Published in final edited form as:

Genet Epidemiol. 2009 ; 33(Suppl 1): S81–S87. doi:10.1002/gepi.20477.

Combining Information from Linkage and Association Methods

Elizabeth E. Marchani¹, Andrea Callegaro², E. Warwick Daw³, and Ellen M. Wijsman^{1,4}

¹ Division of Medical Genetics, Department of Medicine, University of Washington, Seattle, WA ² Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, The Netherlands ³ Division of Statistical Genomics, Washington University School of Medicine, St. Louis, MO ⁴ Department of Biostatistics and Department of Genome Sciences, University of Washington, Seattle, WA

Abstract

Group 12 evaluated approaches to incorporate outside information or otherwise optimize traditional linkage and association analyses. The abundance of available data allowed exploration of identity-by-descent (IBD) estimation, score statistics, formal combination of linkage and association testing, significance estimation, and replication. We observed that IBD estimation can be optimized with a subset of marker data while estimation of inheritance vectors can provide both IBD estimates and a measure of their uncertainty. Score statistics incorporating covariates or combining association and linkage information performed at least as well as standard approaches while requiring less computation time. The formal combination of linkage and association methods may be fruitful, although the nature of the simulated data limited our conclusions. Estimation of significance may be improved through simulation, correction for cryptic relatedness, and the inclusion of prior information. Replication using real data provided consistent results, though the same was not true of simulated data replicates. Overall, we found that increasing the amount of available data limits analyses due to computational constraints and motivates the need to improve methods for the identification of complex-trait genes.

Keywords

score statistic; empirical significance; replication

INTRODUCTION

Ever since the identification of genes influencing complex traits proved to be feasible, as with breast cancer [Miki et al., 1994], Alzheimer's disease [Levy-Lahad et al., 1995; Sherrington et al., 1995], and diabetes [Yamagata et al., 1996], the challenge to locate additional complex-trait genes has been a major theme in genetic epidemiological research. Several general strategies have been used for this purpose, dictated in part by the availability of resources and implicit assumptions about the underlying genetic architecture of complex traits. Most of these strategies are based on either pedigree- or population-based designs.

The strengths and weaknesses of pedigree- and population-based designs are complementary. Pedigree-based designs employ linkage analysis methods based on modeling aspects of the meiotic process. These designs depend only on transmission of

genes within families, and thus are relatively robust to population structure and allelic heterogeneity at trait loci. Pedigree-based designs do not pose a major multiple-testing problem because of the coarse granularity of the meiotic process. Weaknesses include sample acquisition, computational demands, and low power to detect common variants.

Population-based designs use data that reflect the effects of meiotic events occurring over many generations. The advantages of population-based designs include ease of sample acquisition, computational simplicity, and higher power in the presence of a small number of high-frequency trait alleles, especially for alleles with small effects. The resulting fine granularity of the population-level correlation between specific marker alleles and the trait phenotype requires dense marker spacing, which can lead to a severe multiple testing problem for genome-wide studies. Limiting analyses to candidate regions of the genome reduces this multiple testing problem at the cost of genome-wide scope. Population-based studies are also vulnerable to allelic heterogeneity, population and sample structure, and low power to detect associations with rare trait alleles.

The complementary strengths of these two designs suggest that there are advantages to combining both strategies. This was a key theme discussed by the diverse papers of the Group 12 members of Genetic Analysis Workshop 16 (GAW16). Other themes include identity by descent (IBD) and significance estimation, score statistics for linkage analysis, replication, and the integration of information in the search for complex-trait genes.

METHODS

Overall strategies

Group 12 incorporated information on pedigrees, phenotypic traits and their covariates, prior linkage or association signals, and longitudinal phenotypic data. These outside sources of information allowed us to focus analyses on a subset of the available data or to improve the performance of existing methodology by formally incorporating this outside information into the analyses. Group-specific approaches are presented in Table I and summarized below.

Marker data

Given the large amount of data, groups made heterogeneous choices regarding marker use. Some used all 550k SNPs (single-nucleotide polymorphisms) in the real Framingham Heart Study (FHS) markers [Callegaro et al., 2009; Yoo et al., 2009a], while others used subsets of this data [Gray-McGuire et al., 2009; Marchani et al., 2009]. A few groups used the 550k simulated FHS data [Daw et al., 2009; Hendricks et al., 2009], or the 550k North American Rheumatoid Arthritis Consortium (NARAC) data set [Yoo et al., 2009a]. For clarity, the discussion here focuses primarily on the FHS data set.

Individual markers were excluded due to perceived low data quality. Groups excluded markers with low minor allele frequencies and markers with low genotype call rates. Several groups used Hardy-Weinberg equilibrium testing to eliminate markers, while one group excluded markers with high Mendelian error rates [Yoo et al., 2009a].

Most groups used knowledge of linkage disequilibrium to reduce their marker panel, using either a threshold correlation between neighboring markers [Hendricks et al., 2009; Yoo et al., 2009a] or a regular marker interval of 1 to 10 SNPs/cM [Callegaro et al., 2009; Daw et al., 2009; Gray-McGuire et al., 2009; Marchani et al., 2009]. Most groups limited their analyses to chromosomes with prior association or linkage signals for their phenotype of interest.

Pedigrees

Initial data cleaning focused on improving data quality and formatting the data to suit each analytic approach. Most groups used all FHS cohorts, though one pooled the three cohorts [Callegaro et al., 2009] and another used only the Offspring and Generation 3 cohorts [Yoo et al., 2009a]. Several groups limited at least some of their linkage analyses to nuclear families or sibpairs. Subsets of these families were excluded, including whole families missing phenotype data [Gray-McGuire et al., 2009; Li et al., 2009] and individuals with a low genotype call rate or unreliable information for sex determination. Mendelian inconsistencies were removed from the pedigree files, and monozygotic twins were pooled. There was some confusion about the multiple pedigree files included in the data distribution, and so one group reconstructed the pedigree file using parent-offspring information alone [Marchani et al., 2009].

Phenotypes

Phenotypes were adjusted to increase power to detect association or linkage. Many groups used regression to adjust phenotypes for known covariates. Longitudinal data provided several ways to adjust phenotypes, including the use of age-at-onset [Callegaro et al., 2009], using the most recent observation of an age-dependent phenotype [Gray-McGuire et al., 2009], and comparing trait values over time or using average trait values [Hendricks et al., 2009; Li et al., 2009; Yoo et al., 2009a].

Analytical methods

Estimating IBD—Broadly defined, IBD information is used by every linkage analysis method in the context of either meiotic transmissions or sets of alleles shared IBD. For this reason, many groups used existing methods to estimate IBD, including the tools within the SOLAR [Daw et al., 2009; Li et al., 2009], Loki [Hendricks et al., 2009; Marchani et al., 2009], MERLIN [Callegaro et al., 2009; Yoo et al., 2009a], and S.A.G.E. [Gray-McGuire et al., 2009] software packages. A novel method to estimate inheritance vectors conditional on the marker data was also implemented in MORGAN [MORGAN; Marchani et al., 2009]. This method includes a new meiosis sampler and samples sequentially over pedigrees, allowing Markov-chain Monte Carlo (MCMC) sampling for large pedigrees and exact computation for small pedigrees to reduce computational time. The sensitivity of these estimates to the number of markers in the data set was also tested to determine the amount of data that maximized information content while minimizing computational time and multiple testing problems [Marchani et al., 2009].

Score statistics—Three groups used score tests for genetic linkage analysis. Hendricks et al. [2009] compared the score test derived by Tang and Siegmund [2001] to the likelihood-ratio tests (LRTs) for linkage and association, and introduced a combined conditional linkage and association score test, discussed in greater detail below. Callegaro et al. [2009] derived two new nonparametric linkage (NPL) statistics weighted by age at onset. The first test is based on a gamma frailty model that uses a composite likelihood to relieve much of the computational burden for analyzing large pedigrees, while the second test uses a log-normal frailty model which uses a second-order Taylor approximation around the null random effect to approximate the log-likelihood of the model. Gray-McGuire et al. [2009] applied the score test proposed by Wang and Elston [2007] for multivariate linkage analysis. This approach allows both multiple individual-level covariates and a multivariate phenotype constructed from multiple trait values. The score statistics used by Hendricks et al. [2009] and Callegaro et al. [2009] are derived from the probability of the genetic data conditional on the observed phenotype [Tang and Siegmund, 2001]. These score tests are rendered valid under the null hypothesis and robust to departure from normality by this conditional

likelihood approach. In contrast, the score test applied by Gray-McGuire et al. [2009] is derived from the prospective likelihood of the phenotype conditional on the genetic data. This approach maintains correct type I error even when the data are not normally distributed because a robust sandwich-type estimator is used.

Combined linkage and association analyses—Two methods to explicitly integrate linkage and association analyses were explored using the simulated data set. The combined linkage and association score statistic used by Hendricks et al. [2009] takes advantage of the assumed independence between the conditional association score and linkage score under the null hypothesis of no linkage and no association, and is the sum of the chi-square forms of the conditional association and linkage score statistics. The observed value of the combined linkage and association score statistic is then compared with the appropriately weighted chi-square distributions to determine significance. Another group used a single Bayesian MCMC framework to perform segregation, linkage, and association analyses [Daw et al., 2009]. Association between a SNP and the disease was measured as a z -score under the null hypothesis of no effect on the disease variable. The authors model linkage and association in separate terms within the analytical model. If a single marker explained a linkage signal, accounting for that through association should eliminate the linkage signal while still capturing the variance explained by the locus. This strategy could be used to identify the specific contribution of a single locus to a linkage signal.

Estimation of significance—Two groups used simulation methods to obtain empirical p -values. Gray-McGuire et al. [2009] estimated p -values for multivariate linkage analyses by simulating the appropriate asymptotic null distributions [Morris et al., 2008]. Marchani et al. [2009] performed trait resimulation [Igo and Wijsman, 2008], to provide empirical p -values based on the actual marker data for variance-component linkage analyses. The real data were also used in a conditional inheritance vector test to estimate linkage and to describe the uncertainty in the estimated linkage statistic [Di and Thompson, 2009].

Rather than adjust the analytical framework itself, many groups addressed the loss of power introduced by the multiple testing problem by incorporating outside information to reduce the number of markers tested (Table I). A novel method [Choi et al., 2009] was used to adjust p -values from an association test in the presence of known or cryptic relationships [Marchani et al., 2009], addressing correlation in the data set that might lead to spurious p -values. Yoo et al. [2009a] compared the performance of the unadjusted false-discovery rate (FDR) to the weighted FDR (wFDR) [Roeder et al., 2006] and the stratified FDR (sFDR) [Sun et al., 2006]. They used the approach of Storey [2003] to estimate the fraction of tests drawn from the null vs. alternative distributions, yielding results that are similar to a Bonferroni correction when almost all tests are drawn from the null distribution. Both modified FDR approaches use prior evidence such as from linkage analysis results to assign weights by either using the linkage statistic directly (wFDR) or by grouping genomic regions into strata (sFDR). The wFDR is more versatile in utilizing SNP-specific weights, whereas stratum-specific weights can provide robustness to uninformative or misspecified priors Yoo et al. [2009b]. This approach is common, if not formally recognized, e.g., the focused approach employed by Li et al. [2009] can be thought of as an extreme example of the sFDR approach in which some tests are assigned a weight of zero.

Replication—The availability of many markers and phenotypes provided an opportunity to analyze multiple subsets of the data in a process that can loosely be termed replication. We define successful replication broadly and qualitatively, looking for loci that show significant association or linkage signals under one method and that at least approach significance in subsequent analyses. Because no two groups used the same criteria for testing, a more precise evaluation is not possible here. The most obvious form of replication involved

examining multiple subsets of markers, families, or replicates of the simulated data [Daw et al., 2009; Hendricks et al., 2009; Marchani et al., 2009]. Relationships between phenotypes motivated some to compare regions associated with correlated traits or a single phenotype measured at different times [Gray-McGuire et al., 2009; Li et al., 2009]. Still others compared linkage and association signals derived from multiple analytical approaches, including Bayesian oligogenic joint segregation and linkage analysis, regression models, variance-components linkage analysis, association testing, and score statistics (Table I).

RESULTS

Estimating IBD

It is important to identify a threshold representing the number of markers that maximizes the amount of available IBD information without overburdening computational resources. Marchani et al. [2009] found that estimates of IBD sharing among pairs of individuals stabilized as the number of markers genome-wide increased. However, densities of 1 SNP per 3 cM vs. 1 SNP per 1 cM gave similar results, suggesting that increasing marker density beyond this point has little advantage.

Score statistics

The novel score tests evaluated by Group 12 appear to perform at least as well as their predecessors [Callegaro et al., 2009; Gray-McGuire et al., 2009; Hendricks et al., 2009]. Hendricks et al. [2009] found that LRTs and score tests for linkage produced comparable results with similar power. Type I error of the LRT and conditional score tests was inflated under the general 2 degrees-of-freedom genetic model, though using the additive model drastically reduced power for non-additive loci. Callegaro et al. [2009] demonstrated that, although their two weighted NPL score statistics and the unweighted NPL test each have correct type I error, the weighted NPL statistics have considerably more power than the unweighted NPL test. The necessary estimation of multipoint IBD probabilities can be time-consuming in the case of complex pedigrees or large numbers of markers. However, different score tests [Callegaro et al., 2009; Gray-McGuire et al., 2009; Hendricks et al., 2009] can be applied to those estimates, allowing rapid evaluation of many traits.

Combined linkage and association analyses

The integration of linkage and association methodologies proved promising, but the evaluation of these methods was limited by the simulated data available. Hendricks et al. [2009] found the combined linkage and association score statistic to have comparable power to the LRT and conditional association score statistic. This indicates that the combined use of linkage and association methods may compensate for the extra degree of freedom used by the combined score statistic. Daw et al. [2009] found that SNPs associated with the phenotype reduced linkage signals as they moved from the segregation to the association term in the analytical model. Their combined approach was also able to identify approximately one-third of the associated SNPs, primarily those with heritabilities >1%.

Estimation of significance

Trait resimulation produced the expected null-distribution of test statistics and provided a tractable approach to obtaining empirical *p*-values when the computational cost of marker IBD is high [Marchani et al., 2009]. The conditional inheritance vector test was able to use the data to identify two loci within a single linkage peak and to provide what are essentially confidence limits at those positions [Marchani et al., 2009]. The properties of the simulation method used by Gray-McGuire et al. [2009] were described elsewhere [Morris et al., 2008],

but the approach appeared to provide greater precision in determining linkage regions while identifying additional regions of interest.

Association tests that incorporated estimated kinship coefficients provided a simple correction for cryptic relatedness. This computationally straightforward approach required few assumptions and eliminated the excess of spuriously significant test results generated by ignoring relationship information [Marchani et al., 2009]. Similar results were also obtained using pedigree data or marker data alone, suggesting that this approach is useful in a variety of situations.

The modification of significance levels to incorporate prior information appears to be useful. The results obtained were internally consistent, although it is important to temper conclusions made on the basis of real data due to the unobserved underlying truth. For example, for markers strongly associated with a phenotype, the different approaches for adjusting the FDR had little effect on the conclusion, as one would hope [Yoo et al., 2009a]. However, as was noted by the originators of the wFDR [Roeder et al., 2006], very strong linkage signals had a more noticeable effect on the rank of the resulting wFDR than the sFDR. This was more noticeable for the NARAC data than for the FHS data, which had more modest linkage signals and relatively consistent wFDR and sFDR results.

Replication

Many approaches were gratifyingly consistent with each other. Different marker subsets provided similar results for linkage within a data set, even for quantitative trait loci with small effects [Daw et al., 2009]. Similarly, the consistent linkage signals generated by different family subsets suggest that similar effects were present in different families [Marchani et al., 2009]. Correlated phenotypes and repeated measures identified the same genomic regions of association, and may be useful for identifying genes contributing to a phenotype at different life stages, in different environments, or for disease processes defined by multiple sub-phenotypes [Gray-McGuire et al., 2009; Li et al., 2009]. Although a greater specificity was observed among linkage results in the real data [Marchani et al., 2009], linkage and association methods generally agreed with each other when compared within a single simulated data set [Daw et al., 2009; Hendricks et al., 2009]. However, no consistent trends were identified in the location of linkage or association signals across simulated data sets, regardless of analytical approach tested [Daw et al., 2009]. A possible reason for this result is discussed below.

DISCUSSION

Lessons learned

Improved computational speed—Data sets and analytical methods were both modified by participants to improve computational speed. Marker sets at 1/cM density were sufficient to yield consistent IBD estimates, allowing researchers to estimate IBD without using dense marker data [Marchani et al., 2009]. Score tests performed comparably to the LRT, but with much greater computational speed [Hendricks et al., 2009]. Both simulation of the null distribution and trait resimulation provided computationally tractable ways to obtain empirical *p*-values [Gray-McGuire et al., 2009; Marchani et al., 2009]. All of these time-saving developments will prove especially useful when the quest to identify trait loci with small effects requires ever-larger sample sizes, which add computational burden through the effects of sheer numbers.

Improved signal quality—Large data sets provide the luxury of being able to throw away low-quality data while creating replicate data sets to validate linkage and association

signals. In general, this replication or incorporation of additional information into analytical methods proved fruitful, and integration of linkage and association analyses showed promise. However, random effects often determined which loci with very small effects were detectable [Daw et al., 2009; Hendricks et al., 2009]. Association testing while controlling for relatedness within a sample proved valuable: evidence of cryptic relatedness was identified within the FHS [Marchani et al., 2009]. Use of weights or strata to incorporate prior information proved to be a potentially powerful tool, though care is required when choosing weight or stratification thresholds [Yoo et al., 2009a].

Constraints—Although the increasing amount of data available for analysis provides a terrific resource, it also introduces critical constraints. GAW16 participants were limited by data download times and the time necessary to adequately perform quality control and data cleaning. Participants were unable to carry out deep data exploration or quality control analysis, such as looking for double recombinants or phenotypic outliers. Even after reducing the number of markers, families, and phenotypes for study, the time needed to perform analyses on subsets of the data was still lengthy. These practical issues and demands for computational and analytical resources should not be dismissed or overlooked in future analyses, if the value of large public data sets is to be realized.

The inability to replicate linkage or association signals across the simulated data sets demonstrated the difficulty of detecting subtle genetic effects using both linkage and association methods. The simulated data, by design, had multiple genes with very small effect, making such genes difficult to detect. Because the underlying genotypes were the same, one possible explanation is that random variation in the phenotype may alternatively dampen or enhance small genetic signals. Rather than being “false positives,” these signals may represent random noise enhancing a true signal. A better understanding of the underlying causes of these non-replications would be useful for interpretation of analyses of real complex traits.

Future possibilities—Although LRTs can have greater power than score statistics when there are large linkage effects, such large effects are not likely for common complex traits. Score tests are relatively robust to distributional assumptions concerning random effects, which is encouraging considering the impact of random effects in the identification of loci with small effects in the simulated data. However, one disadvantage of score tests is that they tend to be conservative. It is possible that MCMC-sampled inheritance vectors could be used to estimate the variance around IBD estimates, similar to the conditional inheritance vector test used by Marchani et al. [2009]. This variance could then be used to correct the conditional score statistics and thus improve power.

New tools are necessary to improve the power to detect variants with rare allele frequencies or small effect sizes. The development of these tools could be facilitated by the availability of simulated data sets with a greater variety of underlying quantitative trait locus models. Simulated data representing the null hypothesis, with no disease-associated variants on a test chromosome, are also necessary. The usefulness of a simulated data set is also a function of its accessibility to multiple investigators, as is being realized with real data.

Finally, rare variants and genotyping errors can easily be confused. The introduction of error modeling to current methods, combined with the use of pedigree data, might be used to distinguish rare SNPs from genotyping errors. The data cleaning that was necessary for linkage analysis identified many genotyping errors that otherwise would have gone undetected, illustrating another potential advantage of combining pedigree information with association testing. At this point, the importance of identifying such additional errors through use of pedigree information is unknown, but merits future investigation.

CONCLUSIONS

Combining information from linkage and association studies is prevalent in the literature, although rarely formalized. For example, researchers restrict their analyses to regions with previous linkage or association signals, or incorporate that information in the discussion of their results to strengthen their conclusions. The lack of a formal approach in this context means that the underlying assumptions, prior beliefs, and even hypotheses are not obvious and may be overlooked. Prior information should be incorporated explicitly into analytical strategies, with careful attention paid to the corresponding weights. This may help prevent bias in the interpretation of results because the importance of previous signals is set *a priori*. This approach also facilitates a sensitivity analysis in which weights on prior information may be varied and their effects on conclusions determined. This is important because we have no clear guidelines for establishing whether one approach is fundamentally better than another. Future development of decision rules and guidelines for weighting prior information, as well as new methods to integrate pedigree information and association analyses, will be necessary to take full advantage of the complementary information provided by these approaches.

Acknowledgments

The authors thank GAW16 Group 12 members for their comments and assistance. The authors are supported by NIH grants AG05136, HD55782, HD54562, HL30086, GM46255, and AG00258. The Genetic Analysis Workshop is supported by NIH grant GM031575.

References

- Callegaro A, Uh H-W, Helmer Q, Houwing-Duistermaat JJ. New score tests for age-at-onset linkage analysis in general pedigrees. *BMC Proc.* 2009; 3(Suppl 7):S97. [PubMed: 20018094]
- Choi Y, Wijsman EM, Weir BS. Case-control association testing in the presence of unknown relationships. *Genet Epidemiol.* 2009 Epub ahead of print.
- Daw EW, Plunkett J, Feitosa M, Gao X, Van Brunt A, Ma D, Czajkowski J, Province MA, Borecki I. A framework for analyzing both linkage and association: An analysis of Genetic Analysis Workshop 16 simulated data. *BMC Proc.* 2009; 3(Suppl 7):S98. [PubMed: 20018095]
- Di Y, Thompson EA. Conditional tests for localizing trait genes. *Hum Hered.* 2009; 68:139–50. [PubMed: 19439976]
- Gray-McGuire C, Song Y, Morris NJ, Stein CM. Comparison of univariate and multivariate linkage analysis of traits related to hypertension. *BMC Proc.* 2009; 3(Suppl 7):S99. [PubMed: 20018096]
- Hendricks AE, Zhu Y, Dupuis J. Genome-wide association and linkage analysis of quantitative traits: Comparison of likelihood ratio test and conditional score statistic. *BMC Proc.* 2009; 3(Suppl 7):S100. [PubMed: 20017964]
- Igo R, Wijsman EM. Empirical significance values for linkage analysis: Trait simulation using posterior model distributions from MCMC oligogenic segregation analysis. *Genet Epidemiol.* 2008; 32:119–31. [PubMed: 17849492]
- Levy-Lahad E, Wasco W, Poorkaj P, Romano DM, Oshima J, Pettingell WH, Yu CE, Jondro PD, Schmidt SD, Wang K, Crowley AC, Fu YH, Guenette SY, Galas D, Nemens E, Wijsman EM, Bird TD, Schellenberg GD, Tanzi RE. Candidate gene for the chromosome-1 familial Alzheimer's disease locus. *Science.* 1995; 269:973–7. [PubMed: 7638622]
- Li X, Mei L, Yang K, Rotter JI, Guo X. Identifying association under a previous linkage peak on chromosome 16 for body mass index using cross-sectional and longitudinal data of the Framingham Heart Study. *BMC Proc.* 2009; 3(Suppl 7):S101. [PubMed: 20017965]
- Marchani EE, Di Y, Choi Y, Cheung C, Su M, Boehm F, Thompson EA, Wijsman EM. Contrasting identity-by-descent estimators, association studies, and linkage analyses using the Framingham Heart Study data. *BMC Proc.* 2009; 3(Suppl 7):S102. [PubMed: 20017966]

- Miki Y, Swensen J, Shattuckeids D, Futreal PA, Harshman K, Tavtigian S, Liu QY, Cochran C, Bennett LM, Ding W, Bell R, Rosenthal J, Hussey C, Tran T, McClure M, Frye C, Hattier T, Phelps R, Haugenstrano A, Katcher H, Yakumo K, Gholami Z, Shaffer D, Stone S, Bayer S, Wray C, Bogden R, Dayananth P, Ward J, Tonin P, Narod S, Bristow PK, Norris FH, Helvering L, Morrison P, Rostek P, Lai M, Barrett JC, Lewis C, Neuhausen S, Cannonalbright L, Goldgar D, Wiseman R, Kamb A, Skolnick MH. A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science*. 1994; 266:66–71. [PubMed: 7545954]
- MORGAN. <http://www.stat.washington.edu/thompson/Genepi/MORGAN/Morgan.shtml>
- Morris NJ, Elston RC, Stein CM. Calculating asymptotic significance levels of the constrained likelihood ratio test with application to multivariate genetic linkage analysis. *Stat Appl Genet Mol Biol*. 2009; 8(1) Article 39. Epub 2009 Sep 17.
- Roeder K, Bacanu SA, Wasserman L, Devlin B. Using linkage genome scans to improve power of association genome scans. *Am J Hum Genet*. 2006; 78:243–52. [PubMed: 16400608]
- Sherrington R, Rogaev EI, Liang Y, Rogaeva EA, Levesque G, Ikeda M, Chi H, Lin C, Li G, Holman K, Tsuda T, Mar L, Foncin JF, Bruni AC, Montesi MP, Sorbi S, Rainero I, Pinessi L, Nee L, Chumakov I, Pollen D, Brookes A, Sanseau P, Polinsky RJ, Wasco W, Dasilva HAR, Haines JL, Pericakvance MA, Tanzi RE, Roses AD, Fraser PE, Rommens JM, Stegeorgehyslop PH. Cloning of a gene bearing missense mutations in early-onset familial Alzheimer's disease. *Nature*. 1995; 375:754–60. [PubMed: 7596406]
- Storey JD. The positive false discovery rate: A Bayesian interpretation and the q-value. *Ann Stat*. 2003; 31:2013–35.
- Sun L, Craiu RV, Paterson AD, Bull SB. Stratified false discovery control for large scale hypothesis testing with application to genome-wide association studies. *Genet Epidemiol*. 2006; 28:352–67. [PubMed: 15761913]
- Tang HK, Siegmund D. Mapping quantitative trait loci in oligogenic models. *Biostatistics*. 2001; 2:147–62. [PubMed: 12933546]
- Wang T, Elston RC. Regression-based multivariate linkage analysis with an application to blood pressure and body mass index. *Ann Hum Genet*. 2007; 71:96–106. [PubMed: 17227480]
- Yamagata K, Oda N, Kaisaki PJ, Menzel S, Furuta H, Vaxillaire M, Southam L, Cox RD, Lathrop GM, Boriraj VV, Chen XN, Cox NJ, Oda Y, Yano H, LeBeau MM, Yamada S, Nishigori H, Takeda J, Fajans SS, Hattersley AT, Iwasaki N, Hansen T, Pedersen O, Polonsky KS, Turner RC, Velho G, Chevre JC, Froguel P, Bell GI. Mutations in the hepatocyte nuclear factor-1 alpha gene in maturity-onset diabetes of the young (MODY3). *Nature*. 1996; 384:455–8. [PubMed: 8945470]
- Yoo YJ, Pinnaduwaage D, Waggott D, Bull SB, Sun L. Genome-wide association analyses of North American Rheumatoid Arthritis Consortium and Framingham Heart Study data utilizing genome-wide linkage results. *BMC Proc*. 2009a; 3(Suppl 7):S103. [PubMed: 20017967]
- Yoo YJ, Bull SB, Paterson AD, Waggott D, Sun L. The Diabetes Control and Complications Trial/ Epidemiology of Diabetes Interventions and Complications Research Group. Were genome-wide linkage studies a waste of time? Exploiting candidate regions within genome-wide association studies. *Genet Epidemiol*. 2009b Epub ahead of print.

Table 1

Summary of Group 12 approaches

Group 12 manuscript						
	Callegaro et al.	Daw et al.	Gray-McGuire et al.	Hendricks et al.	Li et al.	Marchani et al.
Families	Nuclear	Full pedigree	Sibpairs	Full pedigree	Nuclear, Full pedigree	Full pedigree
Filters ^a	MAF, map information, map density, information content	Map density	Map density, Mendelian consistent	Call rate, sex diagnosis, HWE, Map information, MAF, Mendelian consistent	MAF, Mendelian consistent, HWE, call rate	MAF, HWE, call rate, Mendelian consistent, map density
Scope	Genome-wide	Chromosome	Chromosomes	Genome-wide	Chromosome	Chromosome, genome-wide
Traits ^b	AGE	LDL, HDL	SBP, TG, HDL, LDL	LDL, TG	BMI	HDL
Covariates ^c	Sex	Age, Age ² , Age ³ , Sex	Age, Sex, Age*Sex	Age, Sex	Age, Sex, Cohort	Age, Sex, BMI, Rx, Cohort
Analyses ^d	Regression-based linkage analysis, score test	Bayesian oligogenic joint segregation and linkage analysis, association, VC linkage analysis	Regression-based linkage analysis	Association, VC linkage analysis, score test	Association, VC linkage analysis, quantitative trait TDT, quantitative trait linkage disequilibrium	VC linkage analysis, Bayesian oligogenic joint segregation and linkage analysis, association, IBD explicitly, regression-based linkage analysis
Significance	Asymptotic <i>p</i> -value	asymptotic <i>p</i> -value	Asymptotic <i>p</i> -value, simulation	Asymptotic <i>p</i> -value	Permutation	Asymptotic <i>p</i> -value, relationship-corrected <i>p</i> -value, permutation
						FDR

^aMAF, minor allele frequency; HWE, Hardy-Weinberg equilibrium

^bAGE, age at first of coronary artery disease, diabetes, or death; LDL, low-density lipoprotein; HDL, high-density lipoprotein; TG, triglycerides; SBP, systolic blood pressure

^cRx, cholesterol treatment status

^dVC, variance components; TDT, transmission-disequilibrium test