Genetic Epidemiology 25 (Supplement 1): S18–S28 (2003)

Longitudinal Data Analysis in Pedigree Studies

W. James Gauderman,[1]* Stuart Macgregor,[2] Laurent Briollais,[3] Katrina Scurrah,[4] Martin Tobin,[4] Taesung Park,[5] Dai Wang,[6] Shaoqi Rao,[7] Sally John,[8] and Shelley Bull[3]

[1]*Department of Preventive Medicine, University of Southern California, Los Angeles, California*
[2]*Institute of Cell, Animal and Population Biology, Ashworth Laboratories, University of Edingurgh, Scotland, UK*
[3]*Division of Epidemiology and Biostatistics, Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, Ontario, Canada, and Department of Public Health Sciences, University of Toronto, Toronto, Ontario, Canada*
[4]*Institute of Genetics and Department of Epidemiology and Public Health, University of Leicester, Leicester, England, UK*
[5]*Department of Statistics, Seoul National University, Seoul, Korea and Department of Biostatistics, University of Pittsburgh, Pittsburgh, Pennsylvania*
[6]*Division of Medical Genetics, Medical Genetics Birth Defect Center, Cedars-Sinai Research Institute, Cedars-Sinai Medical Center, Los Angeles, California*
[7]*Center for Cardiovascular Genetics, Department of Cardiovascular Medicine and Department of Molecular Cardiology, Lerner Research Institute, Cleveland Clinic Foundation, Cleveland, Ohio*
[8]*Centre for Integrated Genomic Medical Research, University of Manchester, Manchester, England, UK*

Longitudinal family studies provide a valuable resource for investigating genetic and environmental factors that influence long-term averages and changes over time in a complex trait. This paper summarizes 13 contributions to Genetic Analysis Workshop 13, which include a wide range of methods for genetic analysis of longitudinal data in families. The methods can be grouped into two basic approaches: 1) two-step modeling, in which repeated observations are first reduced to one summary statistic per subject (e.g., a mean or slope), after which this statistic is used in a standard genetic analysis, or 2) joint modeling, in which genetic and longitudinal model parameters are estimated simultaneously in a single analysis. In applications to Framingham Heart Study data, contributors collectively reported evidence for genes that affected trait mean on chromosomes 1, 2, 3, 5, 8, 9, 10, 13, and 17, but most did not find genes affecting slope. Applications to simulated data suggested that even for a gene that only affected slope, use of a mean-type statistic could provide greater power than a slope-type statistic for detecting that gene. We report on the results of a small experiment that sheds some light on this apparently paradoxical finding, and indicate how one might form a more powerful test for finding a slope-affecting gene. Several areas for future research are discussed. *Genet Epidemiol* 25 (Suppl. 1):S18–S28, 2003. © 2003 Wiley-Liss, Inc.

Key words: linkage; heritability; segregation; mixed models; correlation; variance components; Framingham Heart Study; simulation

Grant sponsor: NIH; Grant numbers: ES-10421, 5P30-ES07048.
*Correspondence to: W. James Gauderman, Ph.D., Department of Preventive Medicine, University of Southern California, 1540 Alcazar St., Suite 220, Los Angeles, CA 90089. E-mail: jimg@usc.edu
Published online in Wiley InterScience (www.interscience.wiley.com)
DOI: 10.1002/gepi.10280

INTRODUCTION

Longitudinal studies provide a valuable resource for investigating factors that affect long-term averages and changes over time in a complex trait. Statistical methods that assume independence across observations (e.g., standard linear or logistic regression) are not applicable to longitudinal data, due to the correlation among multiple measurements per subject. More advanced methods were developed to handle this intrasubject correlation [summarized in Diggle et al., 1995], including generalized estimating equations and hierarchical mixed models. These models have enjoyed wide application in epidemiological studies.

Family studies are a valuable resource for investigating genetic factors that influence an outcome. As with longitudinal data, standard statistical models will be inadequate due to the nonindependence in outcomes, in this case among related individuals. In fact, methods of genetic analysis rely on the correlation among family members' outcomes to infer genetic effects. Depending on the study goals and types of data available, the analyst will utilize methods appropriate for analysis of aggregation (e.g., heritability), segregation, linkage, and/or association.

Methods for each of these types of analysis have typically been developed assuming that only one outcome value has been measured on each subject.

The Framingham Heart Study (FHS) represents a marriage of longitudinal and family study designs. The FHS data provided to the Genetic Analysis Workshop 13 (GAW13) participants include repeated measurements of several clinical outcomes (e.g., blood pressure, cholesterol) on 2,885 individuals from 330 pedigrees. Recruitment occurred in two waves, producing two cohorts of individuals within the data set. The original cohort was initiated in 1948. Clinical measurements on this cohort's subjects were scheduled every 2 years to the present, yielding as many as 21 repeated observations on some subjects. The second cohort was initiated in 1971, and included the offspring of original-cohort members. Clinical measurements on these subjects were scheduled every 4 years, yielding up to five repeated observations per subject. The FHS has been a landmark study for advancing our understanding of factors, including diet and lifestyle, that affect coronary outcomes.

Attention recently focused on the analysis of genetic factors that influence coronary outcomes in this data set. Levy et al. [2000] performed a linkage analysis of systolic blood pressure (SBP), using a panel of 399 markers spaced across the genome. They found significant evidence of linkage to a region on chromosome 17, and suggestive linkage signals on chromosomes 5 and 10. In their analysis, Levy et al. [2000] first computed a person-specific residual SBP from a model that included age and other effects, and then utilized these residuals in the program SOLAR [Almasy and Blangero, 1998] to perform a variance-components linkage analysis. The residual used in the linkage analysis for a given subject represented their long-term average SBP, after adjustment for covariates. Their paper did not consider linkage analysis of change (slope) in SBP over time.

There is a relative paucity of methods for genetic analysis of longitudinal data in families. Contributors to GAW13 have developed a wide range of approaches to help fill this gap. Included in the Group 2 contributions are aggregation, segregation, linkage, and association analysis approaches to unraveling genetic effects on both long-term averages and changes over time. Methods were applied to the FHS and to similarly structured simulated data. This paper will describe and compare methods proposed by Group 2 contributors, summarize results of applications to FHS and simulated data, and synthesize the general lessons that were learned and issues that remain.

METHODS

OVERVIEW

Thirteen papers were contributed by Group 2 participants (Table I). Seven contributors applied their methods to the FHS data, with five focusing their primary analysis on SBP and two on body mass index (BMI). Six papers analyzed the simulated data, with four focusing on SBP and two on cholesterol. Additional traits were considered in some papers. All contributions except one included some form of linkage analysis. The analytic approaches are described in some detail below.

NOTATION

We let Y_{ij} denote the measurement of trait Y obtained on subject i at calendar time j, and let T_{ij} denote the corresponding age of the subject at that time. We let X denote one or more covariates, with subscripts included as necessary to indicate whether X represents time-dependent (e.g., BMI) or time-independent (e.g., sex) variables. The methods used by Group 2 contributors can be categorized into one of two general types: a two-step approach, or a joint model approach. These are described below.

TWO-STEP MODELS

Several contributors utilized a two-step approach, consisting of a longitudinal model in the first step, followed by a second-step linkage analysis of one or more statistics derived from the first-step model.

The first-step models had the general form

$$Y_{ij}=a_i+b_i\,T_{ij}+\gamma' X_{ij}+e_{ij} \quad (1)$$

where a_i and b_i are the subject-specific intercept and slope, respectively, and e_{ij} is a residual, assumed to be normally distributed with mean zero and variance σ^2. The slope b_i has the interpretation as the change in Y per increase of 1 year in age. The intercept a_i in this model can be interpreted as the mean of Y when T=0 (i.e., at birth) for a subject with all covariates X_{ij} equal to zero. Transformations to T or X (e.g., centering them on their means) are useful and will not affect b_i or e_{ij}, but will change the estimates and

TABLE I. Summary of data sets and analytic approaches used by Group 2

Data set	Lead author	Cohorts	Reps.	Trait[a]	Markers	Analysis approach[b]	Software[c]
Framingham	de Andrade	1 and 2	N/A	SBP	Ch. 17	L1: longitudinal VC linkage	ACT
	Barnholtz-Sloan	1	N/A	SBP	Ch. 10, 17	L1: linear mixed model; association	SAS
	Briollais	1 and 2	N/A	SPB	All	L2: linear mixed model, VC linkage	SAS/SOLAR
	Cheng	1 and 2	N/A	BMI	All	C2: linear model, VC linkage	SAS/SOLAR
	Gee	1 and 2	N/A	SBP	All	L2: linear model, segregation and linkage	SAS/GAP
	Macgregor	1 and 2	N/A	BMI	All	L1: heritability, VC linkage	SOLAR/ASREML
	Rao	2	N/A	SBP	Ch. 10	L2: principal components, HE linkage	SAS/SAGE
Simulated	Mirea	2	34	SBP	All selected	L2: linear mixed model, HE linkage	SAS/SAGE
						L1: multivariate HE linkage by GEE	SAS
	Scurrah	1 and 2	1	SBP	All	L2: linear mixed model, VC linkage	WinBUGS/Merlin
	Shephard	1 and 2	4, 10, 21	SBP	All	C2: heritability, VC linkage	Stata/SOLAR/GH
	Suh	1 and 2	10	SBP	Selected	L2: linear model, HE linkage	SAS
	Wang	1 and 2	All	Chol	Selected	L2: linear model, HE linkage	SAS/SAGE/GH
	Yang	1 and 2	8	Chol	All	L1: heritability, VC linkage	SOLAR/SAS

[a]Primary trait analyzed. In some contributions, additional traits were considered.
[b]L1, longitudinal one-step approach, with a single model that combines longitudinal and genetic analysis; L2, longitudinal two-step approach, with a first step longitudinal model and separate second step genetic analysis; C2, cross-sectional two-step approach, with a first step model of a selected time point and second step genetic analysis.
[c]GH, GENEHUNTER; GAP, Genetic Analysis Package; Ch, chromosome; Chol, cholesterol; NA, not applicable; see individual papers for descriptions of software programs and references.

interpretation of the a_i. The goal of the first-step analysis was to reduce the data to one observation per subject.

The second-step model was a genetic analysis of a person-specific statistic obtained from the first model. Since there was only one value per subject, standard modern genetic analyses were possible. These included analysis of heritability, segregation, model-free and model-based linkage, and association. For those conducting linkage analysis, most used either the variance- components (VC) approach described by Almasy and Blangero [1998] or the revised Haseman-Elston (HE) approach described by Elston et al. [2000].

Below is a brief summary of the specific approach used by each contributor of a two-step method, highlighting the differences and similarities among contributions.

Briollais et al. [2003]. This contribution expanded the first-step model in Equation (1) to include subject- and family-level models. Letting subscript f denote family, $\bar{T}_{fi}$ be the mean of observed ages for subject i, and $\bar{T}$ be the overall mean age in the sample, they used a three-level model of the form:

Level 1: $Y_{fij}=a_{fi}+b_{fi}\,(T_{fij}-\bar{T}_{fi})+c_{fi}\,(T_{fij}-\bar{T}_{fi})^2+\gamma' X_{fij}+e_{fij}$.

Intercepts

Level 2: $a_{fi}=a_f+\phi\,(\bar{T}_{fi}-\bar{T})+\eta' X_{fi}+e_{fi}$.

Level 3: $a_f=\alpha+e_f$.

Slopes

Level 2: $b_{fi}=b_f+\omega' X_{fi}+h_{fi}$.

Level 3: $b_f=\beta+h_f$.

The intercept and slope residuals *e* and *h* at each level were assumed to have mean zero and unstructured covariance matrix. The second step was a VC linkage analysis conducted on the adjusted mean, using the sum of intercept residuals $e_{fi}+e_f$, and on the slope, using sum of slope residuals $h_{fi}+h_f$. Analyses were conducted on SBP in the FHS data set.

Gee et al. [2003]. This paper utilized the first-step regression model shown in Equation 1, applied to analysis of SBP in the FHS. In addition to the intercepts and slopes, they also derived person-specific standard errors of the intercepts (s_{ai}) and slopes (s_{bi}) from the first-stage model. For a given subject, the magnitude of these standard errors was a function of the length of follow-up, the number and age distribution of measurements during follow-up, and the intraindividual variation in measurements over time. Subjects with longer follow-up tended to have lower estimated standard errors. The second step consisted of a formal segregation analysis of the intercepts and slopes, followed by parametric (LOD score) linkage analysis. The genetic analyses of the intercepts a_i (or slopes b_i) incorporated weights based on s_{ai} (or s_{bi}). Use of these weights allowed subjects with more precise first-step regression parameter estimates to contribute more

information to second-step segregation and linkage parameter estimates.

Mirea et al. [2003]. This paper evaluated the ability to detect linkage in a genome screen using HE analysis applied to several first-step statistics, including the first SBP, last SBP, mean SBP, time-adjusted change between first and last SBP, and linear regression slope of SBP on age. Phenotypic data on Cohort 2 subjects in one replicate of simulated data were utilized, with multiple sibships extracted from the pedigrees. An alternative joint-model analysis was also considered; this approach is described later.

Scurrah et al. [2003]. These authors extended earlier work on generalized linear mixed models [Scurrah et al., 2000] to the longitudinal data setting. The approach utilized a more complex first-step model than that shown in Equation (1), including parameters for polygenic, common family environment, and common sibling environment effects on both the intercepts and slopes. The Markov chain Monte Carlo technique of Gibbs sampling was utilized to fit this model. The subject-specific polygenic residuals for intercepts and slopes were derived from their first step and used in a VC linkage analysis. The method was applied to both cohorts in Replicate 1 of the simulated data.

Wang et al. [2003]. This paper utilized all replicates of the simulated data to perform an analysis of the power to detect linkage using a variety of first-step statistics. They analyzed cholesterol and considered first-visit level, mean level, and slope (the b_i values). Both two-point and multipoint linkage analyses were conducted. They analyzed markers near true trait-causing genes (to evaluate power), and markers on a chromosome not containing any trait-causing genes (to evaluate type I error rates).

Rao et al. [2003]. This contribution focused on Cohort 2 of the FHS and performed three different types of first-step models, each followed by a second-step HE linkage analysis. The first approach repeated the analysis of Levy et al. [2000] and thus focused the second-step genetic analysis on a measure of average SBP. The second approach utilized the model in Equation (1), focusing on slopes. The third was a principal components analysis, in which five separate components estimated in the first step were each utilized in the second-step linkage analysis. The first two components corresponded roughly to the overall mean and slope of SBP and explained most of the variation in the trait, while the remaining three components captured various nonlinear trends.

Shephard et al. [2003]. This contribution utilized the first-step model in Equation 1, but with the slope on age treated as a fixed effect. In other words, the subject-specific slopes b_i were replaced by a single slope parameter β common to all subjects. Subject-specific intercepts a_i were utilized in a second-step VC linkage analysis. Using simulated data, they compared the consistency of linkage results across three separate replicates, and also compared the results to results based on simply using the first-visit value.

Cheng et al. [2003]. This paper analyzed repeated cross-sectional data, and attempted to infer trends in genetic effects across age. Measurements of BMI obtained from FHS participants in 1970, 1978, and 1986 were utilized in three separate VC linkage analyses. The first-step model was analogous to that in Equation (1) without the person-specific slope (b_i) terms. These results were compared to similar analysis using the mean BMI from these three time points.

Suh et al. [2003]. This paper used a first-step model similar to that of Levy et al. [2000], and utilized residuals from this model in an HE linkage analysis. In the linkage analysis, mixed models were used to incorporate a range of correlation structures. In the simplest model, they assumed independence for each pair, as in the standard HE approach. As alternatives, they compared two types of correlation: correlation among sib pairs sharing a common individual, and correlation among all sibs within the same family. The method was applied to SBP in the simulated data.

Collectively, these contributions demonstrated many different approaches for reducing longitudinal data to obtain person-specific statistics for genetic analysis.

JOINT MODELS

In contrast to the two-step methods described above, the goal of these contributors was to simultaneously estimate genetic and longitudinal model parameters. A joint approach is appealing because estimates of genetic and longitudinal parameters will be mutually adjusted for one

another. Additionally, effects that cross models (e.g., interactions between genetic and longitudinal parameters) are more naturally included in a joint model framework. A current limitation of joint models is the increase in computational demands relative to a two-step approach, which can limit the types of analyses that can be considered. Following is a brief summary of the work by joint-model contributors, highlighting these issues in the context of their specific approach.

de Andrade and Olswold [2003]. This paper utilizes the method described by de Andrade et al. [2002], applied to SBP in the FHS. Their mean model had the form

$$Y=\alpha+\gamma'X+a+g+s+e.$$

Here the terms a, g, s, and e represent matrices of additive polygenic, additive major gene, shared-environment, and random-environment effects, respectively. The covariance between pairs of observations was specified using variances of these random effects, with specific contributions depending on the relationship between subjects and the time at which measurements were recorded.

For example, the covariance of observations from two relatives was modeled as a function of π, the observed proportion of alleles shared identical by descent (IBD) at some marker locus, and terms that depend on whether measurements were recorded at the same or different times. No structure with respect to age or calendar year was assumed for the covariance matrix. While such an unstructured covariance matrix is appealing, the number of variance/covariance terms to be estimated grows rapidly as the number of visits increases. Because of this, they restricted each analysis to two time points and focused on chromosome 17 markers.

Yang et al. [2003] and Macgregor et al. [2003]. These two contributions used very similar approaches and will be summarized together. Both papers focused on estimating age-specific heritability across predefined intervals of age. They attempted to solve the computational difficulties alluded to above by modeling the covariance matrix as a smooth function of age. The rationale was based on the fact that repeated observations of a trait are ordered in time, and thus one might expect the variances and covariances of proximal measures to be more similar than measures widely separated in time. Both groups assumed a trait model of the form

$$Y=\alpha+\gamma'X+(a_1+a_2T+a_3T^2+\ldots)+(g_1+g_2T+g_3T^2+\ldots)+(p_1+p_2T+p_3T^2+\ldots)+f+e.$$

The random effects a, g, and e are as described above, while p and f represent permanent environmental effects and time-constant family-specific effects, respectively. Legendre polynomials were used to model the a, g, and p random effects as a function of age. Yang et al. [2003] used a mixture of cubic and linear polynomials applied to age arranged into five age bands, averaging phenotypic and covariate values for subjects with more than one measurement within an age band. Macgregor et al. [2003] used linear polynomials applied to actual adult ages (i.e., 76 bands, one for each age between the ages 20–95). Both papers estimated age-specific total heritability and age-specific heritability attributable to a specific quantitative trait locus (QTL). For the latter, they performed preliminary linkage analysis using a two-step VC approach to identify linked markers. The covariance for the major gene effects (g) was then was modeled as a function of π at a given marker to estimate QTL-specific heritability. Yang et al. [2003] analyzed total cholesterol (TC) in the simulated data, while Macgregor et al. [2003] analyzed BMI, TC, HDLC, and height in the FHS.

Mirea et al. [2003]. In contrast to the above joint-model methods, the unit of analysis in this approach was the sib pair. Focusing on selected loci, they developed an HE-type joint linkage analysis of repeated longitudinal measurements and compared this to their two-step HE approach described above. The joint analysis involved using generalized estimating equations (GEE) to account for serial correlation in repeated measures of the sib-pair trait cross-product over time, ignoring residual correlation among sib pairs within the same family. An advantage of this approach was that, once IBD estimates were obtained, the analysis was possible using standard statistical software, and gene $\times$ time or gene $\times$ age interactions were easily incorporated.

Barnholtz-Sloan et al. [2003]. This contribution was unique as it did not perform linkage analysis, but rather focused on association analysis. A preliminary association analysis was conducted using the binary trait ''high SBP,'' defined as SBP above 140 on two consecutive visits, or reported use of hypertension treatment. A genome scan in the FHS revealed three markers showing

association to this trait. These markers were then analyzed in a joint model for SBP (in its continuous form), using mixed linear regression with random effects to account for family, sibship, and repeated measures.

RESULTS

FRAMINGHAM DATA

Not surprisingly, the various analytic approaches produced many different types of results. Rather than cover each result in detail, we summarize some of the key findings and focus on comparisons/contrasts among findings. We refer to specific marker loci by their chromosome and location in centimorgans (cM), rather than using their specific locus names.

There was much interest in chromosome 17 for SBP, given the LOD score of 4.7 (at 67 cM) observed previously by Levy et al. [2000]. de Andrade and Olswold [2002] were unable to detect any significant linkage to markers on chromosome 17 using their longitudinal VC approach. However, they also repeated the analysis of Levy et al. [2000] on these GAW data, and found a LOD score of 3.0 at position 68 cM on chromosome 17, but only when the sample was restricted to ages 25–75. Briollais et al. [2003] found evidence of linkage on chromosome 17 (62 cM), using intercept residuals in both an unselected (LOD=2.1) and selected (LOD=3.5) sample. Gee et al. [2003] did not find evidence of linkage in this specific region, but reported a modest linkage signal for intercepts to chromosome 17 (100 cM, LOD=1.5).

Evidence of genes on other chromosomes was also detected for those who analyzed SBP. Using first-step model intercepts, Gee et al. [2003] found LOD scores above 2.0 on chromosomes 1 (202 cM and 212 cM), 9 (32 cM), and 10 (125 cM). Their LOD scores were generally larger in analyses that utilized weights based on first-step standard errors, compared with not using weights. Rao et al. [2003] also found linkage evidence at position 125 cM on chromosome 10, using either mean SBP, principal components, or selected cross-sectional observations. Interestingly, Barnholtz-Sloan et al. [2003] found evidence of association (P=0.02) to a marker in this region of chromosome 10 (at 135 cM). Briollais et al. [2003] did not find linkage evidence to any marker on chromosome 10, but did report LOD scores above 2.0 for intercept residuals on chromosomes 2 (38 cM), 3 (79 cM), 8 (37 cM), and 13 (64 cM). This was the only group to also find linkage support for genes that affect SBP slope, on chromosomes 1 (212 cM), 3 (153 cM), and 11 (33 cM). Briollais et al. [2003] reported that the magnitude of their LOD scores at all these markers was quite sensitive to whether they adjusted for BMI in their first-step model. They obtained lower LOD scores in models that did not include BMI.

In two-step analyses of cross-sectional BMI observations, linkage to markers on chromosome 16 was detected by both Cheng et al. [2003] (75 cM, LOD=2.4) and Macgregor et al. [2003] (95 cM, LOD=3.1). Based on subsequent joint-model analysis, Macgregor et al. [2003] reported that the heritability attributable to a gene linked to this 95-cM marker varied substantially across the age range. Specifically, they estimated that 25% of the total variation in BMI could be attributed to this locus at age 20, but this declined to less than 5% for ages greater than 60. On the other hand, they found that a locus linked to total cholesterol (chromosome 20, 24 cM) accounted for a large proportion of variation in cholesterol across all age intervals. Cheng et al. [2003] also found linkage evidence for BMI on chromosomes 3 (181 cM), 6 (146 cM), and 9 (88 cM).

In summary, there was some agreement for genes affecting SBP on chromosomes 1, 10, and 17, and for a gene affecting BMI on chromosome 16. Linkage signals were generally higher for level-type statistics (intercepts, means, and intercept residuals), and most contributors found no evidence for genes affecting slopes. Two questions are suggested from analyses of the FHS data: 1) When are longitudinal data superior to cross-sectional data for genetic analysis? and 2) Do we have adequate power to detect slope genes? With these questions in mind, we turn to results from analyses of the simulated data.

SIMULATED DATA

There were six genes simulated to have direct effects on SBP, three with effects on baseline SBP (b34–b36), and three on slope over age (s10–s12). Slope genes s10 and s12 were simulated to be on the same chromosome.

Performing their analysis without knowledge of the answers, Mirea et al. [2003], Scurrah et al. [2003], and Shephard et al. [2003] were all able to successfully detect some of these genes by linkage analysis, each using different first-step approaches to modeling the longitudinal data. The

performance of the various methods cannot be directly compared, because each paper analyzed different replicates of the simulated data. However, some interesting trends emerged across these contributions with respect to the types of first-step statistics that showed the most significant linkage evidence.

Mirea et al. [2003] found that linkage evidence for baseline genes b34 and b35 was much more significant using visit 1 SBP than using last SBP, mean SBP, or slope of SBP. This is not surprising, given that these genes were simulated to have their effect early in follow-up. What was surprising in their results, however, was that all three slope genes were detected with greater significance using a first-step level-type statistic (e.g., mean SBP or last visit SBP) than by using a first-step slope statistic. Scurrah et al. [2003] reported analogous results. Their most significant linkage evidence was for a marker near slope genes s10 and s12, but the LOD score for this locus was much greater using a first-step intercept residual (LOD=12.9) than using a first-step slope residual (LOD=5.1). They also found suggestive linkage evidence for a marker near slope gene s11, here again using their intercept rather than slope statistic. Shephard et al. [2003] also found strong evidence of linkage near s10 and s12, using the intercepts from their first-step longitudinal model. They reported greater LOD scores using longitudinal data in the first-step model, compared to simply using first-visit SBP, even for detecting baseline genes. In analyses conducted unblinded to the answers, Suh et al. [2003] were also able to detect linkage to slope genes using level-type statistics.

Two contributors analyzed total cholesterol, which was simulated to depend on four baseline genes (b30–b33) and three slope genes (s7–s9). Without knowledge of the answers, Yang et al. [2003] were able to detect linkage to b30, b31, and b32 using visit 1 cholesterol. They were also able to detect slope gene s7, with a slightly higher LOD score using first-step mean (LOD=10.6) than first-step slope (LOD=10.3). In their joint model analysis, they found that heritability was relatively flat across age for baseline genes b30 and b32, but showed a marked increasing trend with age for s7. Wang et al. [2003] were the only contributors to analyze all 100 replicates in a true simulation study. They reported greater power for detecting the baseline genes using exam 1 cholesterol, compared to using mean or slope of cholesterol. They reported the greatest power for detecting slope gene s7 using first-step slope (80%), although first-step mean also provided reasonable power (62%) for detecting this gene. Power was low with any statistic to detect slope genes s8 and s9. Wang et al. [2003] also analyzed several unlinked markers and reported acceptable type I error rates.

Collectively, these simulated-data contributions shed some light on the questions raised by the FHS analyses. Well-selected cross-sectional data (e.g., first or last visit) provided good power for detecting some genes. However, summaries of longitudinal data (e.g., means, slopes) were generally most effective for finding genes, particularly those that affected trends in outcome over time. Somewhat paradoxical was the general finding that level-type statistics (e.g., intercept, mean) provided greater power for detecting slope genes than did slope-type statistics. We now explore this finding further.

A SMALL EXPERIMENT

We performed a small experiment to investigate the use of intercept and slope statistics for detecting a slope gene. We simulated a sample of 1,000 independent individuals. Each individual was randomly assigned a genotype (G) at a slope-affecting locus, with probability 50% each of carrying a normal (G=0) or variant (G=1) genotype. Age (T) was also randomly generated for each subject from a uniform distribution on the range 0–50 years. Conditional on G and T, the trait Y was randomly sampled from a normal distribution with mean $100+\beta^*G^*T$ and variance σ^2. It is clear that under this model, the gene G has no baseline effect, but rather only affects slope.

Can we detect this slope gene with more power using a test based on slope or mean statistics? We first fit a linear model of the form

$$Y=\alpha+\beta_1 G+\beta_2 T+\beta_3 G^*T+e \quad (2)$$

where e was assumed to be normally distributed with mean 0 and variance α^2. The parameter β_3 quantifies the difference in slopes between G=0 and G=1, and the estimated slope β_3 can be used to form a slope-based test of the form $t=\beta_3/\mathrm{se}(\beta_3)$. We then considered a model of the form

$$Y=\alpha+\beta_1 G+e$$

where the parameter β_1 measures the difference in mean Y between genotype groups, with corresponding mean-based test $t=\beta_1/\mathrm{se}(\beta_1)$.

For three different settings of σ (1, 8, and 32, respectively), Figure 1 plots simulated Y vs. T

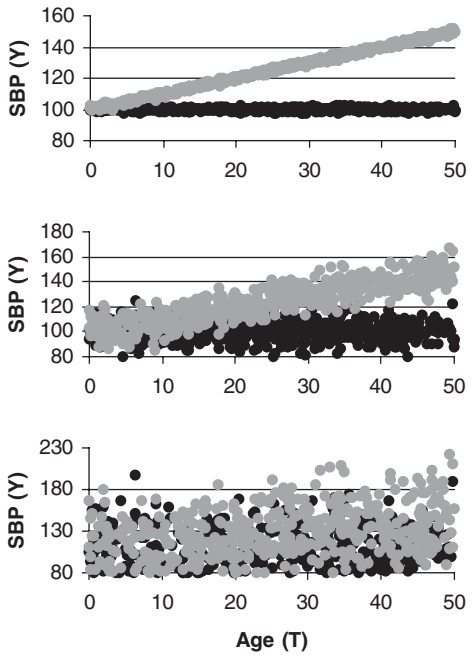

Fig. 1. Simulated SBP, based on model in equation (2), assuming β=1, with σ=1.0 (top), σ=8.0 (center), or σ=32 (bottom). Gray dots have variant genotype (G=1); black dots have G=0.

by genotype G for 1,000 subjects, when β is set to 1.0. The difference in slope of Y on T between G=0 and G=1 is clear when σ=1, but becomes less obvious as σ increases.

Table II gives the expected *t*-statistic for the slope and mean-based tests when the true β=1.0 (as in Fig. 1), and for a larger slope effect (β=3.0). When σ=1, the *t*-statistic (and thus power) is much larger for the slope test than for the mean test. However, as the variance increases, the power of the slope test is dramatically reduced, while the power of the mean test is much less affected. When σ=32, the mean test is more powerful than the slope test both for β=1 and β=3.

The conclusion one can draw from this experiment is that when the residual variance is large, as it is for traits in both the FHS and simulated data, a test based on means can provide greater power to detect a slope-affecting gene than a test based on slopes alone. In practice, many additional factors will determine the relative power of a mean-based to slope-based test, including not only the underlying true effect sizes, but the number of repeated observations and the length of follow-up. Also important will be the relative magnitude of the within- and between-subjects variance of Y.

TABLE II. Expected *t*-statistics for mean- and slope-based tests

		Expected *t*-statistics	
σ	β	Mean test	Slope test
1	1	38.2	229.3
	3	38.4	687.7
8	1	29.9	28.7
	3	37.0	86.0
32	1	11.0	7.2
	3	26.2	21.5

DISCUSSION

Complex traits such as SBP and cholesterol vary with age and likely depend on both genetic and environmental determinants. For such traits, longitudinal data allow one to disentangle genetic and environmental effects, both on the rate of change of the phenotype over time (e.g., slope) and on trait level (e.g., mean). Unlike the FHS, most family studies collect a single cross-sectional measurement on each subject. While this type of data can also be used to analyze mean and slope effects, estimates will be more prone to bias from confounding and more affected by measurement error.

How does the current value of an age-dependent trait depend on genotype? For simplicity, consider two groups of subjects, carriers (C) and noncarriers (N), respectively, of a variant allele at a particular locus. Differences in expected trait value between C and N groups at age T will be a function of their difference at birth plus any difference that accrues between birth and T. There are four possible scenarios: 1) G has no affect at birth or thereafter, 2) G only affects level at birth, 3) G has no affect at birth, but affects development, and 4) G affects both level at birth and development. Without knowing the truth, the analyst is faced with choosing the test statistic that provides the greatest power to detect G.

Although the best statistic to choose will depend on the true situation, these GAW contributions shed some light on the relative robustness of different alternatives. Obviously, any statistic needs to have the correct test size when situation 1 holds. For situation 2, equivalent to a baseline gene in the GAW simulation, only level-type statistics (e.g., mean or cross-sectional value)

provide power. This makes sense, since there is no difference between the C and N groups in slope. When the gene does affect slope (situation 3 or 4), statistics based on slope or change in level over time can be used. However, several contributions and the small experiment indicated that a mean-based statistic can often provide greater power for finding a slope gene than a slope-based statistic.

The reason that a mean-based statistic has any power to detect a slope gene is that a slope gene will typically lead to a difference in the mean of the trait by genotype. This can be seen in Figure 1, for example, where the difference in means is approximately the difference in genotype-specific linear predictions at the midpoint of age (T=25). A notable exception will occur if genotype-specific baseline means are different *and* one slope is positive while the other is negative (graphically, an X-shape rather than the sideways V-shape shown in Fig. 1). However, such an X-shaped relationship is unlikely for most biological systems.

When a slope gene does affect both slope and mean, neither the mean- nor slope-based statistics used by many contributors will be optimal for finding genes that affect rate of change. On the basis of the model in Equation (2), the null hypothesis we should be interested in for such a slope gene is $\beta_1=0$ (no level effect) *and* $\beta_3=0$ (no slope effect). A two-degree-of-freedom likelihood ratio test comparing the likelihood at the joint maximum likelihood estimate (MLE) of β_1 and β_3 to the likelihood with both fixed to zero would be appropriate. This type of test is analogous to previously proposed joint tests in the context of using gene $\times$ covariate interaction information to improve power for detecting linkage [Greenwood and Bull, 1999; Olson, 1999; Gauderman and Siegmund, 2000; Gauderman et al., 2001]. In fact, one can think of the slope parameter β_3 as a measure of gene $\times$ covariate interaction, in this case with age being the covariate.

Careful consideration of covariates will be essential for understanding both environmental and genetic (through G $\times$ E interaction) determinants of complex traits. The current value of an age-dependent trait will likely depend on both current and previous values of environmental covariates. There are several ways covariate information can be included in a model. One can include time-varying covariate values, e.g., smoking status at each visit, directly into a multilevel or joint model. An alternative approach is to incorporate cumulative exposure through a single covariate, e.g., total number of pack years of smoking. One may choose to focus on exposure during a critical period of life, e.g., in utero or early-life exposure to parental smoking. More complicated covariates can also be constructed, e.g., allowing current covariate effects to be modified by previous exposure levels or by genotype. Of course, all these methods depend on the availability of reliable covariate data, which is more likely to derive from longitudinal rather than cross-sectional studies.

In terms of modeling approaches, contributors to this group adopted either a two-step or joint model for the genetic analysis of longitudinal data. In general, a joint model should be preferable for two main reasons. First, parameter estimates in the longitudinal and genetic models are mutually adjusted for one another. Second, a joint model correctly accounts for within-individual and between-individual variability, so that uncertainty in the estimated phenotype (e.g., person-specific intercept or slope) is accounted for during the linkage analysis. While one can weight first-step summary statistics to account for the relative degree of within- and between-subject variance [Gee et al., 2003], such weighting comes about naturally in a joint model.

While a joint modeling approach has theoretical advantages, the two-step approach is attractive for practical (computational) reasons. First-step longitudinal models can be fit using standard statistical software packages (e.g., SAS, SPLUS, and STATA). Once subject-specific summary statistics are abstracted from this first step, a number of available programs can be used for linkage, heritability, or segregation analysis. Commonly used genetic software programs are not designed for longitudinal data analysis, and there is a clear need to develop integrated programs. Regardless of whether a two-step or joint approach is adopted, the analyst should always carefully consider model assumptions, e.g., normality and homoscedascity, since violations can lead to invalid conclusions.

Multilevel modeling, which can take into account the hierarchical structure of the data, may help disentangle the proportion of the trait variability explained by fundamental variation in the mean trait and in the trait slope from the proportion explained by random within-individual variability. Joint modeling in the multilevel model framework is theoretically possible. As an example, the multilevel model of Briollais et al.

[2003] can be expressed as a single mixed model, with the form

$$Y_{fij}=\alpha+\phi(\bar{T}_{fi}-\bar{T})+\beta(T_{fij}-\bar{T}_{fi})+\gamma'X_{fij}+\eta'X_{fi}+\omega'X_{fi} \times (T_{fij}-\bar{T}_{fi})+e_f+e_{fi}+e_{fij}+(h_f+h_{fi})(T_{fij}-\bar{T}_{fi}).$$

This model is easily extended to include additional levels (e.g., sibships within family), with corresponding covariates and random effects. The variance-covariance matrix of the random effects (*e* and *h* values) can be expressed as a function of marker-IBD sharing probabilities among relatives, thus facilitating a test of linkage on intercepts and/or slopes. One could also include a marker genotype as a covariate in the above model, thus also providing tests and estimates of association on trait level and/or slope. This type of model generalizes the hierarchical modeling structure described by Fulker et al. [1999] in the context of cross-sectional data.

In population studies of blood pressure, a significant proportion of blood pressure observations will be affected by hypertensive treatment (HRX). Levy et al. [2000] reported that 15.3% of observations reflected HRX in the FHS. In such observations, measured SBP will be lower than the ''true'' untreated SBP, which will impact estimates of genetic and environmental effects. Members of this group utilized various methods of accounting for this problem. These include ignoring the problem completely, excluding individuals on treatment, including HRX as a covariate, replacing the phenotypes of all individuals on HRX with a single high value, adding a constant (an average HRX effect) to observations on treatment, and imputing post-HRX SBP based on pre-HRX measurements and/or or the SBP of other family members. Some of these approaches will produce biased results, and the extent of the bias is likely to depend on the proportion of individuals on treatment and the actual effects of treatment on those individuals. The advantages and disadvantages of each approach will not be discussed here, and we do not aim to recommend a single best approach, as the problem is still being researched [e.g., Cui et al., 2003]. However, the results of any linkage analysis for such phenotypes will depend on the way in which treatment has been accounted for, and it is an issue that should be considered in population-based studies such as the FHS.

Another important issue in longitudinal studies is that of missing data. All of the contributions in this group ignored the problem of missing data, focusing their analyses on observations with complete outcome and covariate data. It is well-known that the elimination of missing observations can lead to bias if data are not missing completely at random (MCAR), and particularly if there is informative missingness [Little and Rubin, 2002]. An example of informative missingness is cohort dilution, e.g., the elimination of subjects at later ages from the cohort in a nonrandom way with respect to trait genotype. In some situations, one may need to specify a joint model of both the phenotype and the missingness process. This type of analysis was used to model survival and quality-of-life data in cancer patients, when quality of life was not missing at random [Billingham et al., 2001]. Some approaches to dealing with missing data in the FHS have been developed [Badzioch et al., 2003], but this important topic needs further statistical attention.

Understanding the magnitude of within- and between-subject variability in a trait is important in designing a longitudinal family study. When intrasubject variability in a trait is high (as was observed for SBP in the FHS), precision will be increased by having many repeated measurements per subject. On the other hand, when intrasubject variability is low, power will be greater by increasing the number of individuals rather than by increasing the number of measurements. This adds a level of complexity to the design of family studies, for which one also has to consider within- and between-family trait variability. In addition, practical considerations (e.g., stability of the population over time, cost of obtaining measurements) will play heavily in the design of a longitudinal family study.

In conclusion, this group proposed, applied, and evaluated several approaches to the analysis of longitudinal family data. Collectively, our findings confirmed some of those previously reported by Levy et al. [2000], and indicated some additional chromosomal locations that may warrant further investigation. From a methodological standpoint, we described several variations of two-step and joint modeling approaches. Across many different approaches, we found that the use of a mean-based statistic is likely to provide more power for detecting a slope-affecting gene than a slope-based statistic. This finding warrants further study. Also an important topic for future research is the development of models that integrate the estimation of genetic and longitudinal parameters, along with associated software for fitting the models. We encourage readers to see the individual contributions to learn more about each specific method.

REFERENCES

Almasy L, Blangero J. 1998. Multipoint quantitative-trait linkage analysis in general pedigrees. Am J Hum Genet 62:1198–1211.

Badzioch MD, Thomas DC, Jarvik GP. 2003. Summary report: missing data and pedigree and genotyping errors. Genet Epidemiol 25 (Suppl. 1):S36–S42 (this issue).

Barnholtz-Sloan JS, Poisson LM, Coon SW, Chase GA, Rybicki BA. 2003. Analysis of gene × environment interaction in sibships using mixed models. BMC Genet [Suppl] 4:18.

Billingham LJ, Abrams KR, Jones DR. 2001. Simultaneous assessment of quality of life and survival data. In: Stevens A, Abrams KR, Brazier JE, Fitzpatrick R, Lilford R, editors. Advanced handbook of methods in evidence based healthcare. London: Sage Publications. Chap 20, p 352–366.

Briollais L, Tzontcheva A, Bull S. 2003. Multilevel modeling for the analysis of longitudinal blood pressure data in the Framingham Heart Study pedigrees. BMC Genet [Suppl] 4:19.

Cheng R, Park N, Hodge SE, Juo S-HH. 2003. Comparison of the linkage results of two phenotypic constructs from longitudinal data in the Framingham Heart Study: analyses on data measured at three time points and on the average of three measurements. BMC Genet [Suppl] 4:20.

Cui JS, Hopper JL, Harrap SB. 2003. Antihypertensive treatments obscure familial contributions to blood pressure variation. Hypertension 41:207–210.

de Andrade M, Olswold C. 2003. Comparison of longitudinal variance components and regression-based approach for linkage detection on chromosome 17 for systolic blood pressure. BMC Genet [Suppl] 4:17.

de Andrade M, Gueguen R, Visvikis S, Sass C, Siest C, Amos CI. 2002. Extension of variance components approach to incorporate temporal trends and longitudinal pedigree data analysis. Genet Epidemiol 22:221–232.

Diggle PJ, Liang KY, Zeger SL. 1995. Analysis of longitudinal data. Oxford: Clarendon Press.

Elston RC, Buxbaum S, Jacobs KB, Olson JM. 2000. Haseman and Elston revisited. Genet Epidemiol 19:1–17.

Fulker D, Cherny S, Sham P, Hewitt J. 1999. Combined linkage and association sib-pair analysis for quantitative traits. Am J Hum Genet 64:259–267.

Gauderman W, Siegmund K. 2000. Gene-environment interaction and affected sib pair linkage analysis. Hum Hered 52:34–46.

Gauderman W, Morrison J, Siegmund K. 2001. Should we consider gene × environment interaction in the hunt for quantitative trait loci? Genet Epidemiol 21: 831–836.

Gee C, Morrison JL, Thomas DC, Gauderman WJ. 2003. Segregation and linkage analysis for longitudinal measurements of a quantitative trait. BMC Genet [Suppl] 4:21.

Greenwood CMT, Bull SB. 1999. Analysis of affected sib pairs, with covariates—with and without constraints. Am J Hum Genet 64:871–885.

Levy D, DeStefano AL, Larson MG, O'Donnell CJ, Lifton RP, Gavras H, Cupples LA, Myers RH. 2000. Evidence for a gene influencing blood pressure on chromosome 17. Genome scan linkage results for longitudinal blood pressure phenotypes in subjects from the Framingham Heart Study. Hypertension 36:477–483.

Little RJA, Rubin DB. 2002. Statistical analysis with missing data. New York: John Wiley & Sons, Inc.

Macgregor S, Knott S, White I, Visscher P. 2003. Longitudinal variance-components analysis of the Framingham Heart Study data. BMC Genet [Suppl] 4:22.

Mirea L, Bull SB, Stafford J. 2003. Comparison of Haseman-Elston regression analyses using single, summary, and longitudinal measures of systolic blood pressure. BMC Genet [Suppl] 4:23.

Olson JM. 1999. A general conditional-logistic model for affected-relative-pair linkage. Am J Hum Genet 65:1760–1769.

Rao S, Li L, Li X, Moser KL, Guo Z, Shen G, Cannata R, Zirzow E, Topol EJ, Wang Q. 2003. Genetic linkage analysis of longitudinal hypertension phenotypes using three summary measures. BMC Genet [Suppl] 4:24.

Scurrah KJ, Palmer L, Burton P. 2000. Variance components analysis for pedigree-based censored survival data using general linear mixed models (GLMMs) and Gibbs sampling in BUGS. Genet Epidemiol 19:127–148.

Scurrah KJ, Tobin MD, Burton PR. 2003. Longitudinal variance-components models for systolic blood pressure, fitted using Gibbs sampling. BMC Genet [Suppl] 4:25.

Shephard N, Falcaro M, Zeggini E, Chapman P, Hinks A, Barton A, Worthington J, Pickles A, John S. 2003. Linkage analysis of cross-sectional and longitudinally derived phenotypic measures to identify loci influencing blood pressure. BMC Genet [Suppl] 4:26.

Suh YJ, Park T, Cheong SY. 2003. Linkage analysis of longitudinal data. BMC Genet [Suppl] 4:27.

Wang D, Li X, Lin Y-C, Yang K, Guo X, Yang H. 2003. Power of linkage analysis using traits generated from the simulated longitudinal data of the Framingham Heart Study. BMC Genet [Suppl] 4:28.

Yang Q, Chazaro I, Cui J, Guo C-Y, Demissie S, Larson M, Atwood LD, Cupples LA, DeStefano AL. 2003. Genetic analyses of longitudinal phenotype data: a comparison of univariate methods and a multivariate approach. BMC Genet [Suppl] 4:29.