Published in final edited form as:

Genet Epidemiol. 2009 November ; 33(7): 628–636. doi:10.1002/gepi.20414.

The sumLINK statistic for genetic linkage analysis in the presence of heterogeneity

GB Christensen, S Knight, and NJ Camp

University of Utah School of Medicine Department of Biomedical Informatics

Abstract

We present the “sumLINK” statistic—the sum of multipoint LOD scores for the subset of pedigrees with nominally significant linkage evidence at a given locus—as an alternative to common methods to identify susceptibility loci in the presence of heterogeneity. We also suggest the “sumLOD” statistic (the sum of positive multipoint LOD scores) as a companion to the sumLINK. SumLINK analysis identifies genetic regions of extreme consistency across pedigrees without regard to negative evidence from unlinked or uninformative pedigrees. Significance is determined by an innovative permutation procedure based on genome shuffling that randomizes linkage information across pedigrees. This procedure for generating the empirical null distribution may be useful for other linkage-based statistics as well. Using 500 genome-wide analyses of simulated null data, we show that the genome shuffling procedure results in the correct type 1 error rates for both the sumLINK and sumLOD. The power of the statistics was tested using 100 sets of simulated genome-wide data from the alternative hypothesis from GAW13. Finally, we illustrate the statistics in an analysis of 190 aggressive prostate cancer pedigrees from the International Consortium for Prostate Cancer Genetics, where we identified a new susceptibility locus. We propose that the sumLINK and sumLOD are ideal for collaborative projects and meta-analyses, as they do not require any sharing of identifiable data between contributing institutions. Further, loci identified with the sumLINK have good potential for gene localization via statistical recombinant mapping, as, by definition, several linked pedigrees contribute to each peak.

Keywords

Methodology; simulation; multipoint linkage; heterogeneity; empirical testing

INTRODUCTION

Genetic linkage analysis can be an effective tool for identifying disease susceptibility loci. However, locus heterogeneity can counter this effectiveness and is often acknowledged as the single largest issue that hinders the linkage analysis approach. Complex traits may be controlled by numerous genes and therefore statistics that attempt to model or recognize locus heterogeneity are required. The two common methods to address heterogeneity are the heterogeneity LOD statistic (HLOD), which statistically models the heterogeneity with an additional parameter, and phenotypic subset analysis. However, the former may fail to distinguish linked from unlinked pedigrees sufficiently to allow for substantial power increase and suffers from the lack of a precise distribution for assessing significance, and the latter requires a-priori determination of subsets. Beyond heterogeneity, localization also presents a challenge in linkage analysis. Often regions identified by linkage are large

(perhaps 30-50 cM) and the boundaries ill-defined, both of which hinder follow-up studies. A method that can both address locus heterogeneity and produce regions that are useful for localization would be an important addition to the tools already available.

The genetic research community has ascertained a great number of family-based resources for linkage analysis across numerous and varied complex traits. These data repositories represent a tremendous investment of time and resources, and likely contain a wealth of information—much of which has yet to be extracted. In the era of consortia efforts and with great numbers of pedigrees available for specific diseases through collaborative efforts, new approaches and opportunities arise, especially to identify genes that may explain only a very small portion of disease that could not be identified in single studies. Here, we introduce a new approach to locus heterogeneity that focuses on individually powerful pedigrees—something that becomes possible in multi-center collaborative settings and other studies with large numbers of pedigrees. A standard HLOD analysis attempts to statistically separate linked and unlinked pedigrees through an additional parameter in the LOD calculation, α , the proportion of linked pedigrees; however many pedigrees may be uninformative (pedigree-specific contributions surrounding 0) at a locus, and these pedigrees add statistical noise that reduces power. Our new approach uses a pre-defined LOD threshold to simply remove those families that are below the threshold and could be considered “noise”. As such, it could be thought of as a “brute-force” approach to heterogeneity which attempts to gain power by removing noise from the statistical analysis. This method directly addresses the fact that only a small portion of the pedigrees in a data resource may be linked to any true causative locus, and in the process identifies the informative set of families that are most useful for defining and fine mapping the locus. Several recent studies have used statistical recombinant mapping to delimit the boundaries of linkage regions.[Camp, et al. 2007; Camp, et al. 2006; Johanneson, et al. 2008] Recombinant mapping requires that several pedigrees be linked to a region of interest, but it is unclear how many pedigrees should be linked to a locus for it to be considered a reasonable candidate for successful mapping. The sumLINK statistic can address this issue by assigning valid significance probabilities.

Our method focuses on individually powerful pedigrees that are nominally “linked” to a position in the genome and assesses whether the amount of concordance observed across the linked pedigrees at any point in the genome is more than would be expected by chance. Statistical excess of concordance is evidence for an underlying susceptibility locus. An advantage is that by the nature of the procedure, the regions of interest identified by the sumLINK statistic should have multiple pedigrees that can be used to delimit the region using statistical recombinant mapping. Further, in contrast to many other situations, the existence of different genetic marker sets (which often will occur in consortia) is not problematic and may, in fact, lead to some serendipitous pseudo-fine mapping. This method offers additional opportunities to identify disease susceptibility loci and the underlying genes using linkage-based data.

METHODS

sumLINK and sumLOD

Our approach is to identify regions of the genome that display a significant excess of concordance across ‘linked’ pedigrees. The level of concordance is quantified by the sum of the pedigree-specific multipoint LOD scores in the identified linked pedigrees. We consider any pedigree that meets or exceeds a pedigree-specific LOD threshold of 0.588 ($p = 0.05$, “nominal” significant evidence) at a specific genomic position to be “linked” at that position. We have called this linkage-based statistic the “sumLINK,” because it is simply the sum of multipoint LOD scores for linked pedigrees at a given point in the genome. Clearly,

the distribution of the sumLINK statistic varies according to the number and structure of pedigrees in the initial resource and the parameters of the linkage model used to calculate the LOD scores. It is therefore difficult to determine the null distribution of the statistic theoretically; however, empirical methods can be employed to generate the null distribution for any data resource. The creation of the null distribution from which to test significance is outlined below.

To perform a sumLINK analysis it is necessary that linkage results are available for each pedigree at regular intervals across the genome (Figure 1, A). This is possible in many standard linkage software packages that calculate multipoint LOD scores, including Merlin [Abecasis, et al. 2002] and Genehunter-Plus [Kong and Cox 1997]. The sumLINK statistic is then calculated at each position by summing the LOD scores for only those pedigrees that meet or exceed the threshold of 0.588 at each position in the genome. A simplistic example is shown in Figure 2. The null distribution of the sumLINK statistic must represent the chance consistency expected across linked pedigrees; matched for pedigree structure, information content and linkage potential. We achieve this null scenario by using a genome shuffling technique. The shuffling procedure consists of a chromosome randomization step and a genome rotation step. The randomization step begins by randomizing the sequential order of chromosomes for each pedigree. Chromosomes are concatenated end-to-end in this random order to create a 'new' genome (Figure 1, B). The beginning and end of this new genome is connected to form a 'loop'. In the rotation step, a random position in the loop is chosen and the loop rotated such that this position becomes the new starting position and the loop is broken there. This is done for each pedigree separately, and because multipoint LODs were calculated at evenly spaced positions, these new shuffled genomes can again be aligned across pedigrees (Figure 1, C). A null sumLINK statistic can then be calculated at each position across the shuffled genomes. The procedure maintains the continuity and autocorrelation of marker data within chromosomes, but randomizes consistency across pedigrees. The shuffling procedure is repeated a large number of times to determine the null distribution of the statistic for the given data.

Genome-wide significance is determined by the expected frequency of peaks of at least a certain magnitude occurring in the null sumLINK genome scans [Chen and Storey 2006]. All peaks in each null genome are considered. In accordance with guidelines set by Lander and Kruglyak [Lander and Kruglyak 1995] for significance, we consider a peak height that is expected to occur with a frequency no greater than 0.05 per genome as genome-wide significant evidence for linkage, and a peak that occurs with an expected frequency of less than 1.00 per genome as genome-wide suggestive. It is important to note that a false positive rate (FPR) is not a p-value, it is a rate per genome and represents the expected frequency of peaks of at least the specified magnitude under null conditions. For example, $FPR=0.6$ indicates that a similar peak would be expected to occur 0.6 times per genome, which is sufficient evidence for suggestive linkage.

The advantage of the sumLINK is that regions are identified using individually powerful pedigrees, which is more intuitively appealing and perhaps convincing. Further, these independently powerful linked pedigrees can be used for localization using statistical recombinant mapping. In brief, statistical recombinant mapping uses pedigree-specific linkage evidence to estimate the positions of recombinant events on the linked segregating haplotypes which can then be used to delimit the shared genomic region. Aligning these regions across all linked pedigrees localizes the region for further study. Another advantage of sumLINK is that it requires a minimal data set. It is not necessary to know pedigree structures or genotypes, which are required to obtain null statistics by permuting disease status [Chen and Storey 2006], nor is it necessary that all pedigrees be genotyped with the same marker set, so long as the various marker sets are fitted to a common genetic map

before calculating LOD scores. This property makes sumLINK ideal for multi-institutional collaborative research projects. A disadvantage is that the sumLINK will ignore some small pedigrees (e.g. sib-pairs) due to their relative lack of information content. It therefore may be attractive to additionally consider the sumLOD (sum of all positive pedigree LOD scores) as a companion to the sumLINK. The relaxed inclusion threshold of the sumLOD allows the potential for any minimally informative pedigree to contribute to the result. The sumLOD statistic is similar to the previously proposed C statistic [MacLean, et al. 1992], but utilizes multipoint, rather than two-point, inheritance information. The sumLOD has been used previously as a summary measure [Camp, et al. 2003; Horne, et al. 2003; Orr, et al. 2007], but has not been adopted as a test statistic due to the lack of a theoretical distribution. However, our genome shuffling procedure can be used to assess empirical significance of any statistic that is derived as a post-processing step from pedigree-specific LOD values, including the sumLOD.

False Discovery Rate

Often for complex traits, no single significant findings are identified when using conservative family-wise multiple testing corrections and thresholds. It may therefore be useful to identify whether there exists a group of most significant findings that together indicates deviation from the null. The false discovery rate (FDR) evaluates this using the q-value. For example, if a q-value of 0.05 is assigned to the top 40 most significant findings, this indicates that 2 (0.05×40) are likely false positives, and that 38 are true-positives. In this way a group can be identified within which true positive findings likely reside. Using our genome-shuffling method it is possible to estimate empirical FDRs for the observed findings. In particular, this allows us to assess significance accounting for the multiple testing inherent in the multiple models and statistics.

Simulation tests

Simulations under the null hypothesis of no linkage—We tested the sumLINK and sumLOD procedures in data simulated under null conditions in order to assess the validity of our genome shuffling procedure to generate the correct false positive rates. We created 400 two, three and four generation pedigrees typical of the families commonly used in linkage analysis. Each pedigree had a minimum of three affected subjects. Genome-wide genotypes were simulated using random gene-drops based on the genetic map and characteristics of 2,936 autosomal single nucleotide polymorphisms selected from the Illumina 6K SNP array to be free from linkage disequilibrium. This was repeated 250 times. Additional pedigree characteristics are summarized in Table I. For each of the 250 replicates, multipoint parametric linkage statistics were calculated at 1cM intervals for both dominant and recessive inheritance models using Merlin [Abecasis, et al. 2002], and results for each pedigree were extracted. Genome-wide sumLINK and sumLOD statistics were then computed for each replicate, and the empirical significance was assessed with 200 iterations of the genome shuffling procedure. Across the 250 replicates, the median number of pedigrees that contributed to the dominant sumLINK analysis was 153, and the median number that contributed to the recessive sumLINK analysis was 276.

Simulations under the alternative hypothesis of linkage—We illustrate the power of the sumLINK and sumLOD statistics, in comparison to the more standard HLOD, by applying these to data based on the well-documented, simulated genome-wide data from Genetic Analysis Workshop 13 (GAW13) [Almasy, et al. 2003; Daw, et al. 2003]. The GAW13 data were designed to represent random pedigrees (not ascertained for specific disease) and contained several simulated ‘heart disease’ traits. Fifty trait-related genes were simulated, most with common underlying susceptibility alleles and low effect sizes. There were 100 replicates of 330 simulated pedigree structures available, resulting in a total of

33,000 independent pedigrees. Genotypes were simulated for 399 microsatellite markers across the 22 autosomes. We chose to analyze an obesity trait as defined by [Klein, et al. 2003] and sampled from the full set of pedigrees to better represent a linkage resource ascertained for disease. From the full set of 33,000 pedigrees we extracted 5232 independent and minimally informative unilineal pedigrees (those with at least two genotyped subjects classified as obese). For each of the 5232 pedigrees, multipoint parametric linkage statistics were calculated at 1cM intervals for a simple dominant inheritance model using Merlin [Abecasis, et al. 2002]. The pedigrees were divided into two groups based on whether they would be included in a sumLINK analysis (that is, that a minimum LOD score of 0.588 was observed at least once across the genome). Of the 5232, 1056 pedigrees were useful for sumLINK analysis; the remaining 4176 pedigrees were not suitable for sumLINK analysis but remained useful for sumLOD analysis. Pedigree characteristics for each group are summarized in Table I. By sampling from the two groups of pedigrees, we created 100 replicates each containing 200 pedigrees (100 useful for sumLINK and 100 that were not); all sampling was performed with replacement. We then calculated genome-wide sumLOD and sumLINK statistics for each of the 100 replicates, with empirical significance determined by 200 repetitions of the genome shuffling procedure. HLODs were calculated with Merlin. Thresholds of 1.9 and 3.3 were used to determine suggestive and significant HLOD results.

Aggressive Prostate Cancer Case Study

We performed a sumLINK and sumLOD analysis on 190 pedigrees provided by the International Consortium for Prostate Cancer Genetics (ICPCG) [Schaid and Chang 2004] with clinically aggressive prostate cancer. A conventional linkage study of this resource and description of the data was published previously.[Schaid and ICPCG 2006] Dominant and recessive multipoint LOD scores were computed for each pedigree at 1-cM increments throughout the 22 autosomes by Genehunter-Plus using the models (dominant and recessive) as described by Schaid et al. The sumLINK and sumLOD statistics were then calculated at each of the cM positions for both models, and empirical significance of the observed peaks was determined by 1000 repetitions of the genome shuffling procedure. Of the total 190 ICPCG pedigrees, 125 pedigrees achieved linkage evidence of at least 0.588 at some point in the genome with the dominant model and 127 for the recessive model. Hence only these numbers of pedigrees contribute to the sumLINK analysis. All 190 pedigrees reached a LOD score greater than zero at least once in each inheritance model, allowing them all to contribute to the sumLOD analysis under both models.

RESULTS

Simulation tests

Simulations under the null hypothesis of no linkage—Table II illustrates the false positive rates observed under the dominant and recessive models for the sumLINK and sumLOD. All of these results fall within the 95% confidence interval based on 250 replicates and for a Poisson process with rates of 0.05 or 1.0, and illustrate that the genome shuffle procedure to determine significance is valid.

Simulations under the alternative hypothesis of linkage—Results of the power testing are summarized in Table III. All genes that were identified with suggestive evidence at least five times by any one of the three statistics (HLOD, sumLINK or sumLOD) are summarized. The simulated data included only one gene, *Gb11*, which affected baseline weight with a reasonably large effect size. This gene was identified with excellent power with all three statistics (all 99% power). Of the remaining genes, all had very common susceptibility alleles (minor allele frequencies 0.15) and low effect sizes. Subsequently

none were identified particularly well. However, it is interesting to note that of these lower effect size genes that were identified at least 5 times out of the 100 replicates, that the sumLOD and/or sumLINK were always superior, and exhibited significantly more power than the HLOD at eight of the 10 loci. There were no genes that were significantly better identified with the HLOD.

Aggressive Prostate Cancer Case Study

In our real data aggressive prostate cancer case study example, the sumLINK and sumLOD analyses identified significant linkage evidence at two loci (chromosomes 20q and 11q) and suggestive evidence at a third locus (chromosome 2), as shown in Table IV. The peak on chromosome 20 was significant under the dominant inheritance model for both the sumLINK (sumLINK = 13.848, number of linked pedigrees = 17, expected false positive rate (FPR) = 0.005) and the sumLOD (sumLOD = 30.311, number of positive pedigrees = 83, FPR=0.028). The peak on chromosome 11 was significant in the recessive sumLOD analysis (FPR=0.007), with suggestive evidence in other analyses. The sumLINK analysis also identified suggestive linkage evidence on chromosome 2 under both dominant and recessive models (FPR = 0.628 and 0.897, respectively). Figure 3 shows the genome-wide sumLINK results for the dominant model and sumLOD results for the recessive model.

In an attempt to consider the multiple testing inherent from performing both the sumLINK and the sumLOD, both for dominant and recessive models, we considered the false discovery rates. Each centimorgan position in the genome search data (N=3502) was considered as an individual observation and p-values were calculated for every position based on the respective empirical distribution for each of the analyses. When the results of all four analyses are pooled, the top 54 ranked cM positions collectively attained an FDR of 0.1. An FDR of 0.1 indicates that the expected ratio of false:true positives is 1:9. That is, that one tenth of these 54 (or, 5-6 positions) are likely from the null (false positives), but the remaining are likely true positives. FDR will not differentiate which are which; however, in this case example, all 54 positions fall under one of the significant linkage peaks (19, 16 and 19 positions on chromosomes 20 (sumLINK), 20 (sumLOD) and 11 (sumLOD), respectively). Hence, even if all 5-6 false positive findings were from one region, it is still expected to have a true positive in each. In conclusion, the FDR suggests that the linkage peaks on chromosomes 11 and 20 are likely true positive findings after correction for multiple testing.

We applied statistical recombinant mapping to all three regions with at least suggestive genome-wide evidence to delimit the regions of interest. The genotypes used in this multi-center collaborative analysis were derived from several diverse sets of microsatellite markers, generally with an average spacing of 10 cM. On average, therefore, most pedigrees have a genotyped marker within 5 cM of any given cM position on the genetic map. Pedigrees were therefore included in a localization analysis if they achieved LOD ≥ 0.588 within 5 cM of the observed peak. Figure 2 illustrates the by pedigree LOD tracings used in the recombinant mapping for the three regions of interest. Recombinant events are estimated to be at the outermost point of a sharp decline in LOD score, as these positions indicate statistical evidence for a loss of genetic sharing. This point is a conservative estimate for the outer limit of the region where a susceptibility variant may be found. A region bounded by two recombinant events on each side represents an approximate 95% confidence interval for the consensus region.[Camp, et al. 2006] As seen in Figure 4, the linkage peaks on Chromosomes 20, 11, and 2 can each be conservatively localized to regions of 21, 21 and 19 cM, respectively.

DISCUSSION

The sumLINK statistic is a new method aimed at addressing both heterogeneity and localization. The procedure is designed to identify the genomic regions for which an excessive number of powerful pedigrees are concordant. It is an ideal approach for multi-center collaborations or large single-site studies where a large number of pedigrees are available. A distinct advantage of this method is that it does not require collaborating centers to share raw data such as pedigree structures or genotypes; and does not require that each center use the same marker set. Provided a common genetic map is used for analysis, each center can perform their own analyses, calculating multipoint LOD scores at the same equally-spaced increment across the genome. It is only necessary to share these meta data (a multipoint genome scan for each pedigree), which enhances data privacy and security.

An important advantage of the sumLINK is the ability to identify loci that have good potential for gene localization, as several linked pedigrees exist beneath each peak identified. An unexpected benefit of compiling data across centers that used different marker maps is that the resolution of the localization can be higher than any of the individual genetic maps due to the overlaying of data. In our example, even with a low density 10 cM marker map, we were able to localize each region to approximately 20 cM, and these localized regions would be greatly refined with fine-mapping. This method of using the limits of sharing observed within extended pedigrees is intuitively appealing for localization, but may also have theoretical advantages over other common methods. Often so-called “1-LOD” support intervals are reported for linkage peaks generated from a HLOD analysis; however, support intervals should strictly be applied to parameter estimates (the recombination fraction parameter, θ , in the case of linkage statistics) and are relevant in the context of two-point maxLOD statistics that are directly analogous to likelihood ratio tests. The standard practice of a 1-LOD support interval using the value of the statistic itself (usually HLOD) rather than a parameter is not statistically well-grounded, although since θ is a distance parameter has intuitive appeal. In particular, in a HLOD analysis it is not clear whether the statistical noise generated by “unlinked” pedigrees may mask or shift positive linkage evidence. Hence, these “1-LOD” intervals can only be considered as a rough guide.

The shuffling method we have implemented to determine the null distribution is a particularly innovative element of the sumLINK procedure, and may be especially useful to the broader research community. We used the procedure to assess the significance of two genome-wide linkage statistics (sumLOD and sumLINK), but it may have broader applications for testing the significance of other statistics with unknown distributions. It is a simple, elegant, and quick way to create null data for assessing significance. It accounts for variations in pedigree structure as well as the autocorrelation of consecutive loci inherent in genetic linkage data. We developed a post-processing script written in R [RDCT 2006] that calculates the sumLINK and sumLOD statistics, performs the genome-shuffling, generates the empirical null distribution, and tests the significance of observed linkage peaks. Computational time is dependent upon the number of pedigrees and the length of the genomic region being analyzed. The ICPCG data, comprised of 190 pedigrees and 3502 data points from 22 chromosomes, required 21.3 seconds per iteration with a 3.0 GHz Intel Xeon Duo Core 64-bit CPU running R v2.4.1 on Red Hat Enterprise Linux v5. One iteration consists of shuffling all pedigrees, calculating the null sumLINK, sumLOD, and number of pedigrees contributing to each statistic at all data points, and writing out a text file containing these values. Significance is computed in a later step after all shuffling iterations are complete. Our simulated null data (400 pedigrees, 3550 cM, 22 chromosomes) required 60.8 seconds per iteration, and the simulated data sets from GAW13 (200 pedigrees, 3604 cM, 22 chromosomes) required 22.9 seconds per iteration.

Analysis of simulated null data illustrated that the type-I error rate for the sumLINK and sumLOD statistics were all within acceptable boundaries, indicating that the genome shuffling procedure is valid for significance testing. It is interesting to note that the sumLINK and sumLOD statistics did not frequently agree with regard to the locations of statistically significant peaks in the null data, nor did they generally agree with the HLOD. This perhaps indicates that the three statistics are sensitive to different characteristics of the null data.

Analysis of simulated alternative hypothesis data was based on a GAW13 complex model. An obesity phenotype was selected because it is a complex trait simulated with extensive locus heterogeneity. One major weight gene, *Gb11* was easily identified, with both the sumLOD and sumLINK showing good comparability with the HLOD. Power was low for all other genes, but this was not unexpected. Others who analyzed these data reported that the simulated obesity-related genes, particularly those genes affecting change over time, were very difficult to find.[Strauch, et al. 2003; Yoo, et al. 2003] The data creators intentionally made many of the genes challenging and perhaps even impossible to find. [Daw, et al. 2003] Although the power was low, the sumLINK and sumLOD statistics consistently outperformed the HLOD in identifying the minor genes. However, we do not believe that these new statistics should replace the HLOD, rather that our investigation indicates a proof-of-principle that the sumLINK and/or sumLOD are useful companion measures to help identify the best loci for further testing.

Potential limitations of our method include that the genome-shuffling procedure to create the null distribution may not be useful for studies including only a small number of pedigrees due to the limited number of shuffled genomes that can be generated. The shuffling procedure also assumes that information content is approximately constant across the genome, an assumption that may be violated at the telomeres where multipoint information and information content is reduced systematically. We tested robustness to this by removing all the telomeric regions from the ICPCG data and repeating the analysis. We found that because these regions are such a small part of the entire genome, they do not substantially bias the shuffled null genomes and the results were extremely robust. However, given the difference in information content between the sex chromosomes and the autosomes, we suggest the method for autosomal genome scans only. The term “genome-wide” as used in this manuscript refers only to the 22 autosomes. All of the sumLINK and sumLOD analyses we presented were performed using sex-averaged genetic maps. The effect of this assumption on the characteristics of these new statistics has not been investigated here.

In our example case study of the ICPCG aggressive prostate cancer data we identified 3 regions of interest for further follow-up; two with genome-wide significant evidence supported by FDR analysis, and one with suggestive evidence. This performance is very encouraging. A prior linkage study of these data using conventional LOD/HLOD procedures indicated suggestive linkage evidence at the same loci that we identified on chromosomes 11 and 20 (HLODs of 2.40 for a recessive model and 2.49 for a dominant model, respectively). [Schaid and ICPCG 2006] Our method finds superior levels of significance; both loci are genome-wide significant. However, it is certainly notable that Schaid et al. reported that the evidence on chromosome 11 increased to HLOD=3.31 in subset analyses for early age-at-onset pedigrees, and the region on chromosome 20 increased to HLOD= 2.65 in the subset of pedigrees with mean age-at-onset greater than 65 years.[Schaid and ICPCG 2006] Without necessitating the increased multiple testing inherent from subset analyses, the sumLINK was able to identify the more powerful pedigrees and the superior evidence. Our suggestive region on chromosome 2 was not identified using conventional linkage statistics in the previous study.

CONCLUSION

We have proposed a new statistic to identify linkage regions that have promise for localization and follow-up to gene identification. An R-script is available from the authors that can be used to calculate the sumLINK and sumLOD statistics and generate the null distributions to assess significance of each. We do not claim that these statistics are superior, but that there is evidence that they are useful companion statistics to the HLOD. This method is of particular use within the framework of large collaborative data as it requires neither the sharing of raw data nor the use of common marker sets. We believe this is an important additional statistical tool for identifying linkage regions likely to harbor disease predisposition genes.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The computational resources for this project have been provided by the National Institutes of Health (Grant # NCRR 1 S10 RR17214-01) on the Arches Metacluster, administered by the University of Utah Center for High Performance Computing. The International Consortium for Prostate Cancer Genetics is funded by NCI CA89600 (to William B. Isaacs, Head ICPCG). The Genetic Analysis Workshop is funded by R01 GM031575. This work was also supported by NCI CA98364 (to Nicola J. Camp). Bryce Christensen was supported by National Library of Medicine training grant NLM T15 LM07124.

References

- Abecasis G, Cherny S, Cookson W, Cardon L. Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet.* 2002; 30:97–101. [PubMed: 11731797]
- Almasy L, Amos CI, Bailey-Wilson JE, Cantor RM, Jaquish CE, Martinez M, Neuman RJ, Olson JM, Palmer LJ, Rich SM, et al. Genetic Analysis Workshop 13: Analysis of Longitudinal Family Data for Complex Diseases and Related Risk Factors. *BMC Genetics.* 2003; 4(Suppl 1):S1. [PubMed: 14975069]
- Camp NJ, Farnham JM, Allen-Brady K, Cannon-Albright LA. Statistical recombinant mapping in extended high-risk Utah pedigrees narrows the 8q24 prostate cancer locus to 2.0 Mb. *Prostate.* 2007; 67(13):1456–64. [PubMed: 17654497]
- Camp NJ, Farnham JM, Cannon-Albright LA. Localization of a prostate cancer predisposition gene to an 880-kb region on chromosome 22q12.3 in Utah high-risk pedigrees. *Cancer Research.* 2006; 66(20):10205–12. [PubMed: 17047086]
- Camp NJ, Hopkins PN, Hasstedt SJ, Coon H, Malhotra A, Cawthon RM, Hunt SC. Genome-Wide Multipoint Parametric Linkage Analysis of Pulse Pressure in Large, Extended Utah Pedigrees. *Hypertension.* 2003; 43:322–328. [PubMed: 12874090]
- Chen L, Storey JD. Relaxed Significance Criteria for Linkage Analysis. *Genetics.* 2006; 173:2371–81. [PubMed: 16783025]
- Daw EW, Morrison J, Zhou X, Thomas DC. Genetic Analysis Workshop 13: Simulated longitudinal data on families for a system of oligogenic traits. *BMC Genetics.* 2003; 4(Suppl 1):S3. [PubMed: 14975071]
- Horne BD, Malhotra A, Camp NJ. Comparison of linkage analysis methods for genome-wide scanning of extended pedigrees, with application to the TG/HDL-C ratio in the Framingham Heart Study. *BMC Genetics.* 2003; 4(Suppl 1):S93. [PubMed: 14975161]
- Johanneson B, McDonnell SK, Karyadi DM, Hebring S, Wang L, Deutsch K, McIntosh L, Kwon EM, Suuriniemi M, Stanford JL, et al. Fine mapping of familial prostate cancer families narrows the interval for a susceptibility locus on chromosome 22q12.3 to 1.36 Mb. *Hum Genet.* 2008; 123(1): 65–75. [PubMed: 18066601]

- Klein AP, Kovac I, Sorant AJM, B-B A, Doan BQ, Ibay G, Lockwood E, Mandal D, Santhosh L, Weissbecker K, et al. Importance sampling method of correction for multiple testing in affected sib-pair linkage analysis. *BMC Genetics*. 2003; 4(Suppl 1):S73. [PubMed: 14975141]
- Kong A, Cox NJ. Allele-sharing models: LOD scores and accurate linkage tests. *Am J Hum Genet*. 1997; 61:1179–1188. [PubMed: 9345087]
- Lander E, Kruglyak L. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat Genet*. 1995; 11(3):241–7. [PubMed: 7581446]
- MacLean CJ, Ploughman LM, Diehl SR, Kendler KS. A new test for linkage in the presence of locus heterogeneity. *Am J Hum Genet*. 1992; 50:1259–1266. [PubMed: 1598905]
- Orr A, Dubé M, Marcadier J, Jiang H, Federico A, George S, Seamone C, Andrews D, Dubord P, Holland S, et al. Mutations in the UBIAD1 Gene, encoding a potential prenyltransferase, are causal for Schnyder Crystalline Corneal Dystrophy. *PLoS ONE*. 2007; 2(8):e685. [PubMed: 17668063]
- R Development Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2006.
- Schaid D. ICPCG. Pooled genome linkage scan of aggressive prostate cancer: Results from the International Consortium for Prostate Cancer Genetics. *Hum Genet*. 2006; 120(4):471–85. [PubMed: 16932970]
- Schaid DJ, Chang BL. Description of the international consortium for prostate cancer genetics, and failure to replicate linkage of hereditary prostate cancer to 20q13. *Prostate*. 2004
- Strauch K, Golla A, Wilcox MA, Baur MP. Genetic analysis of phenotypes derived from longitudinal data: presentation group 1 of Genetic Analysis Workshop 13. *Genetic Epidemiology*. 2003; 25(Supplement 1):S5–S17. [PubMed: 14635164]
- Yoo YJ, Huo Y, Ning Y, Gordon D, Finch S, Mendell NR. Power of maximum HLOD tests to detect linkage to obesity genes. *BMC Genetics*. 2003; 4(Suppl 1):S16. [PubMed: 14975084]

Figure 1.

Shuffling procedure for creating null distribution. A) Raw test statistic is calculated across un-shuffled data at regular intervals throughout the genome. Figure shows five pedigrees (rows) and four chromosomes (columns). B) Chromosomes are connected end-to-end in random order within each pedigree, and the resulting loop is broken at a random location to create a new starting point. C) The new starting points are aligned, and null statistics can be calculated along the shuffled genome.

Figure 2.

Simplistic illustration of the calculation of the sumLINK statistic. The linkage evidence for three pedigrees (broken lines) are shown across 6 loci. The sumLINK calculated from these three pedigrees is shown by the heavy black line. Pedigree LOD scores that are marked with a diamond meet the threshold for inclusion and these are summed to produce the sumLINK. Pedigree LOD scores marked with a cross do not meet the threshold for inclusion and do not contribute to the sumLINK.

Figure 3. Genome-wide multipoint sumLINK results (dominant model) and sumLOD results (recessive model) for the ICPCG aggressive prostate cancer data.

Figure 4.

LOD traces for each pedigree contributing to the linkage results on chromosomes A) 20-dominant, B) 11-recessive, and C) 2-dominant. Black bars indicate the two-recombinant localization regions.

Table I

Simulated Data Characteristics

	Null Simulation (×250)	Selected GAW13 Simulation Pedigrees		
		All	SumLINK set	SumLOD-only set
Pedigrees	400	5232	1056	4176
Individuals	2673	44,326	13,568	30,758
Persons per ped	6.68 (4 to 16)	8.47 (4 to 25)	12.85 (6 to 25)	7.37 (4 to 22)
Mean generations	2.44 (2 to 4)	2.38 (2 to 4)	2.83 (2 to 4)	2.27 (2 to 4)
Total Affected	1627 (60.9%)	20,889 (47.1%)	6363 (46.9%)	14,526 (47.2 %)
Mean Affected	4.07 (3 to 11)	3.42 (2 to 12)	4.78 (2 to 12)	3.07 (2 to 11)
Total Genotyped	1627 (60.9%)	25808 (58.2%)	7590 (55.9%)	18218 (59.2%)
Mean Genotyped	4.07 (3 to 11)	4.93 (2 to 15)	7.19 (3 to 15)	4.36 (2 to 14)
Marker Type	SNP	STRP	STRP	STRP
Marker Number	2936	399	399	399

Table II

False positive rates estimated from 250 genome-wide replicates under the null hypothesis of no linkage

Statistic	Analysis Model	Statistical FPR threshold surpassed	
		Significant (0.05) [*]	Suggestive (1.00) [†]
sumLINK	Dominant	0.056	0.936
	Recessive	0.060	1.124
sumLOD	Dominant	0.052	1.020
	Recessive	0.052	0.912
OVERALL		0.055	0.998

^{*} 95% CI for 0.05 in 250 replicates is [0.024,0.080]

[†] 95% CI for 1.00 in 250 replicates is [0.876,1.124]

TABLE III

Power to detect at least suggestive linkage evidence in 100 simulations. All loci detected at least 5 times by any of the statistics are shown.

Gene	Affected Trait	Power (%)		
		sumLINK	sumLOD	HLOD
Gb11,Gb4 [*]	Weight-baseline, Height-baseline	99	100	100
Gb2	Height-baseline	17 [†]	22 [†]	3
Gb20	HDL-baseline	9	14 [†]	4
Gb22,Gs3 ^{**}	HDL, Triglycerides	6 [†]	8 [†]	1
Gb15,Gs10, Gs12,Gb37 ^{**}	HDL, Triglycerides, Glucose, SBP, DBP	7 [†]	7 [†]	1
Gs4	Triglycerides, Glucose.	4	7 [†]	1
Gs2	Weight-slope	6	5	3
Gs8	Cholesterol-slope	5	6 [†]	1
Gb24	Triglycerides-baseline	2	6 [†]	0
Gb5	Height-baseline	1	5 [†]	0

^{*} Gb4 was observed independently several times, but was often obscured by the broad peak at Gb11.

^{**} Genes were positioned too closely on the chromosome to discern which was responsible for the linkage signal;

[†] significantly greater power than the HLOD. None of the differences observed between the sumLINK and sumLOD are statistically significant. Gene characteristics can be found in [Daw, et al. 2003].

Table IV

Summary of significant and suggestive linkage peaks

sumLINK						
Model	Chrom.	cM Position	sumLINK	Expected Frequency (FPR)	Number of contributing pedigrees	FDR q-value
Dominant	20	59	13.848	0.005	17	0.017
	2	69	10.837	0.628	13	0.175
Recessive	11	89	10.941	0.598	14	0.128
	2	68	10.624	0.897	14	0.209
SumLOD						
Model	Chrom.	cM Position	sumLOD	Expected Frequency (FPR)	Number of contributing pedigrees	FDR q-value
Recessive	11	89	27.975	0.007	87	0.017
Dominant	20	59	30.311	0.028	83	0.017
	11	89	27.650	0.657	81	0.213