



Comparing Analytic Methods for Longitudinal GWAS and a Case-Study Evaluating Chemotherapy Course Length in Pediatric AML. A Report from the Children's Oncology Group

Marijana Vujkovic^{1*}, Richard Aplenc¹, Todd A. Alonzo², Alan S. Gamis³ and Yimei Li¹

¹ Division of Oncology, Children's Hospital of Philadelphia, Philadelphia, PA, USA, ² Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA, ³ Division of Hematology, Oncology Bone Marrow Transplantation, Children's Mercy Hospitals and Clinics, Kansas City, MO, USA

OPEN ACCESS

Edited by:

Jill Barnholtz-Sloan,
Case Western Reserve University,
USA

Reviewed by:

Peter A. Kanetsky,
Moffitt Cancer Center, USA
Fredrick Schumacher,
Case Western Reserve University,
USA

*Correspondence:

Marijana Vujkovic
vujkovicm@email.chop.edu

Specialty section:

This article was submitted to
Applied Genetic Epidemiology,
a section of the journal
Frontiers in Genetics

Received: 11 April 2016

Accepted: 19 July 2016

Published: 05 August 2016

Citation:

Vujkovic M, Aplenc R, Alonzo TA,
Gamis AS and Li Y (2016) Comparing
Analytic Methods for Longitudinal
GWAS and a Case-Study Evaluating
Chemotherapy Course Length in
Pediatric AML. A Report from the
Children's Oncology Group.
Front. Genet. 7:139.
doi: 10.3389/fgene.2016.00139

Regression analysis is commonly used in genome-wide association studies (GWAS) to test genotype-phenotype associations but restricts the phenotype to a single observation for each individual. There is an increasing need for analytic methods for longitudinally collected phenotype data. Several methods have been proposed to perform longitudinal GWAS for family-based studies but few methods are described for unrelated populations. We compared the performance of three statistical approaches for longitudinal GWAS in unrelated subjects: (1) principal component-based generalized estimating equations (PC-GEE); (2) principal component-based linear mixed effects model (PC-LMEM); (3) kinship coefficient matrix-based linear mixed effects model (KIN-LMEM), in a study of single-nucleotide polymorphisms (SNPs) on the duration of 4 courses of chemotherapy in 624 unrelated children with *de novo* acute myeloid leukemia (AML) genotyped on the Illumina 2.5 M OmniQuad from the COG studies AAML0531 and AAML1031. In this study we observed an exaggerated type I error with PC-GEE in SNPs with minor allele frequencies <0.05, whereas KIN-LMEM produces more than expected type II errors. PC-LMEM showed balanced type I and type II errors for the observed vs. expected *P*-values in comparison to competing approaches. In general, a strong concordance was observed between the *P*-values with the different approaches, in particular among $P < 0.01$ where the between-method AUCs exceed 99%. PC-LMEM accounts for genetic relatedness and correlations among repeated phenotype measures, shows minimal genome-wide inflation of type I errors, and yields high power. We therefore recommend PC-LMEM as a robust analytic approach for GWAS of longitudinal data in unrelated populations.

Keywords: longitudinal analysis, unrelated population, genome wide association analysis, linear mixed effects model, generalized estimating equations

INTRODUCTION

The development of high-throughput DNA genotyping has accelerated the discovery of single nucleotide polymorphisms (SNPs) associated with common traits and diseases (Welter et al., 2014). Genome-wide analysis (GWAS) accounts for multiple testing and underlying genetic structures but does not capture the natural trajectory of phenotypic traits over time (Nyholt, 2004; Weir et al., 2006). Methodologic challenges for longitudinal GWAS include correlation among phenotypic measurements within an individual and missingness of phenotypic measures over time while accounting for phenomena like genetic relatedness, population stratification, and the influence of stationary or time-varying covariates.

The GWAS Catalog has curated over 2000 cross-sectional studies linked to hundreds of diseases and traits but few longitudinal modeling strategies have been evaluated to date (Kerner et al., 2009; Beyene and Hamid, 2014; Welter et al., 2014). A longitudinal approach to GWAS would be of particular benefit in studying dynamic quantitative traits related to chronic diseases, such as bone mineral density, fasting glucose levels, LDL, and HDL cholesterol, triglycerides, arterial stiffness, and systolic and diastolic blood pressure. When investigating genetic variants on a phenotypic outcome type I errors may occur due to population structure in unrelated populations and/or complex pedigree structure among participants in family based studies. The majority of approaches for the analysis of longitudinal data have emphasized family-based studies in large part in response to various Genetic Analysis Workshops calls for analytic approaches to deal with longitudinal phenotypes in the family-based Framingham Heart Study (FHS). Accounting for pedigree structure usually involves polygenic models incorporating between-individual kinship coefficients in the covariance structure for the random effect in generalized linear models. However, analytic methods for longitudinal GWAS in unrelated individuals are relatively sparse. Furlotte et al. proposed a longitudinal GWAS design that explains phenotypic temporal trends and population structure simultaneously using a kinship coefficient matrix-based linear mixed effects model (KIN-LMEM) (Furlotte et al., 2012). However, it is unconventional to account for population stratification using kinship information in unrelated individuals with exception of a few reports (Kang et al., 2010; Zhang et al., 2010). Typically a GWAS accounts for population-specific variations in allele distribution of SNPs (e.g., population stratification) by including principal components (PCs) as covariates in a linear or logistic regression model (Patterson et al., 2006). Recently, Sikorska et al. used a fast conditional two-step approach based on fitting a linear mixed effects model (LMEM) followed by linear regression as a computationally efficient workaround for LMEM with random intercept and random slope (Sikorska et al., 2015). However, this model overcomplicates scenarios where random slopes are unnecessary.

The field of applied bioinformatics heavily relies on easily reproducible or ready-to-go methods due to restrictions of time, resources, and method complexity. Given the current knowledge gap in longitudinal methods for GWAS in unrelated

populations, we present a brief overview of available literature on longitudinal GWAS approaches and compare PC and kinship-based methods to evaluate chemotherapy course length in 2 randomized phase III trials in childhood acute myeloid leukemia (AML). These include: principal component-based generalized estimating equations (PC-GEE), principal component-based linear mixed effects model (PC-LMEM), and kinship coefficient matrix-based linear mixed effects model (KIN-LMEM).

MATERIALS AND METHODS

Study Design

A total of 624 Caucasian patients were included from two randomized phase III trials of the Children's Oncology Group (COG), AAML0531 (Gamis et al., 2014), and AAML1031 (Meshinchi et al., 2012). These trials enrolled children with newly diagnosed AML using standard chemotherapy in combination with investigations agents. DNA was extracted from remission bone marrow aspirates and genotyped on the Illumina 2.5M OmniQuad at the Children's Hospital of Philadelphia, PA. The National Cancer Institute's central institutional review board and institutional review boards at each enrolling center approved both studies; patients and their families provided informed consent or assent as appropriate. The trials were conducted in accordance with the Declaration of Helsinki and registered at <http://www.clinicaltrials.gov> as NCT00372593 (AAML0531) and NCT01371981 (AAML1031).

Mini Review

We aimed to identify all published literature that focused on longitudinal genome-wide association studies (GWAS). Existing publications were identified in PubMed through January of 2016 using the following search terms: "longitudinal," "repeated measures," "linear mixed models," "GWAS," "genome-wide associations," and/or "SNP." Publications not related to longitudinal GWAS were disregarded.

Phenotype

The primary outcome in the current case-study is chemotherapy course length. Start and end dates of the first 4 courses of chemotherapy were extracted from the COG web portal. Course length was defined as the difference in days between start and end date of each course. The start date was the day the chemotherapy regimen was initiated, and end date was considered as the day that absolute neutrophil counts reached 500/mL and platelets recovered to 20,000/mL in the absence of recurrent or persistent leukemia. The 5th course of chemotherapy on AAML-0531 was excluded from analysis to ensure comparability between the two studies. For quality control purposes course duration was set to missing if the number of course days was <20 or >100, or a relapse was observed in the respective course. Covariates collected include age, gender, and treatment arm.

Three Longitudinal Models

In longitudinal analyses the outcome is a set of repeated measures of course duration over time. We use Y_{ij} ($i = 1, \dots, n; j = 1, 2, 3, 4$) to denote the log-transformed course duration for

TABLE 1 | Literature overview of longitudinal GWAS analysis.

References	Design	Outcome	Covariates	Familial relatedness	Missingness	Analysis	Software
Sikorska et al., 2015	Family	Linear	Time-varying	NA	MAR/MCAR	Reduced conditional LMEM + GWAS on random slopes	R
Choi et al., 2014	Family	Binary	Stationary	Random effects	MAR	GEE, family-specific GLMEM	PLINK, SAS, R
Eu-Ahsunthornwattana et al., 2014	Family, unrelated	Linear	Stationary	Variance-covariance structure	MAR	PHENO LIN REG, GWAS on residuals	PLINK, EMMAX, FaST-LMEM, GenABEL
Hossain and Beyene, 2014	Family	Linear	Stationary	Variance-covariance structure	MAR	LMEM	R
Musolf et al., 2014	Family, unrelated	Linear	Time-varying	(Q)TDT	NA	Cluster analysis, GWAS	SAS, TDT-HET
Tan et al., 2014	Family	Linear	Stationary	Kinship matrix	NA	Two-level hierarchical linear model with random coefficients	
Vaitsiakhovich et al., 2014	Unrelated	Delta	Stationary	NA	MAR, MCAR, MNAR + imputation	GWAS on mean change of imputed data	INTERSNP
Wang et al., 2014	Family	Linear	Stationary	NA	MAR + imputation	LMEM and a two-stage approach: random intercept model + GWAS	R, PLINK
Xia and Lin, 2014	Family	Binary	Stationary	Inbreeding coefficient in Bayesian	MAR	Logistic Bayesian LASSO, B-spline, partial GLMEM	R, Hapassoc
Furlotte et al., 2012	Unrelated	Linear	NA	Variance-covariance structure	NA	Modified LMEM	R
Chang et al., 2009	NA	Binary	Time-varying	Random effects	NA	GMM	SAS
Fradin and Fallin, 2009	Family	Binary	Stationary	NA	NA	LOG REG, conditional on risk set	SAS
Kerner et al., 2009	Family	Linear	Stationary	Random effects	NA	GMM, QTL analysis	Mplus, Goldenhelix
Luan et al., 2009	Family	Linear	Stationary	Random effects	NA	LMEM	Stata, SAS
Park et al., 2009	Family	Binary	Stationary	Pedigree membership as covariate	NA	GEE	SAS
Roslin et al., 2009	Unrelated	Linear	Stationary	NA	NA	MLGM, LIN REG	Mplus, PLINK
Yan et al., 2009	Unrelated	Binary	Stationary	NA	NA	PAF	SAS
Zhu et al., 2009	Family	Linear	Time-varying	Sibship group membership	NA	MASAL	MASAL
Aulchenko et al., 2007	Family	Linear	NA	Random effects	NA	GWAS on LMEM phenotype residuals	GRAMMAR

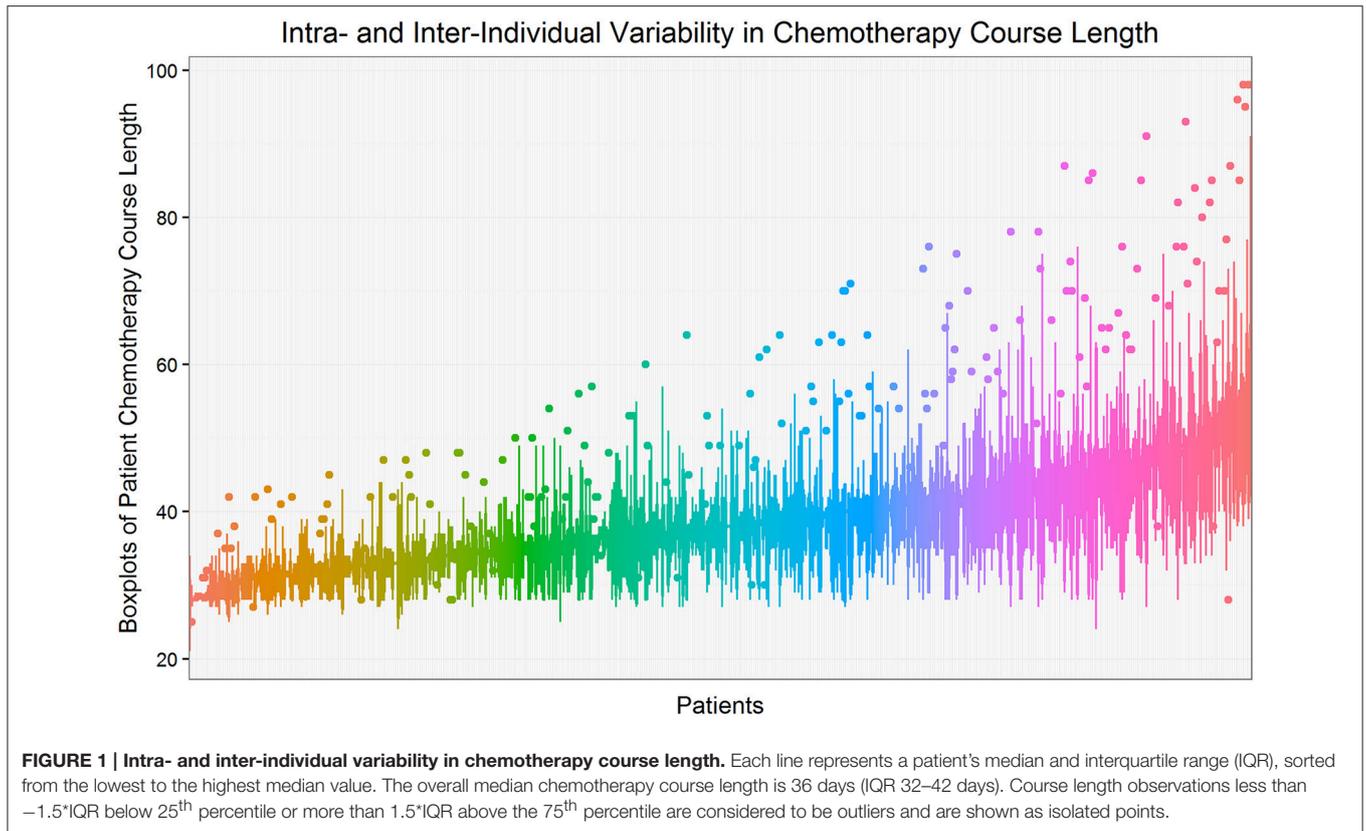
GEE, generalized estimating equation; GLMEM, generalized linear mixed model; GRAMMAR, genome-wide rapid association using mixed model and regression; GWAS, genome-wide association studies; HWE, Hardy-Weinberg equilibrium; LASSO, least absolute shrinkage and selection operator; LMEM, linear mixed model; MAF, minor allele frequency; MCAR, missing completely at random; MGR, missing genotype rate; MNAR, missing not at random; PH, proportional hazard; SNP, single-nucleotide polymorphism; GMM, growth mixture modeling; QTL, quantitative trait locus; LMEM, linear mixed modeling; GEE, generalized estimating equations; MLGM, multivariate linear growth modeling; PAF, population attributable risk fraction; MASAL, multivariate adaptive splines for the analysis of longitudinal data.

subject i at course j . The first model we used is the KIN-LMEM, which is one type of linear mixed effects model (LMEM) and accounts for the individual relatedness by introducing a random effect for each subject that has a covariance matrix estimated from the kinship coefficient (Furlotte et al., 2012). The model assumes that:

$$Y_{ij} = \alpha + \beta \times X_i + u_i + v_{ij} + \varepsilon_{ij} \tag{1}$$

Here X_i represents the state of a particular SNP for subject i and is a fixed effect. u_i is the random effect that captures genetic

relatedness, and assumed to follow a normal distribution with mean 0, variance σ_u^2 , and a correlation of K_{il} between subject i and l (K_{il} is the kinship coefficient). v_{ij} is the random effect for subject i and course j , which follows a normal distribution with mean 0 and variance σ_v^2 . Note that v_{ij} for the same subject but different courses are correlated, i.e., $\text{cov}(v_{ij}, v_{ik}) = \sigma_v^2 D_{jk}$, but v_{ij} for different subjects are independent, i.e., $\text{cov}(v_{ij}, v_{lk}) = 0$. ε_{ij} is the independent error term and follows a normal distribution with mean 0 and variance σ^2 . The three types of random effects u_i , v_{ij} and ε_{ij} are assumed to be independent. This model implies



that the outcomes for the same subject at different courses have a correlation:

$$\text{Corr}(Y_{ij}, Y_{ik}) = (\sigma_v^2 D_{jk} + \sigma_u^2) / (\sigma^2 + \sigma_v^2 + \sigma_u^2);$$

and the outcomes for two different subjects at any course are also correlated due to their genetic relatedness:

$$\text{Corr}(Y_{ij}, Y_{lk}) = \sigma_u^2 K_{il} / (\sigma^2 + \sigma_v^2 + \sigma_u^2)$$

The second model we used is PC-LMEM, a LMEM that accounts for individual relatedness through the use of principal components (PCs) as covariates in the model to adjust for population structure, a common practice in unrelated populations.

The model assumes that:

$$Y_{ij} = \alpha + \beta \times X_i + \gamma \times Z_{ij} + \delta_1 \times P_{1i} + \delta_2 \times P_{2i} + \delta_3 \times P_{3i} + v_i + \varepsilon_{ij} \tag{2}$$

Again X_i is the SNP for subject i as a fixed effect. Z_{ij} is the fixed effect for course. Rather than using a random effect u_i to account for genetic relatedness as in the KIN-LMEM, the PC-LMEM uses the principle components as covariates (three PCs here: P_{1i} , P_{2i} , and P_{3i}). v_i is the random effect to reflect the correlation among the repeated measures within a subject, but takes a simple form of a random intercept, and is assumed to be independent (i.e., $\text{cov}(v_i, v_l) = 0$) and normally distributed (mean 0, variance σ_v^2).

ε_{ij} is the independent error term following normal distribution with mean 0 and variance σ^2 , and the two types of random effects v_i and ε_{ij} are assumed to be independent. This model implies that the outcomes for the same subject at different courses have a correlation:

$$\text{Corr}(Y_{ij}, Y_{ik}) = \sigma_v^2 / (\sigma^2 + \sigma_v^2);$$

but the outcomes for two different subjects at any course are independent since we already adjusted for population substructure through PCs:

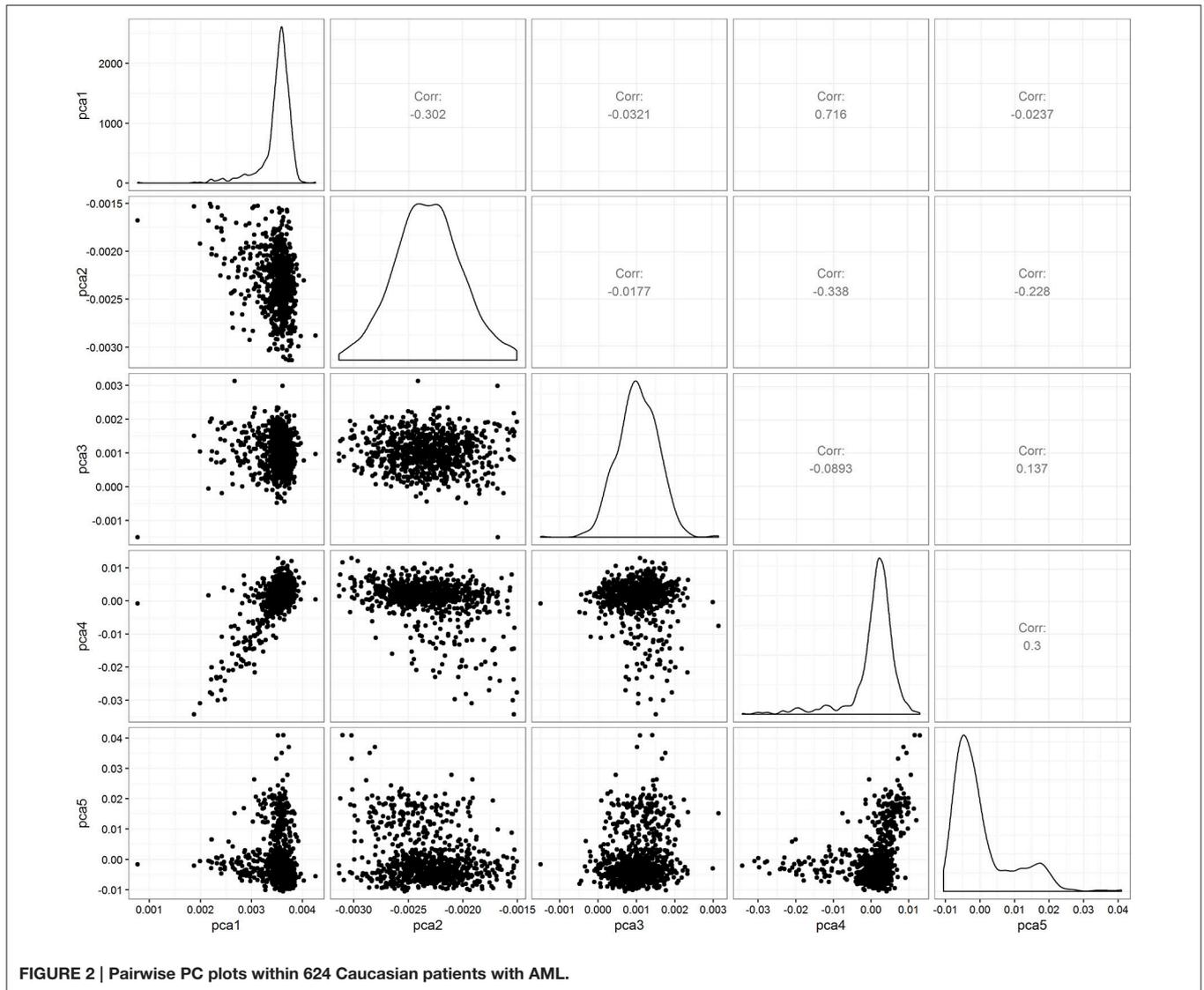
$$\text{Corr}(Y_{ij}, Y_{lk}) = 0.$$

PC-GEE is the third model, an alternative longitudinal model to LMEM. This model adjusts for individual relatedness through the use of PCs and accounts for correlations of repeated measures within a subject through the use of robust covariance structure. Specifically the model assumes that:

$$Y_{ij} = \alpha + \beta \times X_i + \gamma \times Z_{ij} + \delta_1 \times P_{1i} + \delta_2 \times P_{2i} + \delta_3 \times P_{3i} + \varepsilon_{ij} \tag{3}$$

$$\text{Var}(Y_{ij}) = \sigma^2, \text{Corr}(Y_{ij}, Y_{ik}) = \rho, \text{Corr}(Y_{ij}, Y_{lk}) = 0$$

The notations here are the same as in Equation (2), except that we no longer have the random effect v_i and rather explicitly specify an “exchangeable” correlation structure ρ for the repeated



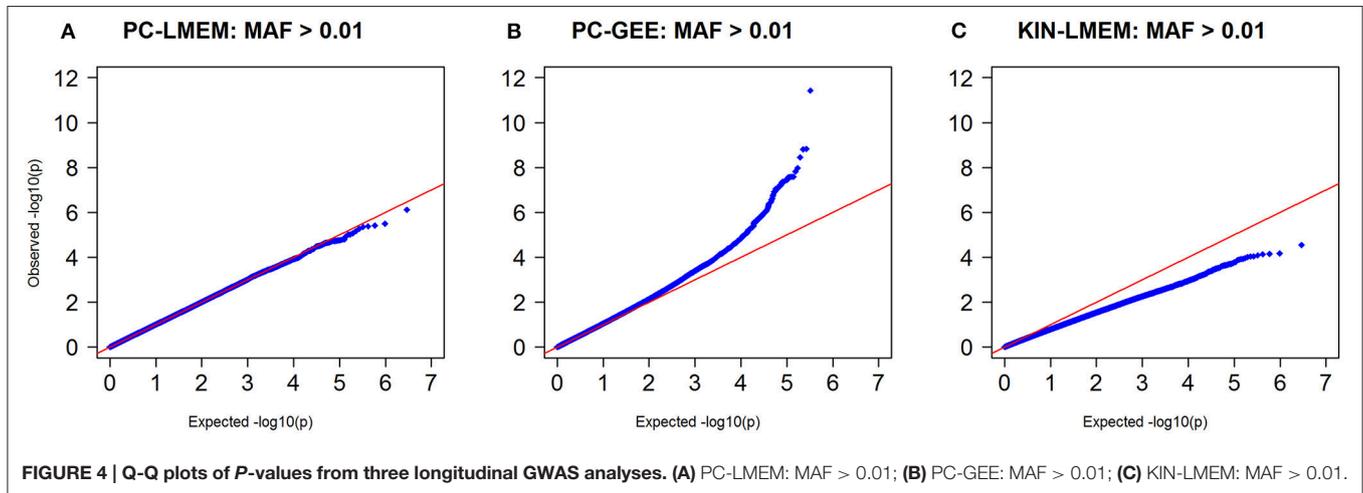
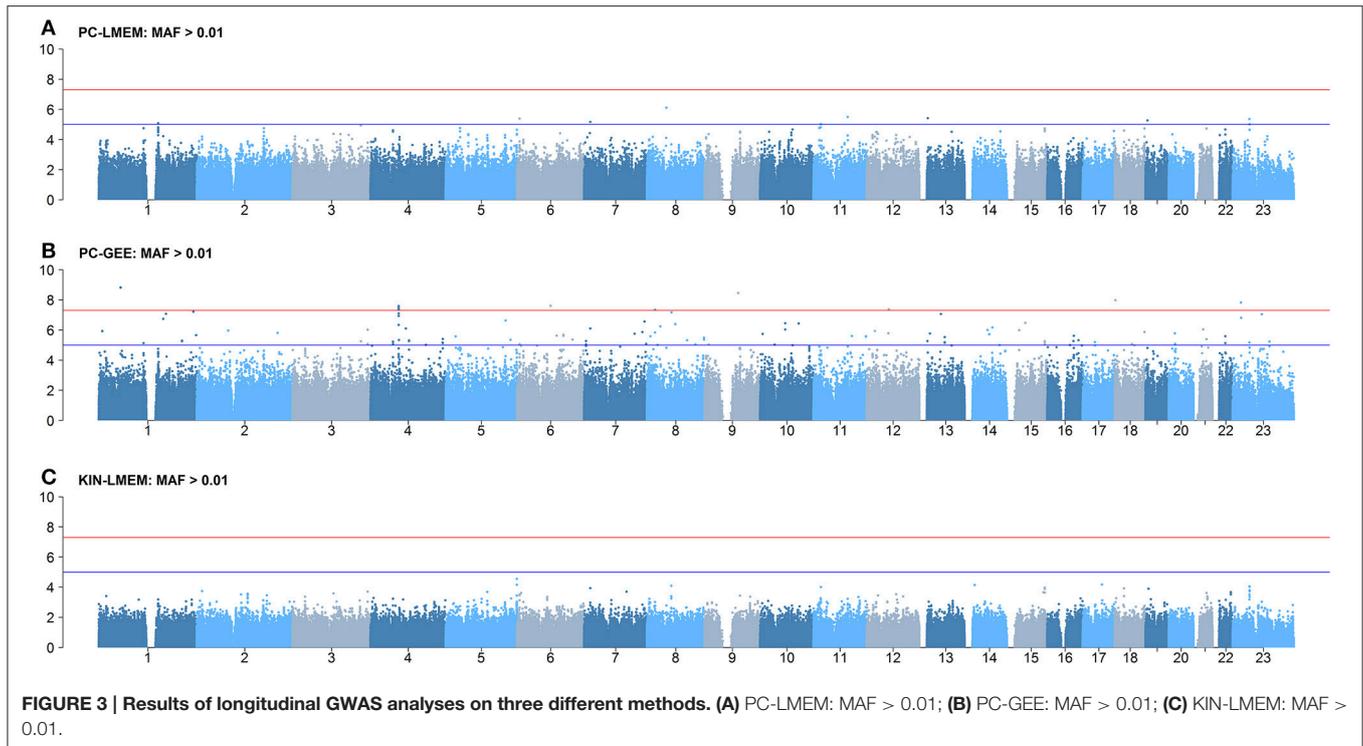
measures within a subject. We note that LMEMs (KIN-LMEM and PC-LMEM) are likelihood-based and valid under the missing at random (MAR) assumption, but PC-GEE is valid only under the missing completely at random (MCAR) assumption.

Statistical Methods

Extensive genotyping quality control checks were performed using PLINK, and SNPs were excluded from analysis in case of (1) call rates < 95%; (2) monomorphic SNPs (MAF < 0.01); and (3) deviation from Hardy-Weinberg Equilibrium ($P < 10^{-5}$) (Purcell et al., 2007). GCTA software was used to identify duplicates among genotyped samples, calculate the kinship coefficients matrix, and ancestral groups were constructed via Principal Components Analysis (PCA), and patients showing familial structure and/or cryptic relatedness were excluded (Yang et al., 2011). The genotype data of the subset of unrelated patients in our cohort was then merged with data from Caucasian participants from HapMap3. Naive Bayes classification was

performed using HapMap3 as the training set. Remaining heterogeneity between individuals of European descent are illustrated with scatterplots between principal components. Longitudinal analyses were performed with R packages *gee* and *lme4*, and the R script provided by Furlotte et al. (2012). Covariates in PC-GEE and PC-LMEM include course number and seven principal components as indicators of ancestry.

We used several metrics to evaluate the performance of the three analytics methods. First, quantile-quantile (Q-Q) plots were estimated with the R package *qq-man* to ensure that the observed P -value distribution follows a Chi-Square null distribution with exception of the extreme tail. Second, Manhattan plots were generated to evaluate the differences in the global pattern of the significance across three methods. Third, correlation coefficients between the observed P -values from different methods were calculated to examine their concordance. Finally, we test the interchangeability of the results by evaluating to what extent the different methods produce similar P -values,

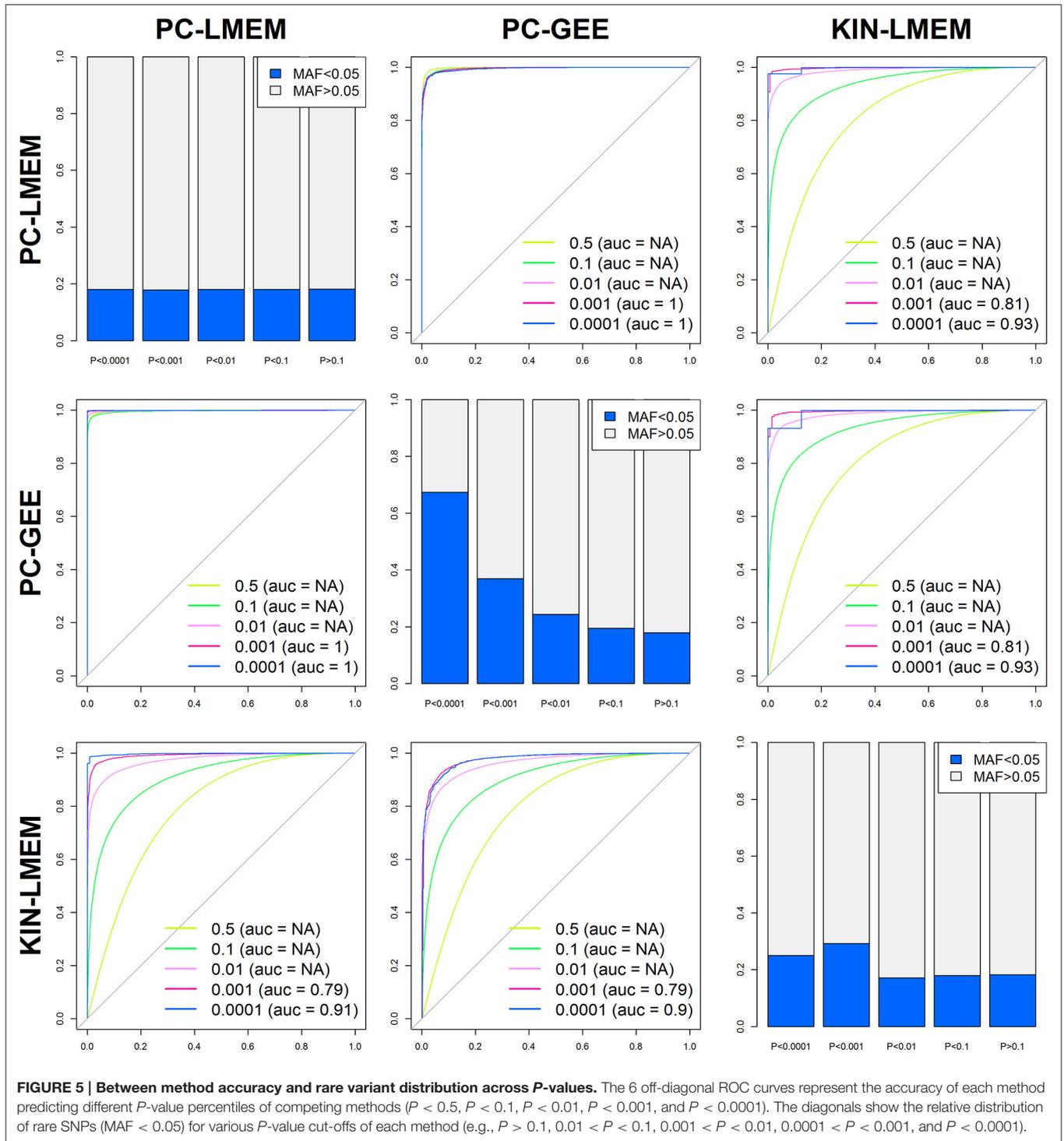


taken at decreasing thresholds of 0.5, 0.1, 0.01, 0.001, and 0.0001. We then summarized the predictive accuracy for each method by their competing alternatives by calculating true positive and false positive rates using the ROCR package in R. We plotted all receiver operating characteristic (ROC) curves considering the area under the curve (AUC) as a measure of predictive performance.

RESULTS

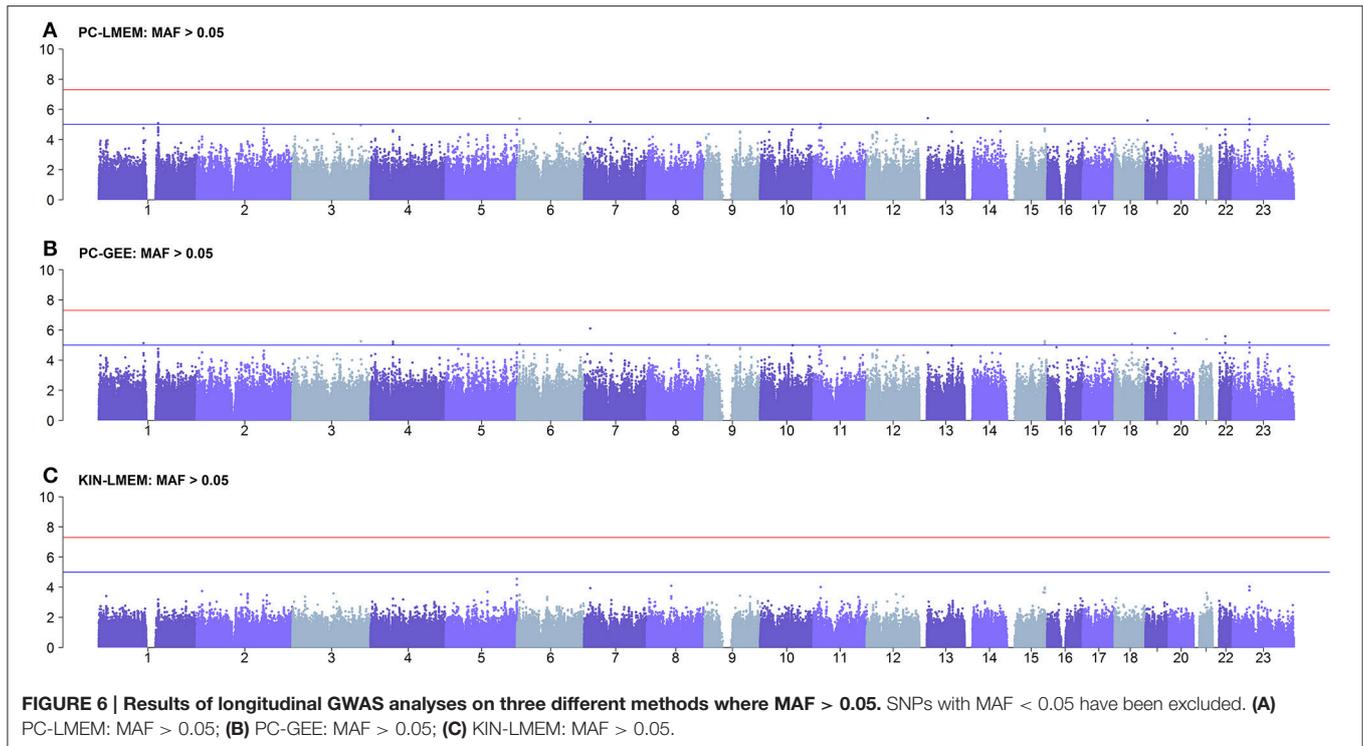
Our search for studies related to longitudinal GWAS yielded 19 results. **Table 1** summarizes longitudinal GWAS evaluation

studies to date. The majority of the study focused on methods accounting for complex pedigree structure structures. Nine studies performed a single-step longitudinal analysis where LMEM and GEE were mostly used (Chang et al., 2009; Kerner et al., 2009; Park et al., 2009; Zhu et al., 2009; Furlotte et al., 2012; Choi et al., 2014; Hossain and Beyene, 2014; Tan et al., 2014). A two-step approach was utilized by others where a first a summary measure for the longitudinal phenotype was extracted and subsequently single observational analysis was performed (Aulchenko et al., 2007; Fradin and Fallin, 2009; Roslin et al., 2009; Yan et al., 2009; Eu-Ahsunthornwattana et al., 2014; Musolf et al., 2014; Vaitiakhovich et al., 2014; Wang et al., 2014; Xia and Lin, 2014). Incidental reports include cluster analysis,



hierarchical multi-level linear modeling, Bayesian LASSO with B-splines, MLGM, and MASAL using various software packages. Model selection, missing data and small sample size were the two most common challenges. There were no loci consistently found the across studies. Linear mixed models were most accurate at confirming known SNPs.

Figure 1 shows the intra- and inter-individual variability in chemotherapy course length for the cohort of 624 pediatric AML patients. Each line represents a patient's median and interquartile range (IQR), sorted from the lowest to the highest median value. The overall median chemotherapy course length is 36 days (IQR 32–42 days). The ancestry plot depicting the relationship



between principal components (Figure 2) illustrates the within-population heterogeneity for all patients of European descent.

Figures 3, 4 show the Manhattan and QQ-plots for PC-GEE, PC-LMEM, and KIN-LMEM. The QQ-plot from PC-GEE shows an increased rate of false positives indicating the necessity for an additional round of genomic control. The KIN-LMEM model suffers of a substantial type II error rate under the null Chi Square distribution. With KIN-LMEM a small cluster of SNPs of below genome-wide threshold P -values was observed on chromosome 5, which was not captured with PC-GEE and PC-LMEM. Finally, PC-LMEM shows an expected pattern of significance with low type I and type II error rates.

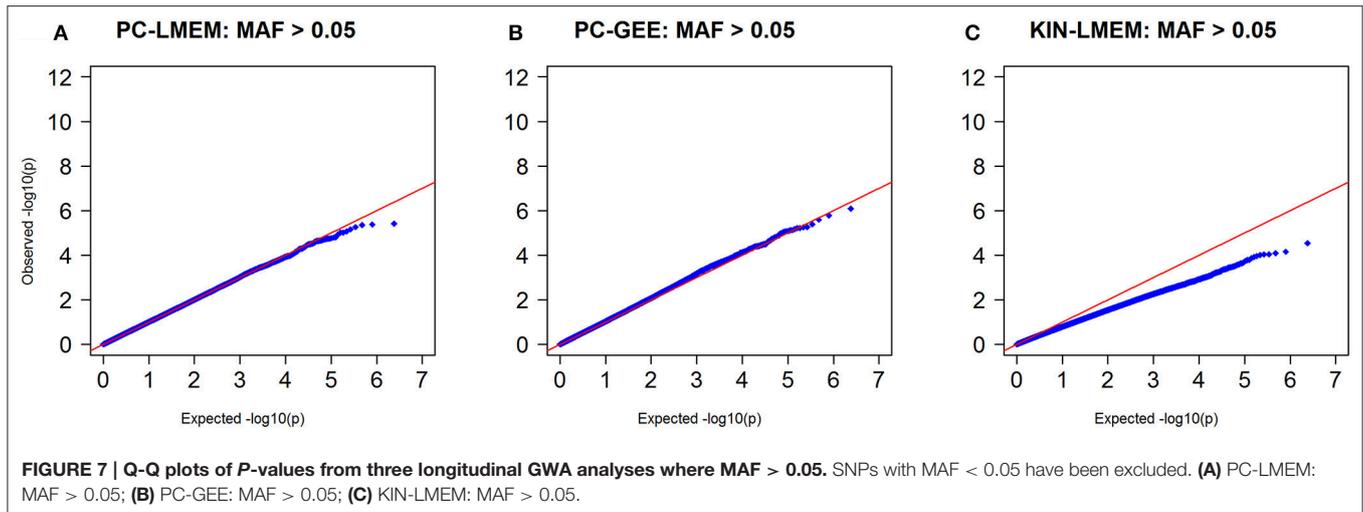
Even though the overall P -values show distinct global patterns as evident from the QQ-plot and Manhattan plot, the three methods show a strong underlying comparability as measured by correlation coefficients. The correlation coefficient between PC-LMEM and PC-GEE P -values exceeds > 0.99. The correlation between PC-LMEM and KIN-LMEM is relatively high ($r^2 = 0.605$), and the same accounts for PC-GEE and KIN-LMEM ($r^2 = 0.603$). Further exploration of between-method similarities and differences are shown in Figure 5. The diagonals in this figure show the relative distribution of rare SNPs (MAF < 0.05) for various P -value cut-offs of each method (e.g., $P > 0.1$, $0.01 < P < 0.1$, $0.001 < P < 0.01$, $0.0001 < P < 0.001$, and $P < 0.0001$). In PC-GEE the lowest P -values ($p < 0.001$) are enriched for rare SNPs. In contrast, PC-LMEM and KIN-LMEM analyses are less affected by SNP prevalence. Off-diagonal plots show ROC curves with AUCs for the prediction of P -value thresholds of the respective method by its competing approaches. The results from PC-LMEM can be predicted with very high accuracy by

PC-GEE across the P -value spectrum (AUC > 0.99 for each P -value threshold), which is in line of expectation due to their extremely high correlation. KIN-LMEM P -values show a more heterogeneous range of predictive performances by both PC-LMEM and PC-GEE, where improvements in predictive accuracy are observed with decreasing P -values. For example, the AUC exceeds 0.99 for the lower range of KIN-LMEM P -values (e.g., $p < 0.01$). In summary, any observed P -value < 0.001 from each method shows to be very accurately predicted by its competing methods (AUC > 0.99).

To confirm whether the inflated type I error in PC-GEE is attributable to rare variants, we recreated the Manhattan and QQ-plots excluding SNPs with MAF < 0.05 (Figures 6, 7). The QQ-plots for PC-LMEM and KIN-LMEM were not affected, however the inflated false positive rate in PC-GEE has normalized and observed P -values follow a Chi-Square null distribution. These results indicate that longitudinal GWAS using GEE is not an appropriate method for investigating rare variant associations. In the longitudinal GWAS analyses restricted to MAF > 0.05, PC-GEE, and PC-LMEM show almost perfect concordance.

DISCUSSION

We compared the performance between three models for GWAS data with a longitudinal, repeatedly measured outcome with missing data from two Phase III clinical trials. The three methods show remarkably different rates of false positives and false negatives under the Chi-Square null distribution. Our results indicate that the PC-LMEM approach shows the better



performance to KIN-LMEM and PC-GEE in this scenario of unbalanced longitudinal data in an unrelated population. Modeling longitudinal data differs from single observations because it needs to account for correlation structures arising from repeated measurements of a trait within an individual, and potential differential missingness of repeated phenotype measures over time.

Two methods to account for genetic relatedness between study individuals are adjustment of PC as covariates (as the PC-LMEM) and the use of a kinship matrix (as the KIN-LMEM). PC-LMEM accounts for population substructure by restriction and genomic control, e.g., first identifying close relatives to remove them from analysis, then correcting for broad sample structure using principal components or spatial information and finally correcting for the residual inflation with genomic control.

The KIN-LMEM method is used in complex pedigrees encompassing both hidden relatedness and population stratification. Recently, linear mixed models based approaches have been proposed as an alternative to PC-based approaches when adjusting for population stratification in studies of unrelated individuals (Kang et al., 2010; Zhang et al., 2010). The covariance structure for the random effect is generally assumed to correspond to that implied by a polygenic model, incorporating the genetic relationship (kinship) between each pair of individuals. The rationale for this approach is that apparently unrelated individuals may nevertheless display distant levels of common ancestry. In our study it seems the KIN-LMEM inflates the type II error rates which may lead to a loss a power and the inability to detect true positives. However, for longitudinal GWAS in family studies, PC-based methods may be insufficient to account for more complex genetic relatedness structures, and KIN-LMEM or other family-based approaches should be considered.

Our observations indicate that GEE is unable to produce a reliable association in SNPs with a low MAF. Even though GEE does not require assumptions on the joint distribution of observed data and random effects it does depend on a correct

specification of covariate-specific means of the outcome. In rare variants the minor allele group contains too few observations to produce reliable marginal population means (Diggle, 2002).

In our study we used chemotherapy course as a categorical covariate and therefore did not use Sikorska method. Sikorska's fast two-step estimation method is recommended in a scenario of continuous time elements and both random intercepts and random slopes are needed. If time (e.g., chemotherapy course) should be considered as categorical and thus random slope is not applicable, using our method is more appropriate. A limitation of our study is the lack of a true genetic association with course duration that inhibits the generation of false negative rates and operational characteristics for comparing the detection of SNPs truly associated with the outcome.

In pediatric AML, chemotherapy course length is a primary determinant of infection risk and shows a substantial intra-individual correlation and inter-individual variability. We previously showed that MTHFR polymorphisms were not associated with a significantly altered risk of chemotherapy course length (Aplenc et al., 2005). However, Murphy et al. showed that in the MTHFR C677CT wildtypes, folic acid was significantly associated with more than two-fold increased neutrophil recovery compared to the CT and TT genotypes (Murphy et al., 2012). Unraveling genetic components for chemotherapy course length has the clinical potential for personalized toxicity monitoring. We anticipate on reanalyzing the data with a larger sample size in future using PC-LMEM.

In conclusion, the false positive rates of the three methods were remarkably different and the PC-LMEM model seems to provide a reliable approximation of the P-value for the effect of SNP on a temporal phenotype. PC-LMEM loses no power compared to the KIN-LEMM and shows no inflation of the type II error rate compared to PC-GEE in SNPs with low MAFs. PC-LMEM holds perhaps the greatest promise for longitudinal GWAS because of its flexibility in accounting for correlation structures and its validity for rare variant association studies. We recommend a practical framework based on PC-LMEM for high-throughput genetic analysis of longitudinal data implemented

in a PLINK/R framework in a high-performance computing environment.

ETHICS STATEMENT

Written informed consent and ethics committee approval were available for all study participants. Informed consent was obtained in accordance with the Declaration of Helsinki. The institutional review boards of all participating institutions approved the clinical protocol, and the COG Myeloid Disease Biology Committee approved all studies.

AUTHOR CONTRIBUTIONS

The authors have been substantial contributions to (1) the conception and design of the study, (2) acquisition of data (MV, TA, AG, RA), (3) analysis and interpretation of data (MV, RA,

YL), (4) drafting the article or revising it critically for important intellectual content (MV, YL, RA, TA, AG), and (5) final approval of the version to be submitted (MV, YL, RA, TA, AG).

FUNDING

Research is supported by the Chair's Grant U10 CA98543-08 and Statistics and Data Center Grant U10 CA180899-02 of the Children's Oncology Group, and Dr. Aplenc's R01 grant 1R01 CA165277, from the National Institutes of Health (NIH). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

ACKNOWLEDGMENTS

The authors thank our clinical and research faculty and staff and the patients and their families for participating.

REFERENCES

- Aplenc, R., Thompson, J., Han, P., La, M., Zhao, H., Lange, B., et al. (2005). Methylenetetrahydrofolate reductase polymorphisms and therapy response in pediatric acute lymphoblastic leukemia. *Cancer Res.* 65, 2482–2487. doi: 10.1158/0008-5472.CAN-04-2606
- Aulchenko, Y. S., de Koning, D. J., and Haley, C. (2007). Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics* 177, 577–585. doi: 10.1534/genetics.107.075614
- Beyene, J., and Hamid, J. S. (2014). Longitudinal data analysis in genome-wide association studies. *Genet. Epidemiol.* 38(Suppl. 1), S68–S73. doi: 10.1002/gepi.21828
- Chang, S. W., Choi, S. H., Li, K., Fleur, R. S., Huang, C., Shen, T., et al. (2009). Growth mixture modeling as an exploratory analysis tool in longitudinal quantitative trait loci analysis. *BMC Proc.* 3(Suppl. 7):S112. doi: 10.1186/1753-6561-3-s7-s112
- Choi, Y. H., Chowdhury, R., and Swaminathan, B. (2014). Prediction of hypertension based on the genetic analysis of longitudinal phenotypes: a comparison of different modeling approaches for the binary trait of hypertension. *BMC Proc.* 8(Suppl. 1):S78. doi: 10.1186/1753-6561-8-S1-S7
- Diggle, P. J. (2002). *Analysis of Longitudinal Data*. Oxford: Oxford University Press.
- Eu-Ahsunthornwattana, J., Howey, R. A., and Cordell, H. J. (2014). Accounting for relatedness in family-based association studies: application to Genetic Analysis Workshop 18 data. *BMC Proc.* 8(Suppl. 1):S79. doi: 10.1186/1753-6561-8-S1-S79
- Fradin, D. D., and Fallin, M. D. (2009). Influence of control selection in genome-wide association studies: the example of diabetes in the Framingham Heart Study. *BMC Proc.* 3(Suppl. 7):S113. doi: 10.1186/1753-6561-3-s7-s113
- Furlotte, N. A., Eskin, E., and Eyheramendy, S. (2012). Genome-wide association mapping with longitudinal data. *Genet. Epidemiol.* 36, 463–471. doi: 10.1002/gepi.21640
- Gamis, A. S., Alonzo, T. A., Meshinchi, S., Sung, L., Gerbing, R. B., Raimondi, S. C., et al. (2014). Gemtuzumab ozogamicin in children and adolescents with de novo acute myeloid leukemia improves event-free survival by reducing relapse risk: results from the randomized phase III Children's Oncology Group trial AAML0531. *J. Clin. Oncol.* 32, 3021–3032. doi: 10.1200/JCO.2014.55.3628
- Hossain, A., and Beyene, J. (2014). Analysis of baseline, average, and longitudinally measured blood pressure data using linear mixed models. *BMC Proc.* 8(Suppl. 1):S80. doi: 10.1186/1753-6561-8-S1-S80
- Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S. Y., Freimer, N. B., et al. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* 42, 348–354. doi: 10.1038/ng.548
- Kerner, B., North, K. E., and Fallin, M. D. (2009). Use of longitudinal data in genetic studies in the genome-wide association studies era: summary of Group 14. *Genet. Epidemiol.* 33(Suppl. 1), S93–S98. doi: 10.1002/gepi.20479
- Luan, J., Kerner, B., Zhao, J.-H., Loos, R. J. F., Sharp, S. J., Muthén, B. O., et al. (2009). A multilevel linear mixed model of the association between candidate genes and weight and body mass index using the Framingham longitudinal family data. *BMC Proc.* 3(Suppl. 7):S115. doi: 10.1186/1753-6561-3-s7-s115
- Meshinchi, S., Hunger, S. P., Aplenc, R., Adamson, P. C., and Jessup, J. M. (2012). Lessons learned from the investigational device exemption review of Children's Oncology Group trial AAML1031. *Clin. Cancer Res.* 18, 1547–1554. doi: 10.1158/1078-0432.CCR-11-2205
- Murphy, N. M., Diviney, M., Szer, J., Bardy, P., Grigg, A., Hoyt, R., et al. (2012). The effect of folinic acid on methylenetetrahydrofolate reductase polymorphisms in methotrexate-treated allogeneic hematopoietic stem cell transplants. *Biol. Blood Marrow Transplant.* 18, 722–730. doi: 10.1016/j.bbmt.2011.09.001
- Musolf, A., Nato, A. Q. Jr., Londono, D., Zhou, L., Matisse, T. C., and Gordon, D. (2014). Mapping genes with longitudinal phenotypes via Bayesian posterior probabilities. *BMC Proc.* 8(Suppl. 1):S81. doi: 10.1186/1753-6561-8-S1-S81
- Nyholt, D. R. (2004). A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *Am. J. Hum. Genet.* 74, 765–769. doi: 10.1086/383251
- Park, Y. M., Province, M. A., Gao, X., Feitosa, M., Wu, J., Ma, D., et al. (2009). Longitudinal trends in the association of metabolic syndrome with 550 k single-nucleotide polymorphisms in the Framingham Heart Study. *BMC Proc.* 3(Suppl. 7):S116. doi: 10.1186/1753-6561-3-s7-s116
- Patterson, N., Price, A. L., and Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genet.* 2:e190. doi: 10.1371/journal.pgen.0020190
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795
- Roslin, N. M., Hamid, J. S., Paterson, A. D., and Beyene, J. (2009). Genome-wide association analysis of cardiovascular-related quantitative traits in the Framingham Heart Study. *BMC Proc.* 3(Suppl. 7):S117. doi: 10.1186/1753-6561-3-s7-s117
- Sikorska, K., Montazeri, N. M., Uitterlinden, A., Rivadeneira, F., Eilers, P. H. C., and Lesaffre, E. (2015). GWAS with longitudinal phenotypes: performance of approximate procedures. *Eur. J. Hum. Genet.* 23, 1384–1391. doi: 10.1038/ejhg.2015.1
- Tan, Q., Hjelmborg, J. V. B., Thomassen, M., Jensen, A. K., Christiansen, L., Christensen, K., et al. (2014). Hierarchical linear modeling of longitudinal pedigree data for genetic association analysis. *BMC Proc.* 8(Suppl. 1):S82. doi: 10.1186/1753-6561-8-S1-S82

- Vaitsakhovich, T., Drichel, D., Angisch, M., Becker, T., Herold, C., and Lacour, A. (2014). Analysis of the progression of systolic blood pressure using imputation of missing phenotype values. *BMC Proc.* 8(Suppl. 1):S83. doi: 10.1186/1753-6561-8-S1-S83
- Wang, S., Gao, W., Ngwa, J., Allard, C., Liu, C. T., and Cupples, L. A. (2014). Comparing baseline and longitudinal measures in association studies. *BMC Proc.* 8(Suppl. 1):S84. doi: 10.1186/1753-6561-8-S1-S84
- Weir, B. S., Anderson, A. D., and Hepler, A. B. (2006). Genetic relatedness analysis: modern data and new challenges. *Nat. Rev. Genet.* 7, 771–780. doi: 10.1038/nrg1960
- Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., et al. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* 42, D1001–D1006. doi: 10.1093/nar/gkt1229
- Xia, S., and Lin, S. (2014). Detecting longitudinal effects of haplotypes and smoking on hypertension using B-splines and Bayesian LASSO. *BMC Proc.* 8(Suppl. 1):S85. doi: 10.1186/1753-6561-8-S1-S85
- Yan, Y., Hu, Y., North, K. E., Franceschini, N., and Lin, D. (2009). Evaluation of population impact of candidate polymorphisms for coronary heart disease in the Framingham Heart Study Offspring Cohort. *BMC Proc.* 3(Suppl. 7):S118. doi: 10.1186/1753-6561-3-s7-s118
- Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M. (2011). GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* 88, 76–82. doi: 10.1016/j.ajhg.2010.11.011
- Zhang, Z., Ersoz, E., Lai, C. Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., et al. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* 42, 355–360. doi: 10.1038/ng.546
- Zhu, W., Cho, K., Chen, X., Zhang, M., Wang, M., and Zhang, H. (2009). A genome-wide association analysis of Framingham Heart Study longitudinal data using multivariate adaptive splines. *BMC Proc.* 3(Suppl. 7):S119. doi: 10.1186/1753-6561-3-s7-s119

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer FS and handling Editor declared their shared affiliation, and the handling Editor states that the process nevertheless met the standards of a fair and objective review.

Copyright © 2016 Vujkovic, Aplenc, Alonzo, Gamis and Li. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.