\mathbf{g}) f(\alpha) d\mathbf{g} d\alpha \\ &= \int_{\alpha} \int_{\mathbf{g}} [\alpha 10^{LOD_i(\theta, \mathbf{g})} + (1-\alpha)] f(\mathbf{g}) f(\alpha) d\mathbf{g} d\alpha \\ &= \left[\int_{\mathbf{g}} \frac{10^{LOD_i(\theta, \mathbf{g})}}{2} f(\mathbf{g}) d\mathbf{g} \right] + \frac{1}{2} \end{aligned}$$

This gives the relationship

$$BR_i(\theta, \alpha) = \frac{BR_i(\theta)}{2} + \frac{1}{2},$$

which means that the impact of including α in the likelihood for a single pedigree is to inflate the BR if $BR_i(\theta) < 1$ (evidence against linkage) and deflate the ratio if $BR_i(\theta) > 1$ (evidence in favor of linkage). Note that this effect depends only upon the size of the Bayes Ratio, or equivalently, the magnitude of the PPL itself, and not upon any other features of the data or the underlying genetics (including whether or not there is linkage). Thus inclusion of α merely attenuates the PPL across the BR range.

The immediate implication is that the heterogeneity-based version of PED would under-perform the homogeneity version across the board, and therefore also NONE (see Results section 4). It therefore seems unnecessary to show additional simulation results for this form of PED.

On the other hand, we note that the selection of a particular form of likelihood based on its sampling behavior begs the question of whether perhaps the heterogeneity likelihood might be doing a better job of representing ‘the truth’, despite its less desirable sampling characteristics. This philosophical question is beyond the scope of the current paper, but see [2] for discussion.

References

- 1 Vieland VJ: Bayesian linkage analysis, or: How I learned to stop worrying and love the posterior probability of linkage. *Am J Hum Genet* 1998;63:947–954.
- 2 Vieland VJ: Thermometers: something for statistical geneticists to think about. *Hum Hered* 2006;61:144–156.
- 3 Logue MW, Vieland VJ: A new method for computing the multipoint posterior probability of linkage. *Hum Hered* 2004;57:90–99.
- 4 Bartlett CW, Vieland VJ: Accumulating quantitative trait linkage evidence across multiple datasets using the posterior probability of linkage. *Genet Epidemiol* 2007;31:91–102.
- 5 Bartlett CW, Vieland VJ: Two novel quantitative trait linkage analysis statistics based on the posterior probability of linkage: Application to the COGA families. *BMC Genet* 2005;6(suppl 1):S121.
- 6 Huang Y, Bartlett CW, Segre AM, O’Connell JR, Mangin L, Vieland VJ: Exploiting gene \times gene interaction in linkage analysis. *BMC Proceedings* 2007;1(suppl 1):S64.
- 7 Yang X, Huang J, Logue MW, Vieland VJ: The posterior probability of linkage allowing for linkage disequilibrium and a new estimate of disequilibrium between a trait and a marker. *Hum Hered* 2005;59:210–219.
- 8 Logue MW, Vieland VJ: The incorporation of prior genomic information does not necessarily improve the performance of Bayesian linkage methods: An example involving sex-specific recombination and the two-point PPL. *Hum Hered* 2005;60:196–205.
- 9 Govil M: Extensions of the posterior probability of linkage: Distributed computation, incorporation of genetic map information, and application to cleft lip and/or palate, Doctoral Thesis 2005; The University of Iowa.
- 10 Huang J, Vieland VJ: Comparison of ‘model-free’ and ‘model-based’ linkage statistics in the presence of locus heterogeneity: Single data set and multiple data set applications. *Hum Hered* 2001;51:217–225.
- 11 Vieland VJ, Wang K, Huang J: Power to detect linkage based on multiple sets of data in the presence of locus heterogeneity: Comparative evaluation of model-based linkage methods for affected sib pair data. *Hum Hered* 2001;51:199–208.
- 12 Wang K, Vieland VJ, Huang J: A Bayesian approach to replication of linkage findings. *Genet Epidemiol* 1999;17(suppl 1):S749–S754.
- 13 Marazita ML, Murray JC, Lidral AC, Arcos-Burgos M, Cooper ME, Goldstein T, Maher BS, Daack-Hirsch S, Schultz R, Mansilla MA, Field LL, Liu YE, Prescott N, Malcolm S, Winter R, Ray A, Moreno L, Valencia C, Neiswanger K, Wyszynski DF, Bailey-Wilson JE, Albacha-Hejazi H, Beaty TH, McIntosh I, Hetmanski JB, Tuncbilek G, Edwards M, Harkin L, Scott R, Roddick LG: Meta-analysis of 13 genome scans reveals multiple cleft lip/palate genes with novel loci on 9q21 and 2q32–35. *Am J Hum Genet* 2004;75:161–173.
- 14 Mitchell LE, Risch N: Mode of inheritance of non-syndromic cleft lip with or without cleft palate: A reanalysis. *Am J Hum Genet* 1992;51:323–332.
- 15 Wyszynski DF, Beaty TH: Review of the role of potential teratogens in the origin of human nonsyndromic oral clefts. *Teratology* 1996;53:309–317.
- 16 Little J, Cardy A, Munger RG: Tobacco smoking and oral clefts: A meta-analysis. *Bull World Health Organ* 2004;82:213–218.
- 17 Ott J: Computer-simulation methods in human linkage analysis. *Proc Natl Acad Sci USA* 1989;86:4175–4178.
- 18 Weeks DE, Ott J, Lathrop GM: SLINK: A general simulation program for linkage analysis. *Am J Hum Genet* 1990;47(suppl 3):A204.
- 19 Smith CAB: Testing for heterogeneity of recombination fraction values in human genetics. *Ann Hum Genet* 1963;27:175–192.
- 20 Clerget-Darpoux F, Bonaïti-Pellié C, Hochez J: Effects of misspecifying genetic parameters in LOD Score Analysis. *Biometrics* 1986;42:393–399.
- 21 Greenberg DA: Inferring mode of inheritance by comparison of LOD scores. *Am J Med Genet* 1989;34:480–486.
- 22 Elston RC: Man bites dog? The validity of maximizing LOD scores to determine mode of inheritance. *Am J Med Genet* 1989;34:487–488.
- 23 Ott J: *Analysis of Human Genetic Linkage*, ed 2, revised. Baltimore, The Johns Hopkins University Press, 1991.

- 24 Govil M, Segre AM, Vieland VJ: MLIP: A multiprocessor linkage analysis system. Proceedings of IEEE 2nd International Multi-Symposiums on Computer and Computational Sciences (IMSCCS 2007); pp 17–24, August 13–15, 2007, Iowa City, IA, USA. Available: <http://ieeexplore.ieee.org/search/srchabstract.jsp?arnumber=4392575&isnumber=4392566&punumber=4392565&k2dockey=4392575@ieeecnfs@query=%28govil+%3Cin%3E+metadata%29+%3Cand%3E+%284392565+%3Cin%3E+punumber%29&pos=0&access=no>
- 25 Huang Y, Segre AM, O'Connell JR, Wang H, Vieland VJ: KELVIN: A 2nd generation distributed multiprocessor linkage and linkage disequilibrium analysis program [abstract 1556]. Presented at the annual meeting of The American Society of Human Genetics; October 9–13, 2006, New Orleans, LA, USA. Available from: <http://www.ashg.org/genetics/ashg06s/index.shtml>.
- 26 Wang H, Segre AM, Huang Y, O'Connell JR, Vieland VJ: Fast computation of human genetic linkage. Proceedings of IEEE 7th Symposium on Bioinformatics and Bioengineering (BIBE 2007); pp 857–863, October 14–17, 2007, Boston, MA, USA. Available: <http://ieeexplore.ieee.org/search/srchabstract.jsp?arnumber=4375660&isnumber=4375521&punumber=4375520&k2dockey=4375660@ieeecnfs&query=%28vianland+%3Cin%3E+metadata%29+%3Cand%3E+%284375520+%3Cin%3E+punumber%29&pos=0&access=no>
- 27 Wang K, Huang J, Vieland VJ: The consistency of the posterior probability of linkage. *Am J Hum Genet* 2000;64:533–553.
- 28 Logue MW, Vieland VJ, Goedken RJ, Crowe RR: Bayesian analysis of a previously published genome screen for panic disorder reveals new and compelling evidence for linkage to chromosome 7. *Am J Med Genet B Neuropsychiatr Genet* 2003;121B:95–99.
- 29 Elston RC, Lange K: The prior probability of autosomal linkage. *Ann Hum Genet* 1975;38:341–350.
- 30 Bartlett CW, Flax JF, Logue MW, Vieland VJ, Bassett AS, Tallal P, Brzustowicz LM: A major susceptibility locus for specific language impairment is located on 13q21. *Am J Hum Genet* 2002;71:45–55.

Erratum

In the article of Guan et al. 'Meta-Analysis of 23 Type 2 Diabetes Linkage Studies from the International Type 2 Diabetes Linkage Analysis Consortium' (*Hum Hered* 2008; 66:35–49), two persons were left out of the listing of the International Type 2 Diabetes Linkage Analysis Consortium. The list for the University of Chicago/University of Texas Health Science Center at Houston group should include Craig L. Hanis and D. Michael Hallman, both of the Human Genetics Center, University of Texas Health Science Center at Houston, Houston, TX, USA.